EKONOMETRIA ECONOMETRICS 4(38) • 2012

ISSN 1507-3866

Paweł Siarka
I 4FS, Żerniki Wrocławskie

METODA ILORAZU ODLEGŁOŚCI – ZAGADNIENIE GRAFICZNEJ PREZENTACJI OBSERWACJI WIELOWYMIAROWYCH

Streszczenie: W artykule zaprezentowano autorską metodę analizy danych wielowymiarowych. Jej koncepcja oparta została na zasadzie redukcji wymiaru przestrzeni w wyniku rzutowania obserwacji na płaszczyznę. Tak uzyskany obraz jest podstawą dalszej analizy wzrokowej badanych obserwacji. Głównym zadaniem zaprezentowanej metody jest dokonanie rotacji obiektów w przestrzeni cech w taki sposób, aby uzyskany obraz uwidocznił ewentualną niejednorodność populacji. Zaproponowane podejście przedstawione zostało na tle metody głównych składowych. Wyniki przeprowadzonych symulacji wykazały, że autorska metoda jest szczególnie skuteczna w procesie wykrywania niejednorodnych populacji. Istotną jej zaletą jest brak konieczności uprzedniej standaryzacji zmiennych, co zapewnia funkcja kryterium warunkująca wybór właściwego rzutowania obserwacji.

Słowa kluczowe: ryzyko kredytowe, rozpoznawanie obrazów, klasyfikacja.

1. Wstęp

Analiza danych wielowymiarowych pociąga za sobą na ogół konieczność zobrazowania obiektów znajdujących się w przestrzeni o wymiarze większym aniżeli trzy. Ograniczenia wynikające z ludzkiej percepcji sprawiają, że geometryczne wyobrażenie o strukturze zbiorów obserwacji jest zazwyczaj niepełne. Stąd naukowcy do wielu lat prowadzą badania nad metodami umożliwiającymi prezentację obiektów wielowymiarowych w prostszej postaci, tj. redukując wymiar przestrzeni. Przekształcenie oryginalnych danych wiąże się oczywiście z utratą części informacji, jakkolwiek jest to koszt, który badacz skłonny jest ponieść w celu uzyskania wiedzy o badanym zjawisku. Zatem jednym z powodów tak szerokiego wachlarza możliwych podejść do zagadnienia analizy danych wielowymiarowych jest ciągłe poszukiwanie sposobów ograniczania strat powstałych w wyniku przekształcania oryginalnych danych.

W literaturze przedmiotu zagadnienie odnoszące się do ogólnego problemu rozpoznawania obrazów bez nauczyciela było wielokrotnie badane. Wyróżnić można trzy podstawowe grupy podejść obejmujących metody opisowe, metody stochastyczne oraz metody graficzne [Jajuga 1990]. Podejścia opisowe obejmują metody

grupowania oraz metody klasyfikacji rozmytej. W ramach grupy metod stochastycznych wyróżnić można podejście klasyfikacyjne oraz mieszankowe. Ostatnia z wyróżnionych grup obejmuje metody graficzne będące przedmiotem niniejszego artykułu. Ich celem jest przedstawianie obserwacji wielowymiarowych na płaszczyźnie, a następnie ich dalsza analiza.

Nowatorskim rozwiązaniem w procesie prezentacji danych wielowymiarowych posłużył się Chernoff [1973], który w tym celu wykorzystał twarze, w ramach których zakodowane zostały poszczególne wartości cech opisujących obiekty. Innym rozwiązaniem szeroko wykorzystywanym przez badaczy są wykresy gwiazdowe, nazywane również sieciami pajęczymi (*spider chart*). W ramach tego podejścia obserwacje wielowymiarowe przedstawiane są na płaszczyźnie w postaci gwiazd, których długości ramion odpowiadają wartościom poszczególnych cech. Metodę tę jako pierwszy wykorzystał G. von Mayr [1877] jeszcze w XIX wieku. Użyteczne podejście zaprezentował Gabriel [1971], nazywa się je metodą podwójnego wykresu (*biplot graphic display*). Koncepcja ta umożliwia prezentację na płaszczyźnie zarówno obserwacji wielowymiarowych, jak i samych cech je opisujących. Kolejnym sposobem prezentacji obserwacji wielowymiarowych jest wykres typu *paraller coordinates,* którego pierwsze wykorzystanie odnotowano jeszcze w XIX wieku. Inselberg, którego można uznać za popularyzatora niniejszej metody, poświęcił jej wiele publikacji [1985]. Podejście to służy prezentacji obserwacji wielowymiarowych przez odznaczanie wartości poszczególnych cech na równoległych osiach. Połączone liniami punkty umożliwiają wzrokową analizę danych. Dalszym rozwojem tej metody zajmowali się m.in. Moustafa oraz Wegman [2002]. Autorzy zwrócili również uwagę na związek pomiędzy tą metodą a metodą Andrewsa [1972], w której obserwacje przedstawiane są przez krzywą analizowaną w przedziale od $-\pi$ do π. Spośród pozostałych spotykanych w literaturze metod służących prezentacji obserwacji wielowymiarowych wymienić należy metodę *hyberbox* zaproponowaną przez Alperna oraz Cartera [1991]*, radical coordinates visualisation* autorstwa Hoffmana [1999]*, table lens* opracowaną przez Fao oraz Carda [1994] oraz pozostałe, jak np. *recursive pattern* [Keim i in. 1995], *spiral techniques* [Keim, Kriegel 1994]*, dimensional starking* [LeBlanc i in. 1990].

Spośród metod służących prezentacji danych wielowymiarowych szczególnie interesujące wydają się te, które oparte są na zasadzie redukcji wymiaru przestrzeni. Wspomniana redukcja realizowana jest przy tym w postaci prostopadłego rzutu obserwacji na płaszczyznę w przestrzeni *m*-wymiarowej. Innymi słowy celem tych metod jest uzyskanie dwuwymiarowego „zdjęcia" wielowymiarowych obiektów. Ich zaletą jest pozostawanie w ścisłym związku z geometryczną reprezentacją obiektów przez punkty w przestrzeni cech. Dzięki temu oryginalne dane nie podlegają tak znacznej transformacji jak w przypadku pozostałych metod, co oddala niebezpieczeństwo błędnego wnioskowania. Kluczowym zagadnieniem jest przy tym takie ułożenie płaszczyzny rzutowania, aby uzyskany obraz uchwycił w jak największym stopniu charakter struktury danych. Przykładem takiego podejścia jest metoda głów-

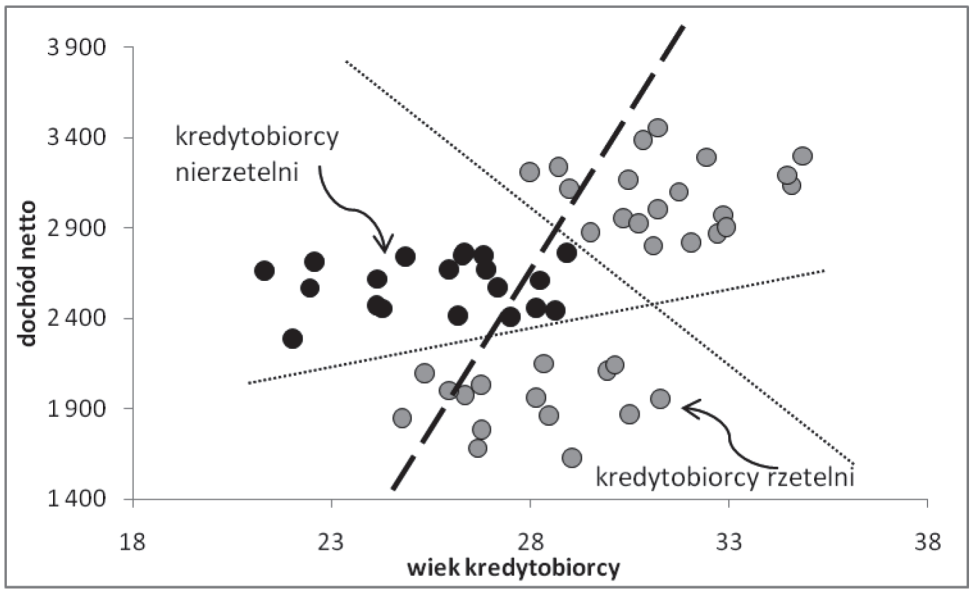
nych składowych zastosowana przez Hotellinga [1933]. Wyznaczone w jej ramach dwie pierwsze składowe odpowiadające największym wartościom własnym macierzy kowariancji umożliwiają prezentację obserwacji na płaszczyźnie.

Celem niniejszego artykułu jest przedstawienie autorskiej metody umożliwiającej graficzną prezentację obserwacji wielowymiarowych na płaszczyźnie. Zadaniem metody jest takie odwzorowanie obserwacji wielowymiarowych w przestrzeni dwuwymiarowej, aby w przypadku braku jednorodności eliptycznej populacji zostało to wyraźnie uwidocznione na powstałym rysunku. Zaproponowane rozwiązanie przedstawiono na tle metody głównych składowych, która wykorzystywana jest w tego typu analizach. Autor przeprowadził szereg symulacji celem porównania wyników własnej metody z metodą głównych składowych. Tym samym szczególna uwaga poświęcona została zweryfikowaniu hipotezy świadczącej o tym, iż w przeciwieństwie do metody głównych składowych proponowane rozwiązanie jest niewrażliwe na wysokie wartości wariancji poszczególnych cech. Stąd możliwe jest tworzenie obrazów dwuwymiarowych pod kątem analizy niejednorodności populacji w odniesieniu do danych oryginalnych (niepoddanych standaryzacji).

2. Problem jednorodności zbiorów w kontekście modeli scoringowych

W celu budowy modeli scoringowych służących ocenie wiarygodności kredytowej klientów banku wykorzystuje się dane historyczne pozyskiwane z systemów bankowych. Znajomość sald zaległości kredytobiorców umożliwia ich podział na grupę rzetelnych oraz nierzetelnych klientów, co wykorzystywane jest następnie w postaci próby uczącej. Na bazie uzyskanych obserwacji estymowane są parametry modelu. Brak jednorodności eliptycznej populacji kredytobiorców dobrych lub złych może być przyczyną niskiej efektywności modeli scoringowych. Wykorzystywane modele scoringowe należą zwykle do klasy tzw. modeli liniowych. W ramach tego podejścia w przestrzeni cech opisujących poszczególnych kredytobiorców budowana jest hiperpłaszczyzna mająca za zadanie jak najlepsze odseparowanie kredytobiorców rzetelnych od nierzetelnych. Gdy jedna lub obie populacje kredytobiorców są niejednorodne, wówczas może okazać się, że wykorzystanie jednego modelu liniowego nie jest wystarczające i prowadzi do wielu nietrafnych prognoz, w rezultacie czego jakość modelu scoringowego znacznie się obniża. Przypadek taki przedstawia rys. 1, na którym w przestrzeni dwóch cech kredytobiorców przedstawiono przykładowe obserwacje reprezentujące klientów banku. Kolorem czarnym oznaczono kredytobiorców nierzetelnych, natomiast kolorem szarym kredytobiorców rzetelnych.

Brak jednorodności eliptycznej kredytobiorców rzetelnych sprawia, że nie jest możliwe wyznaczenie jednej linii prostej, która poprawnie rozdzieliłaby klientów rzetelnych od nierzetelnych. Próby budowy jednego modelu scoringowego skutkują oszacowaniem modelu oznaczonego linią przerywaną. Zasada działania modelu polega na klasyfikacji wszystkich obserwacji znajdujących się na lewo od linii prze-



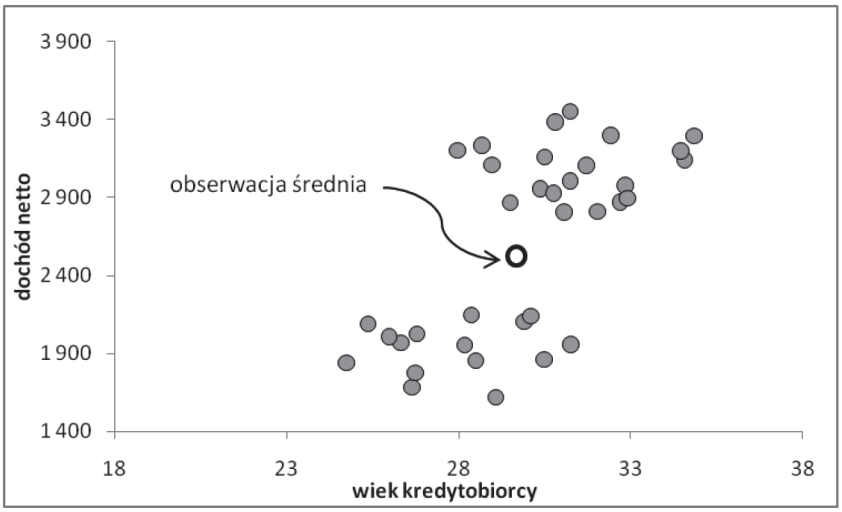
Rys. 1. Prezentacja zagadnienia jednorodności eliptycznej populacji kredytobiorców

Źródło: opracowanie własne.

rywanej jako kredytobiorców nierzetelnych, natomiast pozostałych obserwacji jako kredytobiorców rzetelnych. Na rysunku 1 można zauważyć, że wprawdzie prezentowany model w przeważającej liczbie przypadków poprawnie klasyfikuje obiekty, jednak część obserwacji pozostaje sklasyfikowana błędnie. Warto zwrócić uwagę na fakt, iż posiadając informację o braku jednorodności populacji kredytów rzetelnych, można zbudować dwa modele liniowe (linie kropkowane) zapewniające bezbłędną klasyfikację obserwacji. Zatem kluczowe w procesie budowy modeli scoringowych jest wykrycie ewentualnych skupisk obserwacji świadczących o niejednorodności populacji.

Spośród podejść wykorzystywanych celem oceny jednorodności eliptycznej populacji interesujące rozwiązania stanowią metody umożliwiające prezentację danych w sposób graficzny przez rzutowanie ich na płaszczyznę. Analiza obiektów w przestrzeni o wymiarze większym niż trzy nastręcza wiele kłopotów wynikających z ograniczonej ludzkiej percepcji. Stąd badacz wykorzystujący metody zakładające *a priori* istnienie jednorodności populacji pozostaje w niepewności co do rezultatów swoich badań. Wiele metod wykorzystywanych w ramach statystycznej analizy wielowymiarowej opartych jest na założeniu występowania jednorodności danych, gdzie obserwacje pochodzą z jednomodalnej populacji. W przypadku braku jednorodności obiektów nawet taka miara jak średnia traci swoje właściwości, nie wskazując typowej obserwacji, lecz np. punkt leżący pomiędzy dwiema populacjami, w okolicy którego brak jest jakichkolwiek obserwacji. Przykład zaprezentowany

na rys. 2 dowodzi istotności zagadnienia oceny jednorodności populacji w procesie budowy modeli scoringowych.



Rys. 2. Prezentacja obserwacji średniej dla populacji niejednorodnej

Źródło: opracowanie własne.

W przypadku gdy wymiar analizowanej przestrzeni przekracza trzy, nie sposób posłużyć się geometryczną interpretacją z powodów wcześniej wspomnianych. Stąd wielu badaczy korzysta z szerokiej klasy metod wchodzących w skład analizy skupień opartych na matematycznych kryteriach. Jakkolwiek jednym ze sposobów analizy danych wielowymiarowych jest ich graficzna prezentacja w przestrzeni o wymiarze dostępnym dla ludzkiej percepcji. Warto przy tym pamiętać, że geometryczna interpretacja w ocenie Karla Pearsona powinna być podstawową metodą badania materiału statystycznego, a nie tylko stanowić kolejne narzędzie prezentacji danych.

3. Zagadnienie redukcji wymiarów

Gdy wymiar przestrzeni, w której znajdują się obserwacje, przekracza dwa, wówczas bezpośrednia prezentacja danych na płaszczyźnie nie jest możliwa. Konieczny jest wówczas zabieg polegający na redukcji jej wymiaru. W niniejszym artykule rozważane jest przedstawienie obserwacji wielowymiarowych z wykorzystaniem przestrzeni dwuwymiarowej. Stąd w dalszej części artykułu badana będzie transformacja postaci:

$$Y = X\boldsymbol{a}, \tag{1}$$

gdzie *X* jest macierzą obserwacji o wymiarach $n \times m$ (*n* – liczba obiektów, *m* – liczba cech). *Y* jest macierzą obserwacji po transformacji (o wymiarach $n \times 2$), natomiast $\boldsymbol{a}$ jest macierzą przekształcenia, przy czym:

$$\boldsymbol{a}^T = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1m} \\ a_{21} & a_{22} & \ldots & a_{2m} \end{bmatrix}. \tag{2}$$

Transformacja dana wzorem (1) odwzorowuje macierz *X* w macierz *Y*, której wektory kolumnowe są liniowymi kombinacjami wektorów cech macierzy *X*. Wiersze macierzy *X* stanowią wartości cech poszczególnych obiektów, którymi w przypadku systemów scoringowych są kredytobiorcy. W wyniku transformacji opisanej wzorem (1) uzyskujemy zatem *n* punktów gotowych do prezentacji na płaszczyźnie. Transformacja polegająca na redukcji wymiaru przestrzeni niewątpliwie wiąże się z utratą części informacji zawartych w obserwacjach *m*-wymiarowych, stąd niezbędna wydaje się ocena jakości uzyskanego przekształcenia.

4. Analiza głównych składowych

Jedną z najstarszych metod służących do graficznej prezentacji obserwacji wielowymiarowych jest analiza głównych składowych zaproponowana w 1901 r. przez K. Pearsona [1901], a zastosowana następnie przez Hotellinga. Metoda ta polega na ortogonalnym przekształceniu *m*-wymiarowego układu zmiennych opisujących obserwacje wielowymiarowe na nowy układ zmiennych nieskorelowanych, tzw. głównych składowych. Przekształcenia tego dokonuje się w taki sposób, aby wariancje kolejnych składowych były coraz mniejsze, przy czym całkowita wariancja wszystkich zmiennych wyjściowych jest równa sumie wariancji wszystkich głównych składowych. Oznacza to, że udział wariancji kolejnych głównych składowych w całkowitej zmienności obserwacji wielowymiarowych jest coraz mniejszy.

W metodzie głównych składowych liniowa transformacja wyraża się wzorem:

$$Y = X\boldsymbol{a}, \tag{3}$$

gdzie macierz $\boldsymbol{a}$ jest macierzą przekształcenia ortogonalnego, stąd jej elementy spełniają następujące warunki:

$$\boldsymbol{a}_j^T \boldsymbol{a}_j = 1 \text{ dla } j = 1, \ldots, m, \tag{4}$$

$$\boldsymbol{a}_j \boldsymbol{a}_k = 0 \text{ gdy } j \neq k; \quad j, k = 1, \ldots, m. \tag{5}$$

W przypadku gdy $\boldsymbol{a}_j^T = [a_{j1}, \ldots, a_{jm}]$ oznacza *j*-tą kolumnę macierzy $\boldsymbol{a}$, wówczas transformację można przedstawić w postaci:

$$Y_j = X\boldsymbol{a_j} = a_{j1}{}^T X_1 + ... + a_{jm}{}^T X_m \quad j = 1, ..., m, \tag{6}$$

gdzie: Y_j oznacza *j*-tą główną składową.

Jak można zauważyć, główne składowe są kombinacjami liniowymi zmiennych X_j oraz pozostają nieskorelowane. Macierz a przy zadanych warunkach (4) oraz (5) może być natomiast interpretowana jako przekształcenie polegające na obrocie obiektów wokół początku układu współrzędnych. Podkreślić przy tym należy, że w wyniku tego przekształcenia struktura obiektów pozostaje niezmieniona ze względu na zachowanie odległości pomiędzy obiektami w przestrzeni *m*-wymiarowej. Zmianie ulega natomiast powstały w wyniku rotacji obraz danych widzianych przez obserwatora znajdującego się w miejscu początku układu współrzędnych.

Zgodnie z metodą głównych składowych, aby dokonać transformacji macierzy *X* w *Y,* konieczne jest wyznaczenie macierzy ***a*** przekształcenia ortogonalnego. W podejściu tym wyznacza się kolejno wektory macierzy ***a***, determinujące główne składowe w taki sposób, aby wariancja kolejnych zmiennych była jak największa. Wariancję poszczególnych głównych składowych można zapisać (przy założeniu, że wartości średnie równają się zero):

$$V(Y_j) = Y_j{}^T Y_j = a_j^T X^T X a_j = a_j^T S a_j, \tag{7}$$

gdzie: macierz *S* jest macierzą kowariancji cech stanowiących składowe macierzy *X*.

Na podstawie równania (7) można zauważyć, że szukany wektor a_j, którego celem jest uzyskanie maksymalnej wariancji zmiennej Y_j (długości wektora Y_j), powinien być wektorem własnym macierzy kowariancji *S*. Z definicji wektora własnego macierzy przekształcenia wynika bowiem, że cała siła przekształcenia skupiona jest na jego wydłużeniu przez przemnożenie składowych przez stałą. W tym przypadku warunek ten zapewnia uzyskanie maksymalnej wariancji zmiennej Y_j. W niniejszym podejściu pierwszej głównej składowej Y_j odpowiada wektor własny o największej wartości własnej, natomiast kolejnym składowym odpowiadają wektory własne o coraz mniejszych wartościach własnych.

5. Metoda ilorazu odległości

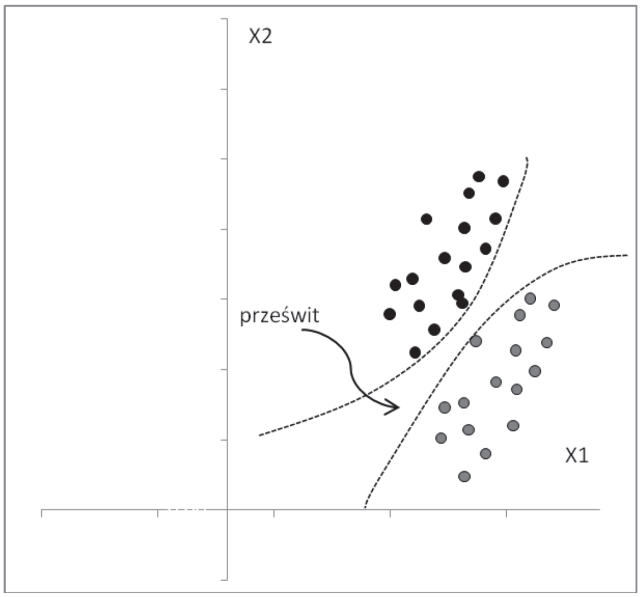
Idea metody głównych składowych sprowadza się do takiego rzutu obserwacji *m*-wymiarowych na płaszczyznę, aby wariancja zmiennych przekształconych była maksymalna. Takie postępowanie jest intuicyjnie oczywiste, gdyż logiczna wydaje się próba „wniknięcia" w zbiór obserwacji i spojrzenia na niego z perspektywy największej dyspersji. Badacz szukający w zbiorze obserwacji naturalnych skupisk ma prawo liczyć na to, że otrzyma obraz ujmujący zależności między obiektami. Intuicyjnie oczywiste jest także to, że niejednokrotnie metoda ta może zawodzić, gdyż szukanie tzw. prześwitów między grupami obserwacji nie musi pokrywać się

z szukaniem maksymalnej wariancji głównych składowych. Niejednokrotnie konieczne jest standaryzowanie zmiennych przed rozpoczęciem analizy głównych składowych. W przypadku bowiem, gdy zabieg standaryzacji nie zostanie przeprowadzony, a jedna ze zmiennych wykazuje istotnie większą bezwzględną zmienność w stosunku do pozostałych zmiennych, wówczas można się spodziewać, że pierwsza wyodrębniona główna składowa będzie „naśladować" cechę o największej wariancji. Stąd konieczny jest często zabieg standaryzacji zmiennych, który jest znaczną ingerencją w strukturę danych.

Aby ominąć problem konieczności standaryzacji danych i jednocześnie umożliwić zaprezentowanie obserwacji w sposób ukazujący istnienie ewentualnych skupisk, opracowana została autorska metoda nazwana metodą ilorazu odległości. Podobnie jak w metodzie głównych składowych szukane jest ortogonalne przekształcenie wielowymiarowych obserwacji, jednak kryterium determinującym przekształcenie zmiennych nie jest maksymalizacja wariancji.

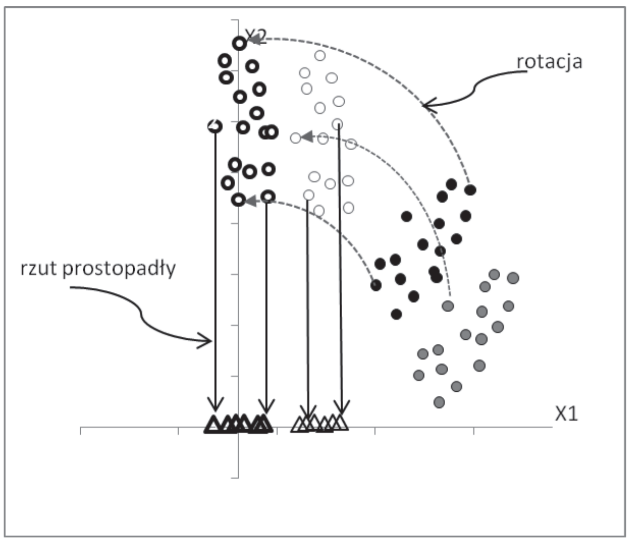
Transformację w proponowanej metodzie przedstawić można w postaci (3), przy warunkach określonych równaniami (4) oraz (5). W celu przedstawienia obserwacji wielowymiarowych na płaszczyźnie konieczne jest wyznaczenie dwóch zmiennych będących liniowymi kombinacjami oryginalnych wektorów cech macierzy X. Istota proponowanej metody sprowadza się do dokonania obrotu obiektów (poszczególnych obserwacji) w przestrzeni m-wymiarowej wokół początku układu współrzędnych w taki sposób, aby funkcja kryterium osiągnęła maksimum. W ten sposób uzyskana zostaje pierwsza zmienna służąca do sporządzenia rysunku na płaszczyźnie. Druga zmienna jest wyznaczana analogicznie, przy zachowaniu warunków przekształcenia ortogonalnego. Efektem opisanych działań jest obraz przedstawiający obserwacje wielowymiarowe, który można interpretować jako ich rzut na płaszczyznę. Ułożenie płaszczyzny rzutowania ustalone jest według wspomnianej funkcji kryterium. Celem wyznaczonego kryterium jest zapewnienie jak najlepszego odwzorowania wielowymiarowych obserwacji, gdzie przez „dobroć" rozumie się uwidocznienie ewentualnej niejednorodności zbiorów przez wystąpienie „prześwitu" pomiędzy zwartymi grupami obiektów. Sytuacja taka przedstawiona została na rys. 3, gdzie widoczne są wyraźne dwa skupiska. W proponowanej metodzie głównym przedmiotem poszukiwań jest właśnie prześwit zaprezentowany na rys. 3.

W celu zilustrowania działania metody przyjmijmy, że badane obiekty są obserwacjami dwuwymiarowymi, zatem przedstawienie ich na płaszczyźnie nie stanowi problemu. Załóżmy przy tym, że nasza percepcja ograniczona jest jedynie do przestrzeni jednowymiarowej. W związku z tym graficzne przedstawienie zbioru obiektów wymaga dokonania redukcji wymiaru przestrzeni. Najprościej jest w tym przypadku posłużyć się prostopadłym rzutem obiektów na pierwszą z osi, tj. X1, co prezentuje rys. 4. Analizując obraz rzutu oryginalnych obserwacji, nie sposób jest wykryć istniejącą niejednorodność danych, bowiem jednowymiarowy obraz, który otrzymano, nie wskazuje na istnienie prześwitu. Konieczne jest zatem dokonanie kolejnego rzutu w nadziei, że uwidoczniony zostanie prześwit świadczący o niejednorodności populacji.



Rys. 3. Prezentacja obiektów wielowymiarowych na płaszczyźnie z widocznymi dwoma skupiskami obiektów

Źródło: opracowanie własne.



Rys. 4. Rotacja i rzut prostopadły obserwacji dwuwymiarowych

Źródło: opracowanie własne.

Rysunek 4 przedstawia możliwy obrót obiektów wokół początku układu współrzędnych. Można zauważyć, że rzut obserwacji po ich rotacji uwidocznił prześwit istniejący pomiędzy skupiskami obiektów. Analogiczny rzut dokonany na podstawie oryginalnych obserwacji spowodowałby, że obiekty obu skupisk „nachodziłyby" na siebie. Analizując uzyskany jednowymiarowy obraz (oś X1), należy stwierdzić, że zaobserwowany prześwit między obserwacjami może być wynikiem występowania dwóch populacji. Zatem na podstawie wzrokowej analizy wykresu wykryta została niejednorodność zbioru danych.

W prezentowanej metodzie ilorazu odległości pożądany obrót ortogonalny jest wyznaczany drogą maksymalizacji funkcji kryterium. Dla przejrzystości wywodu prezentowany wcześniej przykład obejmował przypadek redukcji wymiaru przestrzeni z R2 do R1. Dla celów praktycznych w przypadku danych wielowymiarowych korzystnie jest posłużyć się redukcją wymiaru przestrzeni do R2. Wówczas wciąż możliwa jest wzrokowa ocena uzyskanego obrazu, a strata informacji powstała w wyniku redukcji wymiaru jest mniejsza.

Pierwszą zmienną niezbędną do prezentacji danych na płaszczyźnie wyznacza się według wzoru:

$$Y_1 = X\boldsymbol{a_1} = X_1 a_{11} + ... + X_m a_{1m}, \tag{8}$$

gdzie $\boldsymbol{a_1}^T = [a_{11}, ..., a_{1m}]$ jest wektorem spełniającym warunek $a_1^T a_1 = 1$.

Druga zmienna również jest kombinacją liniową wektorów kolumnowych macierzy *X*, co przedstawia równanie:

$$Y_2 = X\boldsymbol{a_2} = X_1 a_{21} + ... + X_m a_{2m}, \tag{9}$$

gdzie $\boldsymbol{a_2}^T = [a_{21}, ..., a_{2m}]$ spełnia warunek $\boldsymbol{a_2^T a_2} = 1$ oraz $\boldsymbol{a_1}^T \boldsymbol{a_2} = 0$.

Szukany wektor $\boldsymbol{a_1}$ stanowiący element macierzy przekształcenia wyznacza się w sposób iteracyjny w drodze maksymalizacji następującego wyrażenia:

$$P_1 = \frac{I^T \times \left| X\boldsymbol{a_1} - \bar{X}\boldsymbol{a_1} \right|}{\sqrt{(X\boldsymbol{a_1} - \bar{X}\boldsymbol{a_1})^T (X\boldsymbol{a_1} - \bar{X}\boldsymbol{a_1})}}, \tag{10}$$

gdzie $\boldsymbol{I}$ jest wektorem wymiaru $n \times 1$ składającym się z jedynek; $\bar{X}$ jest macierzą wymiaru $n \times m$, której wierszami są wektory średnich dla wszystkich cech (wszystkie wiersze są identyczne), $\left|X\right|$ oznacza wartość bezwzględną z wszystkich elementów macierzy *X*.

Kolejną zmienną wyznacza się analogicznie z zachowaniem warunku ortogonalnego obrotu obiektów. Warto zwrócić przy tym uwagę na fakt, iż wartości P_i zawsze będą większe od jedności oraz mniejsze niż $\sqrt{n}$ (*n* – liczba obserwacji macierzy *X*), co wynika z geometrycznej interpretacji uzyskanych wektorów Y_i.

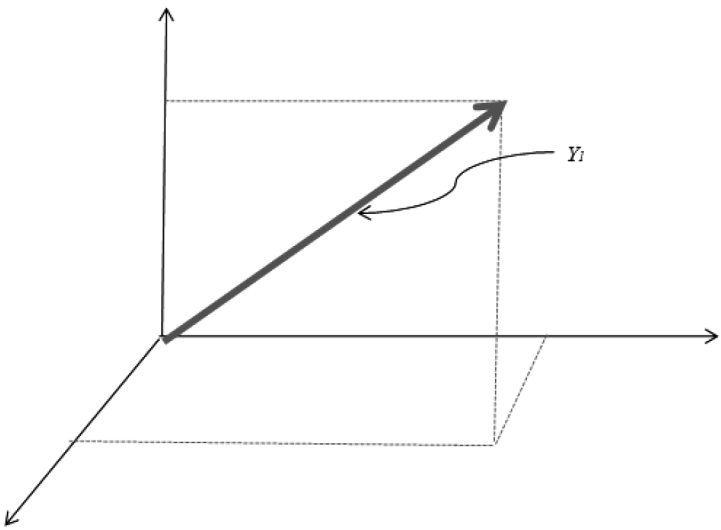
Dobroć przekształcenia można określić z wykorzystaniem wzoru:

$$D = \frac{P_1 + P_2}{Z_1 + Z_2 + ... + Z_m}, \tag{11}$$

gdzie P_1, P_2 obliczane są według równania (10) z zachowaniem odpowiedniej indeksacji, natomiast Z_i wynosi:

$$Z_i = \frac{I^T \left| X_i - \overline{X}_i \right|}{\sqrt{(X_i - \overline{X}_i)^T (X_i - \overline{X}_i)}},$$

gdzie ***I*** jest wektorem wymiaru $n \times 1$ składającym się z jedynek; X_i jest i-tą kolumną macierzy X; $\overline{X}_i$ jest wektorem wymiaru $n \times 1$ składającym się ze średnich obliczonych dla zmiennej X_i.



Rys. 5. Graficzna prezentacja cech Y_1 dla przypadku 3 obserwacji

Źródło: opracowanie własne.

W celu interpretacji maksymalizowanego kryterium określonego wzorem (10) można rozpatrywać je jako stosunek odległości obiektu będącego nową przekształconą (ukrytą) cechą ($X\boldsymbol{a}_1$) od obiektu reprezentowanego przez wektor wartości przeciętnych ($\overline{X}\boldsymbol{a}_1$) w przestrzeni n-wymiarowej. Przy czym odległość w liczniku wyrażona jest w metryce miejskiej, natomiast odległość w mianowniku jest odległością euklidesową. Odległości wyznaczone w różnych metrykach są ze sobą powiązane, jakkolwiek ich iloraz nie jest stały i zależy od macierzy przekształcenia ***a***. Uprosz-

czoną graficzną prezentację pierwszej przekształconej zmiennej (Y_1 – po odjęciu średniej wartości) przedstawia rys. 5. Długość otrzymanego wektora można interpretować jako odchylenie standardowe nowej uzyskanej cechy będącej wynikiem redukcji wymiaru przestrzeni. Suma wartości bezwzględnych współrzędnych analizowanego wektora jest natomiast odległością wyrażoną w metryce miejskiej.

Kryterium określone wzorem (10) wykorzystuje iloraz długości prezentowanego na rys. 5 wektora wyrażonych odpowiednio w metryce miejskiej oraz metryce euklidesowej. Metoda ilorazu odległości zakłada, że przekształcenie jest tym lepsze, im większy jest uzyskany iloraz. Warto zwrócić uwagę na fakt, iż w wyniku przekształcenia powodującego wzrost wariancji cechy Y_1 (na rys. 5 odpowiada to zwiększeniu kwadratu długości prezentowanego wektora) wartość rozważanego ilorazu (wzór (10)) nie ulega zmianie. Właściwość ta jest niezmiernie istotna z punktu widzenia praktycznych zastosowań metody. Zadane kryterium jest bowiem niewrażliwe na przekształcenia „faworyzujące" wysoką wariancję zmiennych wejściowych tak, jak ma to miejsce w przypadku metody głównych składowych. Stąd oryginalne cechy nie muszą być poddawane standaryzacji będącej ingerencją w strukturę danych.

Z punktu widzenia badania jednorodności zbiorów obserwacji korzystne jest, gdy stosunek we wzorze (10) pozostaje jak największy. Rozważmy bowiem przypadek, w którym uzyskujemy dwa przekształcenia, dla których odchylenia standardowe zmiennej Y_1 (długość wektora na rys. 5) są równe lub ich wartości są bardzo zbliżone. Wówczas przekształceniem lepszym wydaje się to, dla którego odległość miejska jest większa. Dzieje się tak, gdyż odległość miejska jest sumą wartości bezwzględnych poszczególnych współrzędnych (powstałych po odjęciu wektora średnich). Maksymalizowane kryterium (10) faworyzuje zatem takie przekształcenia, w ramach których odległości obiektów w przestrzeni m-wymiarowej od wyznaczonej obserwacji średniej pozostają jak najbardziej do siebie zbliżone. Innymi słowy algorytm dba o to, aby wszystkie obserwacje były położone „jednakowo" daleko do średniej. W przypadku dwóch odrębnych skupisk „punkt średni" znajdować się będzie w miejscu szukanego prześwitu. Zabieg taki prowadzi zatem wprost do zaobserwowania ewentualnych prześwitów między dwiema niejednorodnymi grupami, powodując symetryczne rozłożenie obiektów względem średniej.

6. Wyniki symulacji

W celu zweryfikowania skuteczności zaproponowanej metody przeprowadzonych zostało wiele symulacji polegających na wygenerowaniu obserwacji z wykorzystaniem dwuwymiarowego rozkładu normalnego. Jednocześnie zaprezentowano wyniki uzyskane w wyniku zastosowania metody głównych składowych, stąd możliwe było porównanie rezultatów otrzymanych w obu podejściach.

Każda z symulacji odnosi się do populacji osadzonych w przestrzeni dwuwymiarowej, które są liniowo separowalne. Punkty o różnych kolorach należą do odrębnych populacji, jakkolwiek obydwie z prezentowanych metod, tj. metoda ilorazu

odległości oraz głównych składowych, w żaden sposób nie korzystają z informacji określających przynależność obiektów do którejkolwiek populacji. Ich zadaniem jest redukcja wymiaru przestrzeni w celu uzyskania obrazu obiektów przedstawiających „prześwit" pomiędzy zbiorami punktów świadczący o niejednorodności badanych obserwacji. W prezentowanych symulacjach redukcja następuje do przestrzeni jednowymiarowej, zatem interpretacja uzyskanych wyników polega na analizie obiektów z perspektywy osi poziomej, tj. X1. Metoda graficznej prezentacji obiektów jest tym lepsza, im ich rzut na oś X1 ujawni większy „prześwit". Innymi słowy, konieczne jest dokonanie takiej rotacji punktów względem początku układu współrzędnych, która uwypukli brak jednorodności.

Łącznie przeprowadzono trzy grupy symulacji odzwierciedlających różnorodne położenie obu populacji względem siebie. Pierwsza z nich zaprezentowana została na rys. 8, gdzie wygenerowane obiekty można zaobserwować na rysunkach *a, b* oraz *c*. Parametry rozkładu normalnego wykorzystanego w procesie generowania obserwacji, gdzie dla uproszczenia przyjęto, że zmienne wyjściowe są niezależne, przedstawione zostały w tab. 1. Różnica pomiędzy wymienionymi wariantami polega na równoległym przesunięciu obiektów oznaczonych kolorem czarnym, co skutkuje uzyskaniem większego „prześwitu". Podejście takie umożliwia ocenę wpływu utworzonego „prześwitu" na jakość uzyskanych wyników w obrębie obu metod.

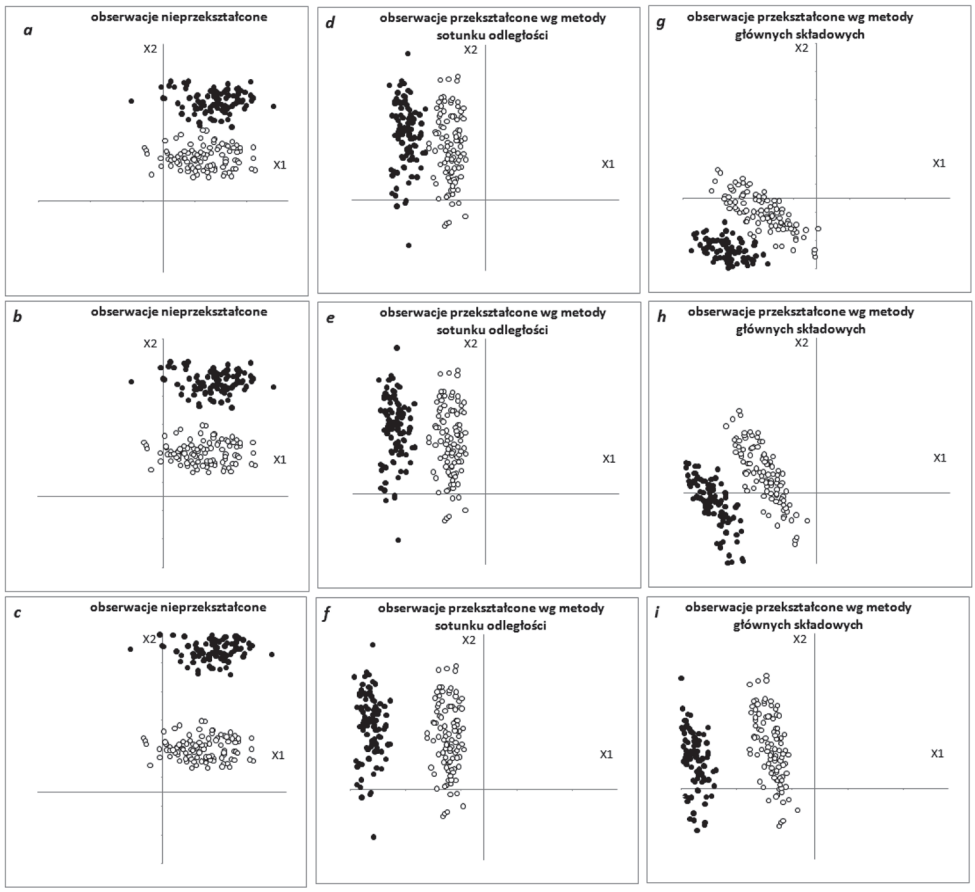
Tabela 1. Parametry populacji generowanych w ramach symulacji – wariant I

	wariant *a*				wariant *b*				wariant *c*			
	Populacja 1		Populacja 2		Populacja 1		Populacja 2		Populacja 1		Populacja 2	
	X1	X2	X1	X2	X1	X2	X1	X2	X1	X2	X1	X2
wartość przeciętna	4,0	3,0	5,0	7,0	4,0	3,0	5,0	8,0	4,0	3,0	5,0	10,0
odchylenie standardowe	2,5	0,8	2,5	0,8	2,5	0,8	2,5	0,8	2,5	0,8	2,5	0,8

Źródło: opracowanie własne.

Na rysunku 6 wykresy oznaczone jako *d, e* oraz *f* obrazują wyniki uzyskane z wykorzystaniem metody ilorazu odległości, podczas gdy g, *h* oraz *i* przedstawiają wyniki metody głównych składowych. Zatem interpretacja rezultatów otrzymanych dla populacji przedstawionych na wykresie *a* odnosi się do analizy wyników znajdujących się w tym samym wierszu na rysunku, czyli wykresów *d* oraz *g*.

Na rysunku 6 można zauważyć, że metoda ilorazu odległości zobrazowana przez rotację obiektów zaprezentowaną na wykresach *d, e* oraz *f* wyróżnia się niezmiernie wysoką skutecznością w prezentacji istniejącego „prześwitu". Uzyskane w wyniku maksymalizacji zaproponowanego kryterium obrazy, niezależnie od zadanej w symulacji odległości pomiędzy populacjami, w każdym przypadku uwidoczniły szukany „prześwit". Wyniki uzyskane w metodzie głównych składowych przedstawiają się odmiennie. W przypadku *a*, gdy obie populacje znajdują się stosunkowo blisko siebie, metoda głównych składowych nie poradziła sobie z zadaniem wykrycia „prześwitu", co prezentuje wykres *g*. Stopniowe oddalanie populacji od siebie sprzy-



Rys. 6. Wyniki analizy symulacyjnej w wariancie I

Źródło: opracowanie własne.

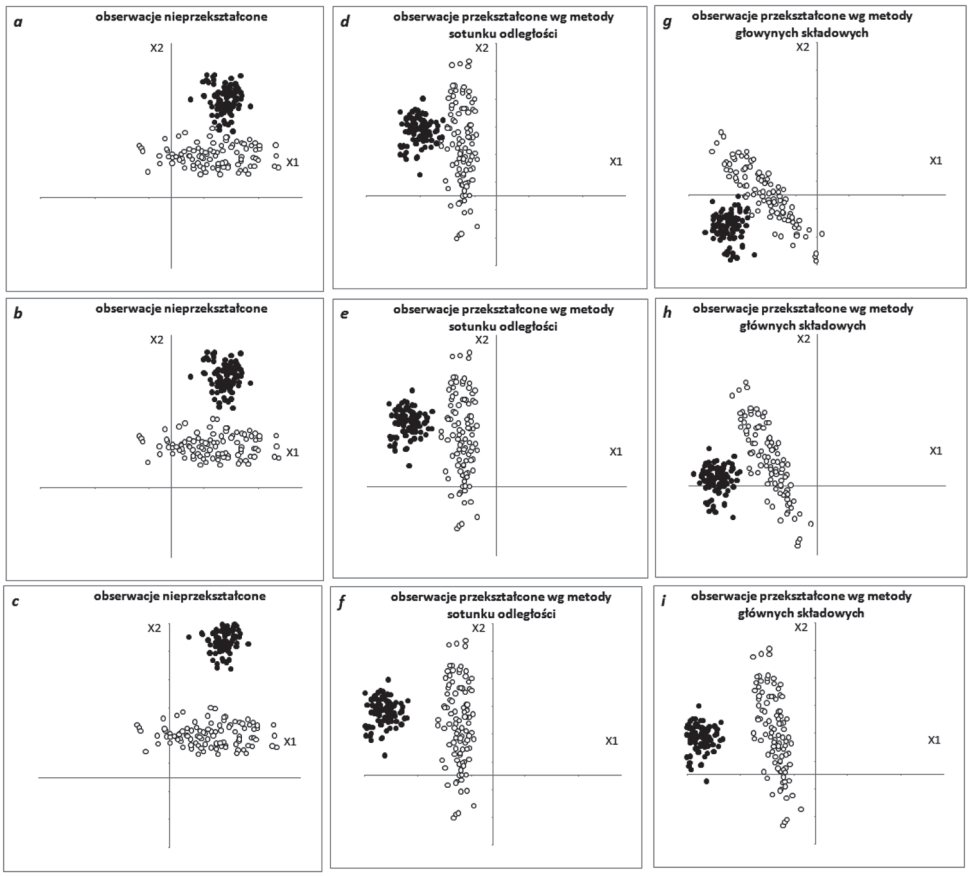
ja jednak uzyskaniu lepszych rezultatów. Na wykresie *h* widzimy, że wprawdzie „prześwit" uzyskany na osi X1 nie jest jeszcze widoczny, jednak przeprowadzona rotacja jest lepsza niż poprzednia. Na wykresie *i* z perspektywy osi X1 można natomiast zauważyć wyraźny prześwit, jakkolwiek nie jest on aż tak szeroki jak w przypadku wyników uzyskanych we wszystkich wariantach metody ilorazu odległości.

Drugi wariant symulacji również obejmuje dwie populacje, z czego ta oznaczona kolorem czarnym różni się wartością wariancji jednej z cech. Wygenerowane według tego schematu obserwacje przestawione zostały na rys. 9 na wykresach *a, b* oraz *c*. Analogicznie jak w przypadku wcześniejszym różnice między nimi wynikają z wielkości „prześwitu", co można zauważyć na podstawie danych zawartych w tab. 2.

Tabela 2. Parametry populacji generowanych w ramach symulacji – wariant II

	wariant *a*				wariant *b*				wariant *c*			
	Populacja 1		Populacja 2		Populacja 1		Populacja 2		Populacja 1		Populacja 2	
	X1	X2	X1	X2	X1	X2	X1	X2	X1	X2	X1	X2
wartość przeciętna	4,0	3,0	5,0	7,0	4,0	3,0	5,0	8,0	4,0	3,0	5,0	10,0
odchylenie standardowe	3,0	0,8	1,0	1,0	3,0	0,8	1,0	1,0	3,0	0,8	1,0	1,0

Źródło: opracowanie własne.



Rys. 7. Wyniki analizy symulacyjnej w wariancie II

Źródło: opracowanie własne.

Podobnie jak w przypadku wyników symulacji w wariancie I również tu metoda ilorazu odległości okazała się niezmiernie skuteczna, przewyższając wyniki uzyskane metodą głównych składowych. Ta ostatnia okazała się skuteczna dla danych wygenerowanych zgodnie ze scenariuszem *c,* co zostało zobrazowane na wykresie *i.*

W przypadku tym uzyskany „prześwit" jest widoczny, czego nie można stwierdzić na podstawie wyników zobrazowanych wykresami *g* oraz *h*.

Ostatni z wariantów przeprowadzonej symulacji przedstawiony został na rys. 8. W tym przypadku wygenerowano obserwacje w taki sposób, aby w obydwu populacjach wariancje drugiej zmiennej X2 były znacznie większe aniżeli zmiennej X1. Przypadek ten jest niezmiernie interesujący, ponieważ jest odwrotnością scenariusza I. Wcześniej zaobserwowanie prześwitu wymagało takiej rotacji obiektów w przestrzeni, aby wariancja szacowana łącznie dla wszystkich obserwacji zbliżała się do jej minimum. To tłumaczy w znacznej mierze, dlaczego metoda głównych składowych okazała się wówczas nieskuteczna. W istocie faworyzuje ona bowiem taką rotację, która zapewni maksymalną wariancję pierwszej składowej. Dlatego też przypadek III umożliwia zweryfikowanie hipotezy świadczącej o tym, iż być może metoda ilorazu odległości jest skuteczna jedynie wobec specyficznych zbiorów obserwacji.

Na rysunku 8 wykresy *a, b* oraz *c* przedstawiają wyjściowe warianty obejmujące wygenerowane obserwacje zgodnie z parametrami rozkładu normalnego zawartymi w tab. 3.

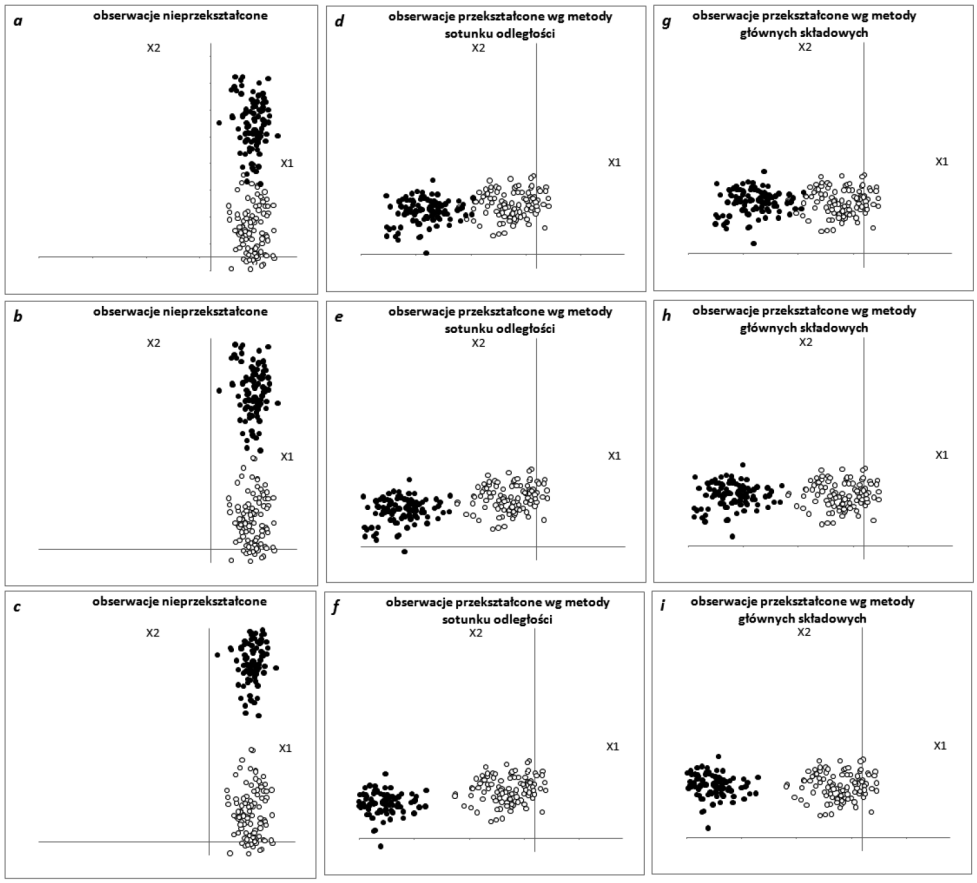
Tabela 3. Parametry populacji generowanych w ramach symulacji – wariant III

	wariant *a*				wariant *b*				wariant *c*			
	Populacja 1		Populacja 2		Populacja 1		Populacja 2		Populacja 1		Populacja 2	
	X1	X2	X1	X2	X1	X2	X1	X2	X1	X2	X1	X2
wartość przeciętna	4,0	2,0	4,0	10,0	4,0	2,0	4,0	12,0	4,0	2,0	4,0	14,0
odchylenie standardowe	1,0	2,0	1,0	2,0	1,0	2,0	1,0	2,0	1,0	2,0	1,0	2,0

Źródło: opracowanie własne.

Uzyskane wyniki w ramach symulacji III (rys. 8) wskazują, że rotacje w ramach metody głównych składowych umożliwiają zaobserwowanie pożądanego „prześwitu" analizowanego z punktu widzenia jednego wymiaru, tj. osi X1. Również metoda ilorazu odległości umożliwiła uzyskanie wyników sprzyjających zaobserwowaniu niejednorodności populacji. Obie metody w tej symulacji wykazały się zatem wysoką skutecznością niezależnie od zadanej odległości pomiędzy populacjami.

W ramach wariantu III analiza wzrokowa nie umożliwia jednoznacznej oceny co do wyższości którejkolwiek z metod, jednak uzyskane wyniki w obu przypadkach są satysfakcjonujące. Potwierdza to tym samym tezę, iż metoda ilorazów odległości jest skuteczna również w przypadkach wymagających takiej rotacji obiektów, która zapewnia maksymalną wartość wariancji szacowanej w ramach zredukowanej przestrzeni jednowymiarowej. Stąd kolejny raz zaproponowane kryterium w ramach metody ilorazu odległości potwierdziło swoją skuteczność.



Rys. 8. Wyniki analizy symulacyjnej w wariancie III

Źródło: opracowanie własne.

7. Podsumowanie

Zagadnienie występowania zjawiska jednorodności populacji warunkuje możliwość wykorzystania większości metod dyskryminacyjnych stosowanych w bankowości, obronności czy medycynie. Metody graficzne, które umożliwiają rzutowanie obserwacji wielowymiarowych na płaszczyznę bez konieczności ingerencji w ich strukturę, stanowią interesujące podejście. Główną ich zaletą jest możliwość prostej oraz intuicyjnej interpretacji uzyskanych wyników. Podkreślić przy tym należy, że analizie poddawane są nieprzekształcone w wyniku standaryzacji lub innych zabiegów normalizujących zmienne. Zatem podejście to polega na wykonaniu „fotografii" obiektów wielowymiarowych celem zaobserwowania braku jednorodności populacji. Utrata części informacji w wyniku rzutowania obiektów na przestrzeń o mniej-

szym wymiarze jest nieuchronna, jakkolwiek nie wyklucza możliwości wykrycia szukanych „prześwitów".

Uzyskane w przeprowadzonych symulacjach wyniki wskazują, że zaproponowana metoda ilorazu odległości jest skuteczna. Analiza porównawcza wykazała ponadto, że podejście wykorzystujące metodę głównych składowych w wielu przypadkach okazuje się znacznie gorsze. Zaprezentowana w świetle otrzymanych wyników metoda ilorazu odległości może być użytecznym narzędziem w celu analizy jednorodności danych wielowymiarowych. Maksymalizowana w ramach niniejszej metody funkcja okazała się niewrażliwa na występujące różnice w wariancjach cech, co umożliwia analizę oryginalnych danych. Podobnie jak w przypadku innych graficznych metod tego typu brak jest jednak obiektywnego kryterium stwierdzającego niejednorodność. Jakkolwiek wzrokowa analiza wydaje się w tym przypadku wystarczająca.

Literatura

Alpern B., Carter L., *Hyperbox*, Proceedings of the 2nd IEEE Conference on Visualization '91, 1991.

Andrews D., *Plots of high-dimensional data*, "International Biometric Society" 1972, no 18(1).

Chernoff H., *The use of faces to represent points in k-dimensional space graphically*, "Journal of American Statistic Association" 1973, no 68.

Fao R., Card S.K. *The Table Lens: Merging Graphical and Symbolic Representations in an Interactive Focus + Context Visualization for Tabular Information*, Proceedings of the SIGCHI Conference on Human Factors in Computer Systems: Celebrating Interdependence, 1994.

Gabriel R.K., *The biplot graphic display of matrices with application to principal component analysis*, "Biometrika"1971, no 58.

Hoffman P.E., *Table Visualizations: A Formal Model and Its Applications, Doctoral Dissertation*, Computer Science Department, University of Massachusetts at Lowell, 1999.

Hotelling H., *Analysis of a complex of statistical variables into principal components*, "Journal of Educational Psychology" 1933, no 24.

Inselberg A., *The plane with parallel coordinates*, "Visual Computer" 1985, no 1(4).

Jajuga K., *Statystyczna teoria rozpoznawania obrazów*, PWN, Warszawa 1990.

Keim D.A., Driegel H.-P., Ankerst M., *Recursive Pattern: A Technique for Visualizing Very Large Amounts of Data*, Proceedings of the 6th IEEE Conference on Visualization '95, 1995.

Keim D.A., Kriegel H.-P., *VisDB: database exploration using multidimensional visualization*, "IEEE Transactions on Computer Graphics and Applications" 1994, vol. 14, no 5.

LeBlanc J., Ward M.O., Wittels N., *Exploring N-Dimensional Databases*, Proceedings of the 1st IEEE Conference on Visualization '90, 1990.

Moustafa R., Wegman E., *On Some Generalizations of Parallel Coordinate Plots. Seeing a million, A Data Visualization Workshop*, Rain am Lech, Germany 2002.

Pearson K., *On lines and planes of closest fit to systems of points in space*, "Philosophical Magazine" 1901, no 2.

Von Mayr G., *Die Gesetzmäßigkeit im Gesellschaftsleben Oldenbourg*, München 1877.

DISTANCES RATIO METHOD − THE ISSUE OF GRAPHICAL PRESENTATION OF THE MULTIDIMENSIONAL OBSERVATION

Summary: The article presents the author's method of multidimensional data analysis. Its concept is based on the principle of reduction of the space dimension by projecting the observation on two-dimensional plane. The resulting image is the basis for the further visual analysis of the observation. The main objective of the presented method is to rotate the observations in such a way that the resulting image has highlighted the possible heterogeneity of the population. The proposed approach is presented against the background of the method of principal components. The results of the simulation showed that the author's method is particularly effective in the detection of heterogeneous populations. An important advantage is no need for the prior standardization of variables, which provides a criterion function used for projection of observation.

Keywords: credit risk, pattern recognition, classification.