Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie

Wydział Zarządzania

Statystyczna analiza danych

Porównanie różnych metod skalowania wielowymiarowego

Mateusz Grzelik

14 grudnia 2024

1 Wstęp

W pracy postaramy się opisać oraz zilustrować działanie trzech metod skalowania wielowymiarowego:

- Klasyczne skalowanie wielowymiarowe - Jest to metoda oparta na analizie wartości własnych macierzy odległości. Celem klasycznego skalowania jest odwzorowanie danych w przestrzeni o mniejszej liczbie wymiarów, w taki sposób, aby możliwie dokładnie zachować odległości euklidesowe między punktami.
- Skalowanie niemetryczne Kruskala - stanowi rozwinięcie metody klasycznej, skupiające się na zachowaniu monotoniczności relacji między odległościami. W przeciwieństwie do klasycznego MDS, metoda ta nie jest liniowa.
- Skalowanie Sammona - koncentruje się na minimalizacji różnic między odległościami w przestrzeni wysokowymiarowej i ich odwzorowaniami w przestrzeni niskowymiarowej. W tej metodzie szczególną uwagę zwraca się na dokładne odwzorowanie mniejszych odległości.

2 Teoretyczna analiza

2.1 Klasyczne skalowanie wielowymiarowe

Idea klasycznego skalowania wielowymiarowego opiera się na następującym twierdzeniu dowiedzionym przez Younga i Householdera [1]:

Niech będzie dana macierz $D = [d_{ij}]_{n\times n}$, taka, że $\forall_{i,j} : d_{ij} = d_{ji}$ oraz macierz $B = HAH$, gdzie $H = I_n - \frac{1\cdot 1^T}{n}, A = -\frac{1}{2}d_{ij}^2$. Wówczas na to, by macierz D była macierzą odległości punktów w przestrzeni euklidesowej potrzeba i wystarcza, by macierz B była dodatnio półokreślona (wszystkie elementy większe lub równe zero).

Twierdzenie to jest uogólnieniem nierówności trójkąta. Algorytm wykorzystywany w funkcji `cmdscale()` w R przebiega następująco [2]:

1. Wyznaczamy wyżej wspomnianą macierz $B = HAH$, gdzie $H = I_n - \frac{1\cdot 1^T}{n}, A = -\frac{1}{2}d_{ij}^2$ (I_n - macierz jednostkowa, 1 - wektor jedynek, d_{ij} - elementy macierzy odległości)
2. Wyznaczamy wartości własne $\lambda_1, ..., \lambda_p$ macierzy B oraz odpowiadające im wektory własne $v_1, ..., v_p$.
3. Macierz współrzędnych punktów X otrzymujemy korzystając ze wzoru $X = V\Lambda^{\frac{1}{2}}$, gdzie $V = (v_1, ..., v_p)$ - macierz wektorów własnych, Λ - macierz diagonalna, na przekątnej posiadająca wartości własne $\lambda_1, ..., \lambda_p$

Metoda ta ma znaczącą przewagę nad pozostałymi dwoma - nie jest iteracyjna, zatem czas jej wykonania jest nieporównywalnie krótszy, nawet dla bardzo dużych zestawów danych.

2.2 Skalowanie niemetryczne Kruskala

Metoda skalowania wielowymiarowego Kruskala opiera się na minimalizacji statystyki nazywanej STRESS danej następującym wzorem [3]:

$$S = \sqrt{\frac{\sum_{i<j}(d_{ij} - \hat{d}_{ij})^2}{\sum_{i<j} d_{ij}^2}}$$

gdzie d_{ij} jest odległością i-tego obiektu od j-tego obiektu w oryginalnej przestrzeni, a $\hat{d}_{ij}$ w przestrzeni t-wymiarowej. Skalując n punktów w wielowymiarowej przestrzeni do t-wymiarowej przestrzeni wartość S jest traktowana jako funkcja zmiennych $x_{11}, ..., x_{1t}, x_{n1}, ..., x_{nt}$, czyli współrzędnych punktów w przestrzeni t-wymiarowej. Zauważmy, że jest ich dokładnie $n \cdot t$. Funkcja ta jest minimalizowana z użyciem metody najszybszego spadku, polegającej na obliczeniu (ujemnego) gradientu S:

$$\nabla S = \left[-\frac{\partial S}{\partial x_{11}}, ..., -\frac{\partial S}{\partial x_{nt}}\right]$$

a następnie wykonaniu niewielkiej zmiany współrzędnych w kierunku, w którym wartość owego gradientu jest największa, aż do momentu gdy będzie on bliski zeru (oznacza to minimum lokalne).

W celu opisania procedury wprowadzimy następujące oznaczenia:

$$S^* = \sum (d_{ij} - \hat{d}_{ij})^2$$

$$T^* = \sum d_{ij}^2$$

$$S = \sqrt{\frac{S^*}{T^*}}$$

Działanie algorytmu przedstawia się następująco:

1. Losowo dobieramy wstępną konfigurację punktów w t-wymiarowej przestrzeni i obliczamy dla niej wartość S.
2. Obliczamy gradient korzystając ze wzoru (dla odległości euklidesowej):

$$g_{kl} = S \sum_{i,j} (\delta_{ki} - \delta_{kj}) \left[\frac{d_{ij} - \hat{d}_{ij}}{S^*} - \frac{d_{ij}}{T^*} \right] \frac{x_{il} - x_{jl}}{d_{ij}}$$

   gdzie δ to tzw. delta Kroneckera (przyjmuje wartość 1 dla równych indeksów, 0 w pozostałych przypadkach).
3. Jeżeli wartość gradientu jest satysfakcjonująco mała, algorytm kończy działanie. W przeciwnym razie wykonujemy niewielki krok w kierunku w którym gradient jest największy i dla nowych współrzędnych wracamy do punktu 2.

2.3 Skalowanie Sammona

Skalowanie Sammona opiera się na minimalizacji błędu danego wzorem [4]:

$$E = \frac{1}{\sum_{i<j}^{N} d_{ij}} \sum_{i<j}^{N} \frac{(d_{ij} - \hat{d}_{ij})^2}{d_{ij}}$$

gdzie d_{ij} jest odległością i-tego obiektu od j-tego obiektu w oryginalnej przestrzeni, a $\hat{d}_{ij}$ w przestrzeni t-wymiarowej. Niech m oznacza numer iteracji, a $x_{ij}(m)$ współrzędne punktów w przestrzeni t-wymiarowej w m-tej iteracji. Problem optymalizacyjny jest rozwiązywany metodą najszybszego spadku, podobnie jak w przypadku skalowania Kruskala. Jednak ze względu na różnice w postaci minimalizowanej funkcji procedura przedstawia się nieco inaczej:

1. Dobieramy wejściową konfigurację punktów ($x_{ij}(m)$) w przestrzeni t-wymiarowej, rzucając pełnowymiarowy obiekt na t współrzędnych o największych wariancjach.
2. Wyznaczamy $x_{pq}(m+1)$ dane wzorem:

$$x_{pq}(m+1) = x_{pq}(m) - (MF) \cdot \frac{\frac{\partial E}{\partial x_{pq}}}{\frac{\partial^2 E}{\partial^2 x_{pq}}}$$

   gdzie MF to parametr dobierany przez użytkownika. W cytowanym artykule J.W. Sammon ustalił go empirycznie na poziomie 0.3 lub 0.4. W funkcji `sammon()` w pakiecie MASS w R jego domyślna wartość wynosi 0.2. Pochodne cząstkowe są dane wzorami:

$$\frac{\partial E}{\partial x_{pq}} = \frac{-2}{\sum_{i<j}^{N} d_{ij}} \sum_{j=1, j\neq p}^{N} \left[\frac{d_{pj} - \hat{d}_{pj}}{d_{pj}\hat{d}_{pj}} \right] (x_{pq} - x_{jq})$$

$$\frac{\partial^2 E}{\partial x_{pq}^2} = \frac{-2}{\sum_{i<j}^{N} d_{ij}} \sum_{j=1, j\neq p}^{N} \frac{1}{\hat{d}_{pj} d_{pj}} \left[(d_{pj} - \hat{d}_{pj}) - \frac{(x_{pq} - x_{jq})^2}{d_{pj}} \left(1 + \frac{d_{pj} - \hat{d}_{pj}}{\hat{d}_{pj}} \right) \right]$$

   3. Jeżeli gradient ($\frac{\partial E}{\partial x_{pq}} : \frac{\partial^2 E}{\partial^2 x_{pq}}$) jest dla każdej współrzędnej bliski zera, to algorytm kończy działanie. W przeciwnym razie powrót do punktu 2.

3 Tesserakt

W tej sekcji podejmiemy próbę wizualizacji 4-wymiarowego obiektu będącego uogólnieniem trójwymiarowego sześcianu z użyciem skalowania wielowymiarowego. Podobnie, jak dwa kwadraty połączone czterema krawędziami przez trzeci wymiar tworzą sześcian, tak dwa sześciany połączone ośmioma krawędziami przez czwarty wymiar tworzą tesserakt. Te intuicyjną konstrukcję tesseraktu postaramy się uwypuklić na poniższych rysunkach.

Jeden z 16 wierzchołków tesseraktu osadzimy w punkcie $(0, 0, 0, 0)$ 4-wymiarowego układu współrzędnych, a wszystkie jego krawędzie będą zawierały się lub będą równoległe do jednej z osi. Tak skonstruowany tesserakt o boku długości 1 będzie miał wierzchołki w następujących punktach:

x	**y**	**z**	**w**
0	0	0	0
0	0	0	1
0	0	1	0
0	0	1	1
0	1	0	0
0	1	0	1
0	1	1	0
0	1	1	1
1	0	0	0
1	0	0	1
1	0	1	0
1	0	1	1
1	1	0	0
1	1	0	1
1	1	1	0
1	1	1	1

Tabela 1: Współrzędne wierzchołków tesseraktu. Można traktować je jako kolejne liczby od 0 do 15 zapisane w systemie binarnym.

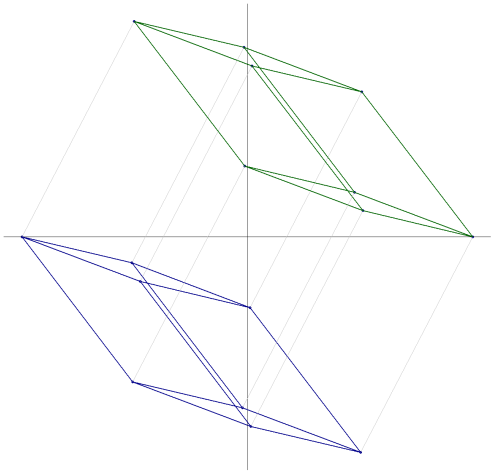
Zauważmy, że:

$$\forall_{i,j} : d_{ij} \in \{0, 1, \sqrt{2}, \sqrt{3}, 2\}$$

gdzie d_{ij} to odległość i-tego wierzchołka od j-tego wierzchołka tesseraktu. Zatem elementy powyższego zbioru będą jedynymi elementami macierzy odległości.

3.1 Klasyczne skalowanie

Dla lepszej interpretowalności rysunku poszczególne zrzutowane krawędzie tesseraktu zostały pokolorowane. Zielonym kolorem oznaczono wszystkie krawędzie, które w oryginalnej przestrzeni leżały na płaszczyźnie $x = 1$, na niebiesko leżące na płaszczyźnie $x = 0$, a szarym pozostałe.



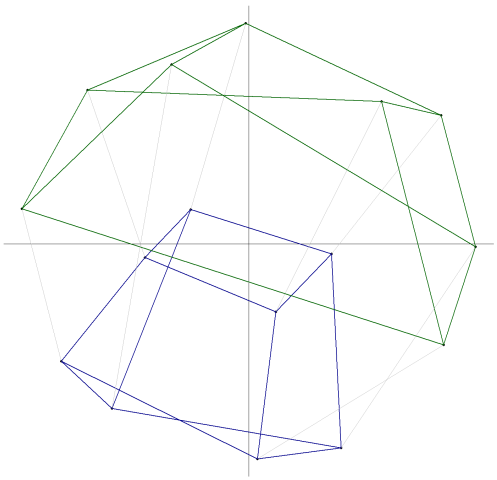
Rysunek 1: Tesserakt skalowany klasyczną metodą

Powyższy rysunek dobrze obrazuje omówioną powyżej intuicyjną konstrukcję tesseraktu, jako połączenie wierzchołków dwóch sześcianów (zielony i niebieski) ośmioma odcinkami poprzez czwarty wymiar (na szaro). Wartość statystyki STRESS wyniosła 41.24%. Sugeruje ona słabe dopasowanie rzutu [5].

Metoda skalowania Kruskala dobrze odwzorowuje kształt wielowymiarowej struktury (krawędzie oryginalnie równoległe pozostały równoległe, bo metoda jest liniowa), może jednak zaburzyć stosunki odległości pomiędzy punktami - w oryginalnej przestrzeni każda z widocznych na rysunku krawędzi miała długość 1, po rzucie widać znaczne dysproporcje.

3.2 Skalowanie niemetryczne Kruskala

Na rysunku zachowano konwencję kolorystyczną z powyższego punktu.

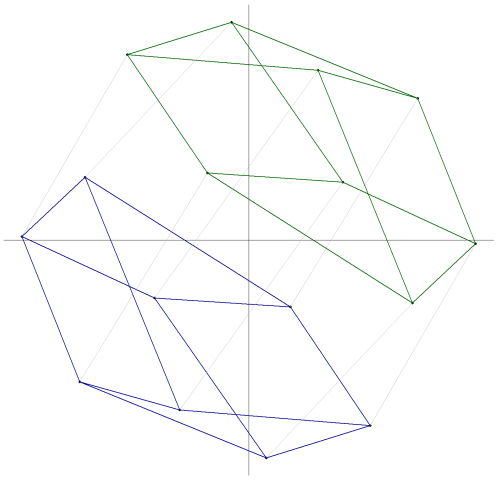


Rysunek 2: Tesserakt skalowany metodą Kruskala

Wartość statystyki STRESS wyniosła w tym przypadku 26.36%. Wartość ta sugeruje słabe dopasowanie skalowania. Na ilustracji widać znaczne dysproporcje odległości, również sam kształt jest mocno zaburzony. Metoda Kruskala skupia się na zachowaniu kolejności odległości - ponieważ w przypadku tesseraktu jest dużo parami równych odległości, algorytm słabo radzi sobie z ich odwzorowaniem.

3.3 Skalowanie Sammona

Na rysunku zachowano konwencję kolorystyczną z powyższych punktów.



Rysunek 3: Tesserakt skalowany metodą Sammona

W przypadku metody Sammona STRESS wyniósł około 9.76%. Wartość ta sugeruje dobre dopasowanie rzutu [4]. Również z rysunku widać, że zaznaczone krawędzie o oryginalnej długości 1 różnią się od siebie mniej niż w

przypadku skalowania Kruskala. Zachowana (choć lekko zniekształcona) jest także intuicyjna konstrukcja tesseraktu - widoczne są dwa "sześciany"połączone ośmioma krawędziami. Zwróćmy uwagę, że nie jest zachowana równoległość odpowiednich krawędzi - metoda ta jest nieliniowa.

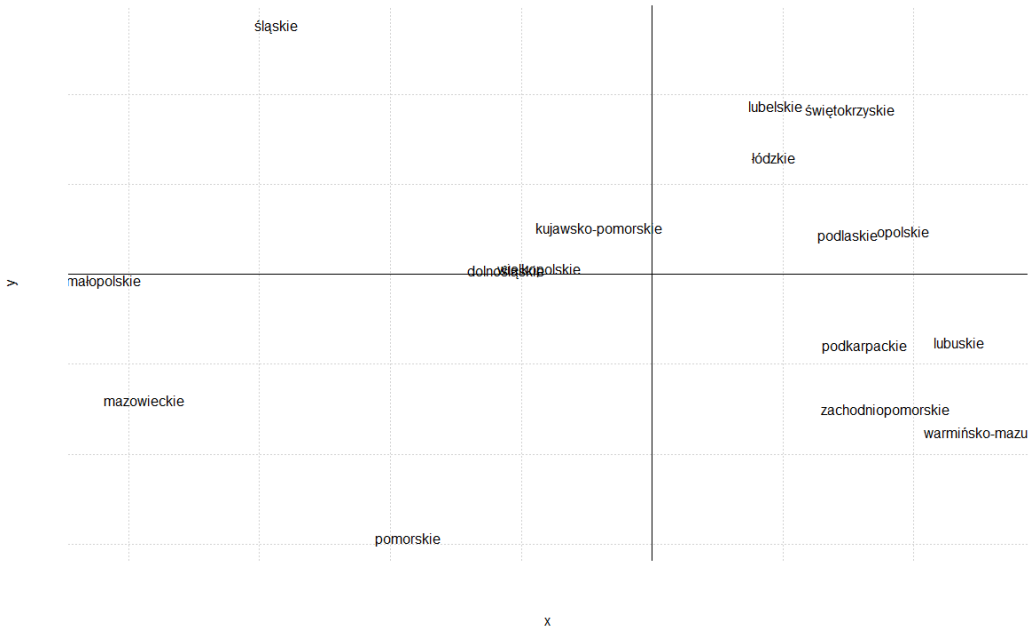
4 Przykład praktyczny - polskie województwa

4.1 Dane

Szczegółowa analiza została przeprowadzona w projekcie pierwszym, w celu interpretacji skalowania przypomnijmy jedynie ostateczny ranking z poprzedniego projektu:

1	śląskie
2	małopolskie
3	kujawsko-pomorskie
4	dolnośląskie
5	mazowieckie
6	wielkopolskie
7	pomorskie
8	opolskie
9	świętokrzyskie
10	lubelskie
11	podkarpackie
12	łódzkie
13	zachodniopomorskie
14	warmińsko-mazurskie
15	lubuskie
16	podlaskie

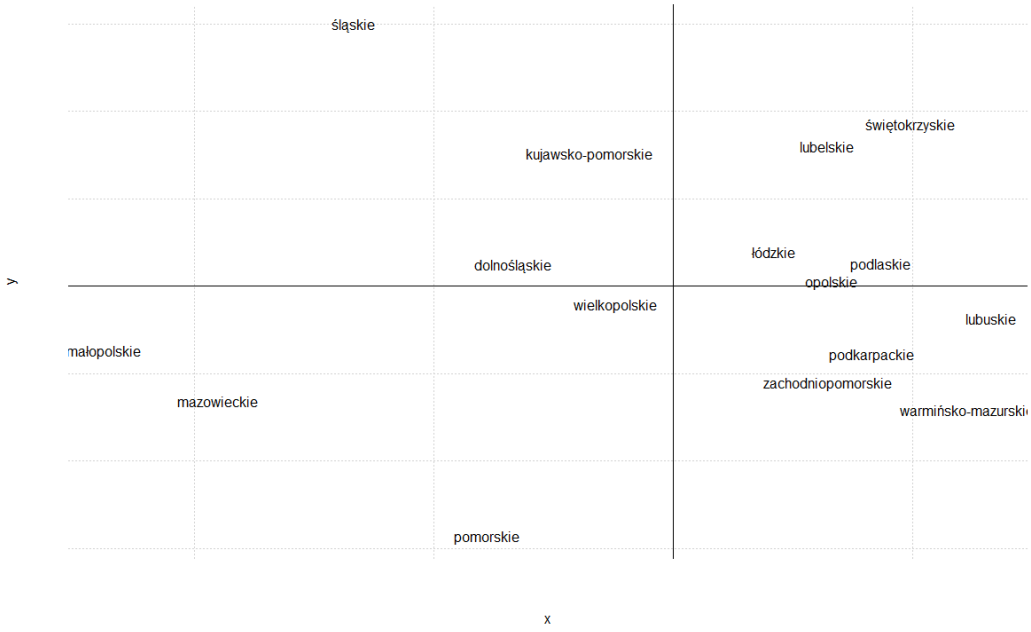
4.2 Skalowanie klasyczne



Rysunek 4: Wyniki skalowania metodą klasyczną

STRESS wyniósł 36.71%. Dopasowanie rzutu jest słabe, zatem rysunek jest słabo interpretowalny. Mimo to widać znacznie wyróżniające się województwo śląskie oraz małopolskie (są daleko od pozostałych, ale w róznych kierunkach - można zatem wnioskować, że cechy powodujące ich wysoką pozycję w rankingu są różne).

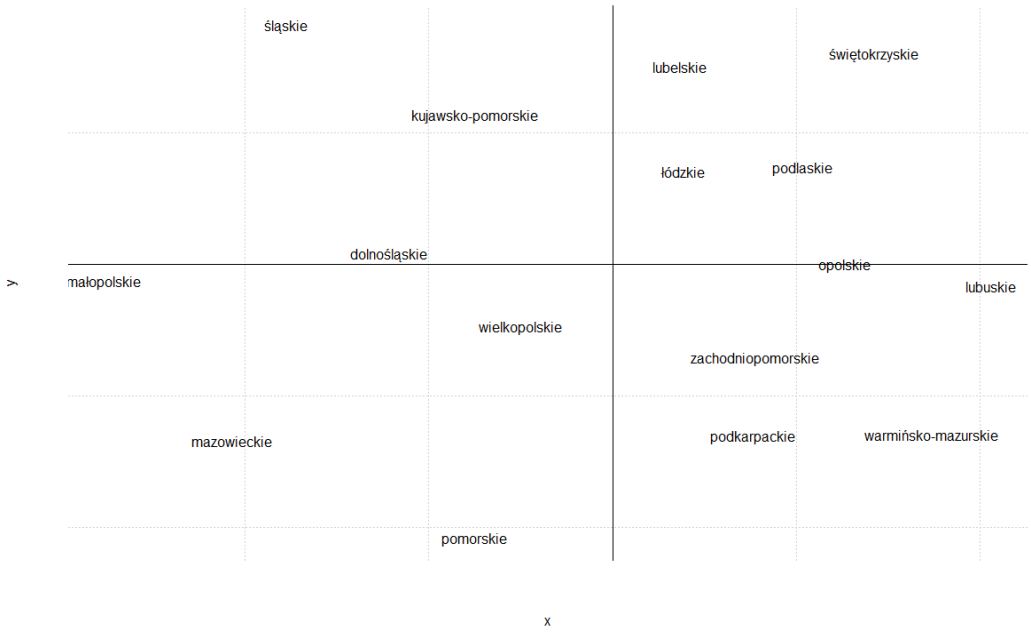
4.3 Skalowanie niemetryczne Kruskala



Rysunek 5: Wyniki skalowania metodą Kruskala

STRESS równy 12.56% sugeruje akceptowalne dopasowanie rzutu. Widoczne jest skupisko województw w pierwszej i czwartej ćwiartce układu współrzędnych, ponadto zbliżone są województwa małopolskie i mazowieckie. Pamiętajmy jednak, że metoda ta ma na celu odwzorowanie kolejności województw, co może prowadzić do znacznego zniekształcenia struktry (jak widzieliśmy na przykładzie tesseraktu). Rzeczywiste rozmieszczenie miast w przestrzeni może zatem wyglądać nieco inaczej.

4.4 Skalowanie Sammona



Rysunek 6: Wyniki skalowania metodą Sammona

STRESS wyniósł 5.87%. Wartość ta oznacza dobre dopasowanie rzutu. Rysunek można zatem z powodzeniem interpretować. W trzeciej ćwiartce układu widzimy województwa wysoko rozwinięte - małopolskie, mazowieckie, pomorskie oraz wielkopolskie. Znacząco wyróżnia się województwo śląskie, które w rankingu poziomu życia plasowało się wysoko, głównie ze względu na dostęp do rekreacji oraz obiektów sportowych. Na uwagę zasługuje także fakt, że rozkład miast jest asymetryczny względem pionowej osi (mimo że metoda Sammona rzutuje środek ciężkości wielowymiarowej struktury na środek układu). Województwa znajdujące się po prawej stronie charakteryzują się na ogół niższym poziomem życia (co pokazaliśmy w pierwszym projekcie), zatem można wnioskować że brak miast wysuniętych skrajnie na prawo jest pozytywną cechą.

5 Wnioski

W obydwu analizowanych przypadkach metoda Sammona najlepiej poradziła sobie z odwzorowaniem odległości pomiędzy obiektami. Należy jednak zauważyć, że metoda ta, ponieważ jest nieliniowa, nie odwzorowuje kształtu wielowymiarowej struktury. Problemu tego nie ma klasyczne skalowanie, jednak tutaj często zaburzeniu ulegają stosunki odległości. Mając do dyspozycji wyniki klasycznego skalowania wielowymiarowego oraz skalowania Sammona, obydwa o STRESS-ie w okolicach 5-10% (czyli dopasowanie dobre) za bardziej interpretowalny uznać można ten pierwszy.

Bibliografia

[1] Gale Young i A. S. Householder. "Discussion of a set of points in terms of their mutual distances". W: *Psychometrika* 3.1 (1938), s. 19–22. ISSN: 1860-0980. DOI: 10.1007/BF02287916. URL: https://doi.org/10.1007/BF02287916.

[2] K.V. Mardia. "Some properties of clasical multi-dimesional scaling". W: *Communications in Statistics - Theory and Methods* 7.13 (1978), s. 1233–1241. DOI: 10.1080/03610927808827707.

[3] J. B. Kruskal. "Nonmetric multidimensional scaling: A numerical method". W: *Psychometrika* 29.2 (1964), s. 115–129. ISSN: 1860-0980. DOI: 10.1007/BF02289694. URL: https://doi.org/10.1007/BF02289694.

[4] John W. Sammon. "A Nonlinear Mapping for Data Structure Analysis". W: *IEEE Transactions on Computers* C-18 (1969), s. 401–409. URL: https://api.semanticscholar.org/CorpusID:43151050.

[5] J. B. Kruskal. “Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis”. W: *Psychometrika* 29.1 (mar. 1964), s. 1–27. DOI: 10.1007/bf02289565.