Jest to zapis odczytu wygłoszonego na XXVI Szkole Matematyki Poglądowej: *Twierdzenia z pogranicza.*

Obrazy zależności

Andrzej DĄBROWSKI, Wrocław

Mapy zależności

Obraz wart jest tysiąc słów – często więc posługujemy się przedstawieniem tego, co chcemy powiedzieć, za pomocą obrazu, wykresu czy diagramu. *Możliwość zobrazowania graficznego zależności między jakimiś wielkościami, zobaczenia tej zależności, bardzo wzmacnia siłę intelektualną badającego taką zależność.* [1, str. 150] Gdy wyniki obserwacji, są wyrażone liczbami, możemy użyć wynalezionej przez Kartezjusza metody współrzędnych. (Pod warunkiem, że tych współrzędnych nie jest za dużo. Najlepiej, gdy są to dwie współrzędne.) Ale co zrobić, gdy nasze obserwacje dotyczą zjawisk wyrażonych w języku opisowym? Jedyne, co możemy zrobić w takiej sytuacji, to zapisać, jak często pojawiły się w naszych obserwacjach poszczególne kategorie obserwowanych zjawisk. Macierz, zawierającą takie informacje, nazywa się *tablicą kontyngencji.*

W grupie 100 osób, w której było 50 kobiet i 50 mężczyzn, przeprowadzono ankietę na temat ulubionego sposobu spędzenia czasu wolnego. Ankietowani mieli do wyboru jedną z trzech odpowiedzi A, B, C. Wyniki ankiety zamieszczone są w tablicy kontyngencji:

	A	B	C	Razem
kobiety	30	10	10	50
mężczyźni	10	40	0	50
Razem	40	50	10	100

Chcielibyśmy znaleźć taki sposób przedstawienia danych, aby móc uwidocznić związek, o ile taki istnieje, między płcią a sposobem spędzenia czasu wolnego. Najprostsze, co przychodzi do głowy, to przypisanie odpowiedziom pochodzącym od kobiet jakiejś wartości punktowej, podobnie odpowiedziom od mężczyzn, czy też przypisanie wartości punktowej każdej odpowiedzi typu A, B, C. Oczywiście, takie przypisanie musi być w jakimś sensie rozsądne. Przyjrzyjmy się skutkom wyboru skali punktowej dla tych danych.

Przypiszmy na początek odpowiedziom, dawanym przez kobiety, 2 punkty a przez mężczyzn 1 punkt. Jakie wartości punktowe powinniśmy przypisać wtedy odpowiedziom A, B i C? Odpowiedź A zgromadziłaby 70 punktów ($70 = 30 \cdot 2 + 10 \cdot 1$), B – 60 punktów, C – 20 punktów. Gdybyśmy więc każdej odpowiedzi A przypisali 1,75 punktu ($1{,}75 = 70/40$), odpowiedzi B – 1,2 punktu, a C – 2 punkty, to otrzymalibyśmy łączną liczbę punktów za każdy typ odpowiedzi, zgodną z wymyślonym przez nas systemem punktowania dla płci.

Przyjmijmy więc ten system punktowania odpowiedzi A, B, C za obowiązujący i zobaczmy, czy potwierdzi się wyjściowy system punktacji. Odpowiedzi kobiet zgromadziłyby 84,5 punktu ($84{,}5 = 30 \cdot 1{,}75 + 10 \cdot 1{,}2 + 10 \cdot 2$), a odpowiedzi mężczyzn 65,5 punktu ($65{,}5 = 10 \cdot 1{,}75 + 40 \cdot 1{,}2 + 0 \cdot 2$), co w przeliczeniu na jedną odpowiedź daje 1,69 punktu dla kobiet i 1,31 punktu dla mężczyzn. Z tego wynika, że nasza propozycja punktacji za odpowiedzi według płci nie była dobra, bo nie byliśmy w stanie uzgodnić jej z punktacją za odpowiedzi A, B, C.

Spróbujmy opisać zastosowaną procedurę w sposób ogólny. Niech x oznacza wektor (wektory będziemy zawsze pisać w kolumnach) punktów za odpowiedzi według cechy, opisanej w wierszach tablicy kontyngencji (w naszym przypadku jest to płeć) a y – wektor punktów za kategorie cechy, opisanej w kolumnach. Ocena punktowa dla cechy kolumnowej, generowana przez wektor x, jest wektorem, którego współrzędne można wyrazić wzorem $K^T x$, gdzie w j-tej kolumnie macierzy K występują *częstości* pojawiania się odpowiedzi z j-tej kolumny, symbol T oznacza transpozycję macierzy. Podobnie, ocena punktowa dla cechy wierszowej, generowana przez wektor y, jest wektorem, którego współrzędne można wyrazić wzorem Wy, gdzie w i-tym wierszu macierzy W występują częstości pojawiania się odpowiedzi z i-tego wiersza. Macierze K

Częstość zdarzenia jest to liczba pojawień się tego zdarzenia, podzielona przez liczbę wszystkich zdarzeń.

i W będziemy nazywać odpowiednio *profilami kolumnowymi* i *wierszowymi*. Z definicji profili wynika, że suma elementów w każdej kolumnie profilu kolumnowego wynosi 1, podobnie jak suma w każdym wierszu profilu wierszowego.

Dla naszego przykładu profile mają postać

$$K = \begin{bmatrix} 0{,}75 & 0{,}2 & 1 \\ 0{,}25 & 0{,}2 & 0 \end{bmatrix}, \qquad W = \begin{bmatrix} 0{,}6 & 0{,}2 & 0{,}2 \\ 0{,}2 & 0{,}8 & 0{,}0 \end{bmatrix}$$

a oceny punktowe, generowane przez wektory $x^T = [2 \quad 1]$ i $y^T = [1{,}75 \quad 1{,}2 \quad 2]$ mają postać

$$K^T x = \begin{bmatrix} 0{,}75 & 0{,}25 \\ 0{,}2 & 0{,}8 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 1{,}75 \\ 1{,}2 \\ 2 \end{bmatrix}, \qquad \begin{bmatrix} 0{,}6 & 0{,}2 & 0{,}2 \\ 0{,}2 & 0{,}8 & 0{,}0 \end{bmatrix} \begin{bmatrix} 1{,}75 \\ 1{,}2 \\ 2 \end{bmatrix} = \begin{bmatrix} 1{,}69 \\ 1{,}31 \end{bmatrix}$$

Warunek zgodności ocen, generowanych przez x i y tworzy układ równań z dwoma nieznanymi wektorami x i y: $y = K^T x$, $x = Wy$. Po podstawieniu y do drugiego wzoru otrzymamy równanie $x = WK^T x$, równoważne równaniu jednorodnemu $(I - WK^T)x = 0$, gdzie I oznacza macierz jednostkową 2×2. Macierz

$$I - WK^T = \begin{bmatrix} 0{,}31 & -0{,}31 \\ -0{,}31 & 0{,}31 \end{bmatrix} = 0{,}31 \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

jest kwadratowa, symetryczna i ma wyznacznik 0, a więc równanie zawierające niewiadomą x ma nieskończenie wiele rozwiązań postaci $x^T = a[1 \quad 1]$ dla dowolnej stałej a. Od razu otrzymamy też rozwiązanie $y^T = a[1 \quad 1 \quad 1]$, co daje nam zgodny system punktacji, dający te same wartości punktowe dla każdej kategorii cechy wierszowej jak i kolumnowej. Trzeba przyznać, że to rozwiązanie może rozczarować, tym bardziej, że nietrudno zauważyć (bo suma w kolumnach/wierszach profili jest równa 1), iż zachodzi twierdzenie:

Wektory $x^T = a[1 \quad 1 \quad \ldots \quad 1]$ *i* $y^T = a[1 \quad 1 \quad \ldots \quad 1]$ *są rozwiązaniem układu równań* $y = K^T x$, $x = Wy$, *gdzie* K *i* W *są profilami danej macierzy kontyngencji* D.

Takie nieinteresujące rozwiązania nazwiemy trywialnymi. Ta trywialność wynikła z bardzo sztywnego warunku $y = K^T x$, $x = Wy$ definiującego zgodny system punktacji. Warto przecież zauważyć, że dwa wektory: x i równoległy do niego wektor λx, w istocie stanowią równoważny system punktacji.

W takim razie możemy wprowadzić definicję

układem punktacji zgodnym z macierzą kontyngencji D *o* w *wierszach i* k *kolumnach, nazywamy parę nietrywialnych wektorów* x *i* y *spełniających układ równań*

$$(1) \qquad \begin{cases} \mu y = K^T x \\ \lambda x = Wy \end{cases}$$

dla pewnych niezerowych skalarów μ *i* λ. *Macierze* K *i* W *są kolumnowymi i wierszowymi profilami* D.

Rozwiązanie układu (1) metodą podstawienia daje nam równanie

$$WK^T x = \lambda\mu x$$

co jest równoważne stwierdzeniu, że x jest wektorem własnym macierzy WK^T, odpowiadającym niezerowej wartości własnej $\lambda\mu$. Macierz WK^T jest nieujemnie określoną macierzą kwadratową o rozmiarze $w \times w$, a wartość własna $\lambda\mu$ jest liczbą rzeczywistą dodatnią.

Podobnie, y jest wektorem własnym macierzy $K^T W$, o rozmiarze $k \times k$, odpowiadającym tej samej wartości własnej $\lambda\mu$. A więc każdej wartości własnej WK^T odpowiada para wektorów własnych (x, y) odpowiednio macierzy WK^T i $K^T W$ realizujących, zgodny z macierzą kontyngencji D, system punktacji. Par tych jest co najwyżej $\min(w, k)$.

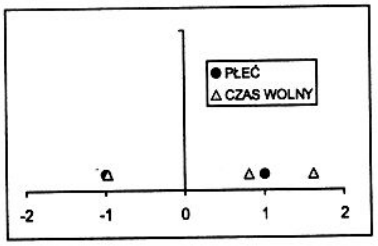
Warto zauważyć, że układ (1) jest równoważny układowi

$$\begin{cases} \rho y^* = K^T x^* \\ \rho x^* = W y^* \end{cases} \tag{2}$$

dla pewnego $\rho > 0$. Wystarczy bowiem podstawić $\rho = \sqrt{\lambda\mu}$, $x^* = x$, $y^* = \sqrt{\frac{\lambda}{\mu}}y$, a otrzymamy układ (2). Parametry ρ, x^*, y^* stanowią trójkę rozwiązań układu (2) (a więc rozwiązanie zagadnienia zgodnego z D systemu punktacji cech). Wektor x^* jest wektorem własnym macierzy WK^T, odpowiadającym niezerowej wartości własnej ρ^2, a y^* jest wektorem własnym macierzy K^TW, odpowiadającym tej samej wartości własnej ρ^2. Jedną z wartości własnych jest $\rho = 1$, której odpowiada rozwiązanie trywialne dla wektorów x i y. Tak więc nietrywialnych rozwiązań jest co najwyżej $\min(w, k) - 1$.

W naszym przykładzie, $WK^T = \begin{bmatrix} 0,69 & 0,31 \\ 0,31 & 0,69 \end{bmatrix}$, wartości własne są równe 1 i 0,38. Wektory własne, odpowiadające tym wartościom mają współrzędne $[1 \quad 1]$ i $[1 \quad -1]$. Pierwsze rozwiązanie, jako trywialne odrzucamy. Pozostaje nam jedno nietrywialne rozwiązanie $x^T = [1 \quad -1]$ (odtąd będziemy podawać rozwiązania równań (2) i opuszczać gwiazdki w oznaczeniach wektorów) z $\rho = \sqrt{0,38} \approx 0,616$. Tak więc poprawna punktacja, przyznana za odpowiedzi kobiet wynosi 1, a za odpowiedzi mężczyzn -1 (z dokładnością do współczynnika proporcjonalności). Punktację za odpowiedzi A, B, C otrzymamy z równań (2):

$$u = \frac{1}{\rho}K^T x = \frac{1}{0,616}\begin{bmatrix} 0,75 & 0,25 \\ 0,2 & 0,8 \\ 1 & 0 \end{bmatrix}[1 \quad -1] = \begin{bmatrix} 0,811 \\ -0,970 \\ 1,622 \end{bmatrix}$$



Mając jednoznaczny system punktacji dla płci i odpowiedzi o wykorzystaniu czasu wolnego (jest tylko jeden nietrywialny zestaw wektorów (x, y)) możemy je narysować na osi liczbowej. Przyjęło się, że rysujemy je we wspólnym układzie osi. Nie jest to bez znaczenia, gdyż bliskość punktów, odpowiadających parom wartości obu cech, jak się za chwilę okaże, świadczy o ich częstym współwystępowaniu.

Przyjmując tę interpretację widzimy, że aktywność typu A związana jest z kobietami, typu B z mężczyznami, natomiast aktywność C jest słabo związana z płcią, choć najbliżej tu do punktu, reprezentującego odpowiedzi kobiet. Do tych wniosków mógł dojść od razu uważny czytelnik tablicy kontyngencji. W innych sytuacjach tak prosta analiza tych tablic nie jest jednak możliwa.

W rachunku prawdopodobieństwa niezależność dwóch zdarzeń oznacza, że prawdopodobieństwo jednoczesnego ich zajścia jest iloczynem prawdopodobieństw zajścia każdego z nich.

W naszym przykładzie oznaczałoby to niezależność płci i sposobu wykorzystania czasu wolnego.

Zaproponowany wyżej sposób wyznaczenia punktów, zgodny z macierzą kontyngencji ma jeszcze jedną zaletę – w bardzo istotny sposób wiąże się z zagadnieniem niezależności cechy, występującej w wierszach i cechy występującej w kolumnach. W tablicy kontyngencji typowy element n_{ij} jest liczbą obserwacji takich, że i-ta wartość cechy wierszowej zachodzi jednocześnie z j-tą wartością cechy kolumnowej. Niezależność cech wierszowej i kolumnowej oznacza, że częstość jednoczesnego zajścia dowolnej pary wartości jest iloczynem częstości zajścia każdej z nich. Oznacza to spełnienie, dla każdego i oraz j, równości

$$\frac{n_{ij}}{n} = \frac{n_{i\cdot}}{n}\frac{n_{\cdot j}}{n}, \tag{3}$$

gdzie $n_{i\cdot}$ jest liczbą obserwacji w i-tym wierszu, $n_{\cdot j}$ – liczbą obserwacji w j-tej kolumnie, a n łączną liczbą wszystkich obserwacji.

W rzeczywistości wzór (3) nie musi zachodzić. Niezgodność z postulatem niezależności można ocenić, obliczając dla każdego elementu tabeli kontyngencji błąd względny

$$e_{ij} = \frac{\frac{n_{ij}}{n} - \frac{n_{i\cdot}}{n}\frac{n_{\cdot j}}{n}}{\frac{n_{i\cdot}}{n}\frac{n_{\cdot j}}{n}},$$

albo jeszcze lepiej jego kwadrat e_{ij}^2. Średni (oczekiwany) kwadratowy błąd względny, gdy cechy wierszowa i kolumnowa są niezależne, można obliczyć ze

wzoru

$$\Phi^2 = \sum_{i=1}^{w}\sum_{j=1}^{k} \frac{n_{i\cdot}}{n}\frac{n_{\cdot j}}{n} e_{ij}^2.$$

Wielkość $n\Phi^2$ oznacza się symbolem χ^2, związanym z powszechnie znanym testem χ^2 Pearsona.

Wielkość Φ^2, zwana inercją, mierzy, na ile obserwacje, zapisane w tablicy kontyngencji są niezgodne z postulatem niezależności cech. Im większa inercja – tym bardziej obserwacje przeczą hipotezie niezależności.

Aby móc ustalić, czy inercja jest duża, używa się tzw. miary V Craméra, zdefiniowanej wzorem $V = \Phi/\sqrt{\min(w,k) - 1}$. Miara V ma wartość maksymalną, równą 1, gdy cechy są w najwyższym stopniu zależne, 0 – gdy są niezależne. W naszym przykładzie, V=0,616.

W naszym przykładzie macierz błędów względnych ma postać

	A	B	C
kobiety	0,5	−0,6	1
mężczyźni	−0,5	0,6	−1

a inercja Φ^2 wynosi 0,38. Uważny Czytelnik przypomni sobie, że z wartością 0,38 spotkał się przy okazji wyznaczania zgodnego systemu punktacji. Nie jest to przypadek. Zachodzi bowiem twierdzenie:

Dla macierzy kontyngencji D suma wartości własnych macierzy WK^T (lub K^TW) różnych od 1 jest równa inercji Φ^2.

Każda wartość własna ma swój udział w inercji. Im większy udział – tym bardziej zależność między cechami opisuje odpowiadająca jej para wektorów własnych (x, y). Tak więc systemy punktowe dla cech wierszowych i kolumnowych możemy uporządkować względem ich malejącego udziału w wyjaśnianiu zależności. Udział ten na ogół wyraża się w procentach całkowitej inercji.

Zazwyczaj rysuje się wykres dwuwymiarowy, opierając się na dwóch pierwszych parach wektorów własnych: parze $x_1^T = [x_{11} \quad x_{12} \quad \dots \quad x_{1w}]$ i $y_1^T = [y_{11} \quad y_{12} \quad \dots \quad y_{1k}]$, odpowiadających największej wartości własnej ρ_1^2, oraz parze $x_2^T = [x_{21} \quad x_{22} \quad \dots \quad x_{2w}]$ i $y_2^T = [y_{21} \quad y_{22} \quad \dots \quad y_{2k}]$, odpowiadających drugiej wartości własnej ρ_2^2. Punkt P_i o współrzędnych (x_{1i}, x_{2i}) reprezentuje i-tą wartość cechy wierszowej, punkt o współrzędnych (y_{1j}, x_{2j}) reprezentuje j-tą wartość cechy kolumnowej. O zależności między wartością cechy wierszowej i kolumnowej świadczy kąt między wektorami umiejscowionymi w początku układu współrzędnych i końcach w punktach P_i oraz Q_j: im ten kąt mniejszy tym większa zależność (kwadrat cosinusa tego kąta jest równy kwadratowi współczynnika korelacji). O braku zależności świadczy rozwarty lub prosty kąt między tymi wektorami.

Osie układu współrzędnych też mają swoją interpretację: są one związane z tymi wartościami cech, które są reprezentowane przez punkty leżące najbliżej osi.

Metoda obrazowania zależności między cechami opisowymi nosi nazwę *analizy odpowiedniości* (nazywanej też analizą korespondencji). Została ona stworzona przez Francuza, libańskiego pochodzenia, Jeana-Paula Benzécriego, w latach 60 dla zastosowania w badaniach językoznawczych.

Weźmy, jako przykład, zestawienie odczytów wygłoszonych na pierwszych 17 Szkołach Matematyki Poglądowej. Odczyty te zostały przeze mnie podzielone na 5 grup tematycznych: G (geometria, topologia), A (analiza, algebra, teoria liczb), D (dydaktyka), Z (zastosowania matematyki i probabilistyka), H (historia). (Podział jest arbitralny; odczyty na pograniczu podziałów zostały przydzielone do wszystkich grup, do których należą.)

Oto tablica kontyngencji tych danych

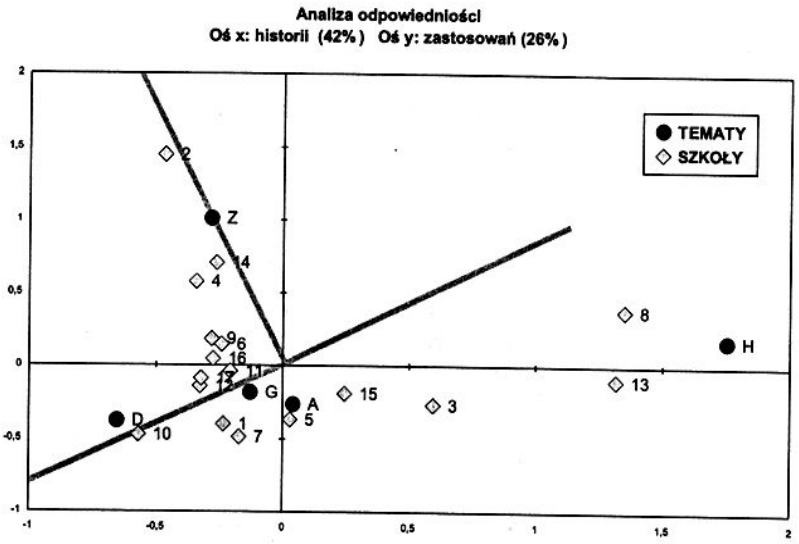
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
G	10	0	6	7	2	4	7	2	9	6	8	8	5	1	5	7	6
A	9	0	4	1	12	7	5	4	1	9	3	3	8	6	5	4	9
D	3	3	0	0	2	1	1	0	0	15	0	2	0	0	1	1	4
Z	1	11	0	5	1	4	0	3	3	2	2	2	0	6	1	3	4
H	0	1	3	0	1	0	0	9	0	0	0	0	11	0	2	0	0

Również wartość $\chi^2 = 248{,}7$ przy tym rozmiarze tablicy oznacza istotną zależność miedzy tematyką a numerem Szkoły.

Inercja dla tych danych wynosi 0,8548, miara Craméra $V = 0,4623$, co potwierdza istotną zależność tematyki i numeru Szkoły.

Pierwsze dwa wektory własne o wartościach własnych 0,3612 i 0,2201 wyczerpują odpowiednio 42% i 26% inercji całkowitej, co oznacza, że oparcie się na tych dwóch systemach punktacji wyjaśnia 68% inercji, będącej, jak pamiętamy miarą stopnia zależności cech.

Wykres punktów, reprezentujących obie cechy: tematykę i numer Szkoły w układzie dwóch pierwszych wektorów własnych przedstawia poniższy rysunek:



Na wykresie zaznaczony został kierunek, związany z zastosowaniami matematyki, co pozwala zauważyć, które Szkoły były mocno związane z tą tematyką. Były to Szkoły 2 (szkoła całkowicie probabilistyczna), 14 (*Aproksymacje*) i 4 (*Geometria innym*). Szkoły wewnątrz kąta prostego utworzonego przez kierunek zastosowań były z tą tematyką blisko związane. Wszystkie punkty, reprezentujące Szkoły, leżące w drugiej półpłaszczyźnie nie są z tą tematyką związane.

Układ współrzędnych związany jest z osią historii i geometrii (oś odciętych) i osią algebry i zastosowań (oś rzędnych). Gęste ułożenie punktów wzdłuż osi odciętych wskazuje, że tematyka geometryczno – historyczna dominowała w tych 17 Szkołach.

Ułożenie punktów w podkowę wskazuje na szczególną postać tablicy kontyngencji. Po przegrupowaniu wierszy tak, aby odzwierciedlały porządek rzutu cech wierszowych na oś odciętych (D, Z, G, A, H) otrzymamy macierz z dużą liczbą obserwacji na przekątnej, co świadczy o tym, że dużo Szkół było monotematycznych i były one ściśle związane z wybranymi przeze mnie grupami tematów.

Mapy różnic

Dane są wzajemne odległości (w km) między: Amsterdamem, Berlinem, Paryżem, Warszawą i Rzymem:

	A	B	P	W	R
A	0	675	523	1261	1807
B	675	0	1176	586	1735
P	523	1176	0	1690	1574
W	1261	586	1690	0	1867
R	1807	1735	1574	1867	0

Odtwórz mapę, na której położone są te miasta.

Wydaje się, że jest to nadzwyczaj proste zadanie konstrukcyjne. Wystarczy skonstruować trójkąt z bokami, łączącymi dowolne trzy wierzchołki,

np. trójkąt ABP o bokach AB = 675, AP = 523, BP = 1176, następnie trójkąty ABW (punkty A,B,W leżą na prostej) i ABR. Do konstrukcji tych pięciu punktów wykorzystaliśmy jedynie odległości, umieszczone w zacieniowanych polach tablicy. Niestety, pozostałe wielkości są różne od odległości między skonstruowanymi przez nas punktami. Dla Paryża i Warszawy odległość z konstrukcji wynosi 1756 km (różnica 66 km), dla Paryża i Rzymu 1835 km (różnica 261 km!), dla Warszawy i Rzymu 1878 km (różnica 11 km). Jest to jeszcze jeden dowód na niepłaskość Ziemi i na to, że tych pięciu punktów nie da się umieścić na płaskiej mapie nie deformując odległości. Gdybyśmy się jednak uparli i chcieli narysować je na płaskiej mapie, to nasze zadanie należałoby nieco zmodyfikować:

Skonstruuj mapę płaską, na której odległości między miastami najmniej różnią się od zadanych.

Metody, które przedstawiamy poniżej pozwalają skonstruować taką mapę, że maksymalna różnica wynosi 27 km.

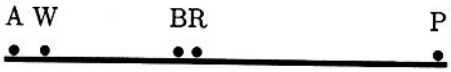
W wielu praktycznych sytuacjach chcemy zilustrować na mapie różnice między elementami danej zbiorowości, opisanej przez określony zestaw cech. Na przykład, chcemy zilustrować różnice pomiędzy elektoratem kilku partii, notując dla każdej z nich rozkład wykształcenia, płci, miejsca zamieszkania (miasto, wieś), nazwy regionów, w których partia uzyskała największe i najmniejsze poparcie. Różnicę między parą partii powinna wyrażać funkcja rozbieżności, której argumentami są obserwowane cechy, a wartościami liczby rzeczywiste nieujemne. Otrzymamy, podobnie jak w przykładzie z miastami, tablicę rozbieżności. Na podstawie tej tablicy chcemy skonstruować mapę taką, że odległości między punktami, symbolizującymi partie, będą jak najdokładniej zbliżone do miary rozbieżności między nimi.

Ocenimy różnice między liczbą mieszkańców miast, których położenie na płaskiej mapie odtwarzaliśmy na podstawie tabeli odległości. Naturalną miarą rozbieżności między dwoma miastami będzie tu moduł różnicy liczby ludności tych miast.

Ludność w mln.	A	B	P	W	R
A (1,1)	0	2,8	9,6	0,5	2,9
B (3,9)	2,8	0	6,8	2,3	0,1
P (10,7)	9,6	6,8	0	9,1	6,7
W (1,6)	0,5	2,3	9,1	0	2,4
R (4,0)	2,9	0,1	6,7	2,4	0

Dla każdych trzech punktów nierówność trójkąta jest równością. Jest to oczywiste, gdyż punkty są scharakteryzowane jedną cechą, a odległość między nimi jest euklidesową odległością na prostej.

Łatwo sprawdzić, że każde trzy punkty muszą leżeć na linii prostej, a więc mapa, reprezentująca rozbieżności między ludnością tych miast, będzie jednowymiarowa :



Pierwszy użył seriacji do datowania znalezisk W.M.F. Petrie w 1899 roku. Technika komputerowa oraz nowoczesne metody statystyczne (program *Bonn Archeological Statistical Package*) pozwalają użyć metody seriacji w bardzo trudnych do analizy archeologicznej przypadkach.

Jednowymiarowość, choć bardzo rzadko spotykana, umożliwia uporządkowanie obiektów według podobieństwa cechy. Uporządkowań takich, noszących nazwę seriacji, szuka się w archeologii. Na przykład, uporządkowanie skorup naczyń według podobieństwa motywów pozwala na datowanie znalezisk. Na ogół ten wymiar jest większy i próba umieszczenia mapy zróżnicowania na płaszczyźnie, lub w przypadku seriacji na prostej, wymaga znalezienia najlepszej reprezentacji o danym wymiarze, mało deformującej rozbieżności.

W naszym przykładzie z odległościami miast, odległości każdych trzech z nich spełniają nierówność trójkąta, co może świadczyć o tym, że odległości były mierzone po prostej, a nie po linii geodezyjnej.

Na przykład, trzy pary rozbieżności 1,2,5 nie pozwalają zbudować trójkąta.

Najczęściej nie tylko mamy problem z wymiarami, ale również z tym, że rozbieżności nie spełniają nierówności trójkąta, co powoduje, że nie istnieje przestrzeń euklidesowa, w której odległości między punktami są równe rozbieżnościom.

Środek ciężkości układu punktów jest w punkcie, którego współrzędne są średnimi arytmetycznymi punktów układu.

Przypuśćmy, że punkty $P_1, P_2, \ldots, P_n$ o p współrzędnych leżą w przestrzeni p-wymiarowej. Początek układu współrzędnych wybierzemy w środku ciężkości tego układu punktów. Jaki warunek musi spełniać ich macierz odległości euklidesowych?

Oznaczmy przez X macierz $X = \lfloor x_{ij} \rfloor$ rzędu $n \times p$, gdzie $\lfloor x_{ij} \quad x_{i2} \quad x_{ip} \rfloor$ oznaczają współrzędne punktu P_i. Niech d_{rs}^2 oznacza kwadrat odległości

euklidesowej między punktami P_r i P_s. Wtedy

$$d_{rs}^2 = \sum_{j=1}^{p}(x_{rj} - x_{sj})^2 = \sum_{j=1}^{p} x_{jr}^2 + \sum_{j=1}^{p} x_{sj}^2 - 2\sum_{j=1}^{p} x_{rj}x_{sj}.$$

Oznaczając przez $Q = [q_{rs}]$ macierz XX^T otrzymamy równość

$$q_{rs} = \sum_{j=1}^{p} x_{rj}x_{sj}$$

oraz tożsamość

$$d_{rs}^2 = q_{rr} + q_{ss} - 2q_{rs}. \tag{4}$$

Dodając ostatnie równania stronami najpierw po r, potem po s, a następnie jednocześnie po r i s otrzymamy układ równości:

$$\begin{cases} \sum_r d_{rs}^2 = A + nq_{ss} \\ \sum_s d_{rs}^2 = A + nq_{rr} \\ \sum_s \sum_r d_{rs}^2 = 2nA \end{cases}, \tag{5}$$

gdzie $A = \sum_r q_{rr}l$.

W równościach tych uwzględniono, że warunek położenia początku układu współrzędnych w środku ciężkości punktów jest równoważny zachodzeniu, dla każdego r, równości $\sum_{j=1}^{p} x_{rj} = 0$.

Redukując parametr A z równań (5) i wstawiając do równania (4) otrzymamy tożsamość

$$q_{rs} = -\frac{1}{2}(d_{rs}^2 - d_{r\cdot}^2 - d_{\cdot s}^2 + d_{\cdot\cdot}^2), \tag{6}$$

gdzie $d_{r\cdot}^2 = \frac{1}{n}\sum_s d_{rs}^2$, $d_{\cdot s}^2 = \frac{1}{n}\sum_r d_{rs}^2$, $d_{\cdot\cdot}^2 = \frac{1}{n^2}\sum_r \sum_s d_{rs}^2$,.

Wzór ten oznacza, że

dysponując macierzą odległości punktów potrafimy odtworzyć macierz Q, która jest kwadratem macierzy X, równym XX^T.

Macierz Q jest symetryczna, a jako kwadrat macierzy X jest nieujemnie określona. Z ogólnej teorii takich macierzy wynika, że istnieje macierz ortogonalna T oraz macierz diagonalna Λ takie, że $Q = T\Lambda T^T$. Kolumny macierzy T są wektorami własnymi, a elementy przekątnej Λ odpowiednimi nieujemnymi wartościami własnymi macierzy Q (jest ich $n-1$, bo rząd macierzy Q jest co najwyżej równy $n-1$). W macierzy T można tak przestawić kolumny $t_1, t_2, \ldots, t_{n-1}$ i odpowiadające im wartości własne $\lambda_1, \lambda_2, \ldots, \lambda_{n-1}$, aby zachodziły nierówności $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{n-1}$. Co więcej, prawdziwe jest następujące stwierdzenie

Niech dana będzie macierz rozbieżności $D = [d_n]$ dla $r, s = 1, 2, \ldots, n$. Jeżeli istnieje zbiór punktów $P_1, P_2, \ldots, P_n$, dla których D jest macierzą odległości euklidesowych, to ich współrzędne są wierszami macierzy X, wyznaczonej ze wzoru:

$$X = T\sqrt{\Lambda},$$

gdzie macierz Q, obliczoną z (6) można przedstawić w postaci

$$Q = T\Lambda T^T$$

dla pewnej macierzy ortogonalnej T i diagonalnej Λ. Liczba współrzędnych punktów P_i jest równa liczbie niezerowych wartości własnych Q.

Ograniczając się do pierwszych dwóch kolumn macierzy X otrzymamy dwuwymiarową reprezentację macierzy rozbieżności D. Jest to optymalna reprezentacja w tym sensie, że spośród wszystkich prostopadłych rzutów układu punktów $P_1, P_2, \ldots, P_n$ na płaszczyznę najmniejszy błąd, równy

$\lambda_3 + \lambda_4 + \ldots + \lambda_{n-1}$, ma rzut na płaszczyznę, rozpiętą przez pierwsze dwie kolumny macierzy X.

Aby istniała euklidesowa reprezentacja macierzy rozbieżności D, powstała z niej macierz Q musi być nieujemnie określona. W przeciwnym przypadku taka reprezentacja nie istnieje. Wtedy szuka się „w pobliżu" macierzy D takiej, że powstała z niej macierz Q jest nieujemnie określona i dla niej szuka się odpowiedniej reprezentacji.

Metoda, opisana w tym punkcie, znana jest pod nazwą skalowania wielowymiarowego. Pozwala ona nadać cechom niekoniecznie liczbowym ich walorów punktowych. Jest więc podobna do metody analizy odpowiedniości, gdzie punktacja ma opisać siłę zależności między dwiema cechami. W metodzie skalowania wielowymiarowego odległości między punktami wyrażają rozbieżności między wartościami jednej cechy.

Dla ilustracji posłużymy się przykładem, w którym wykorzystamy metodę skalowania wielowymiarowego do analizy pewnych obserwacji w językoznawstwie.

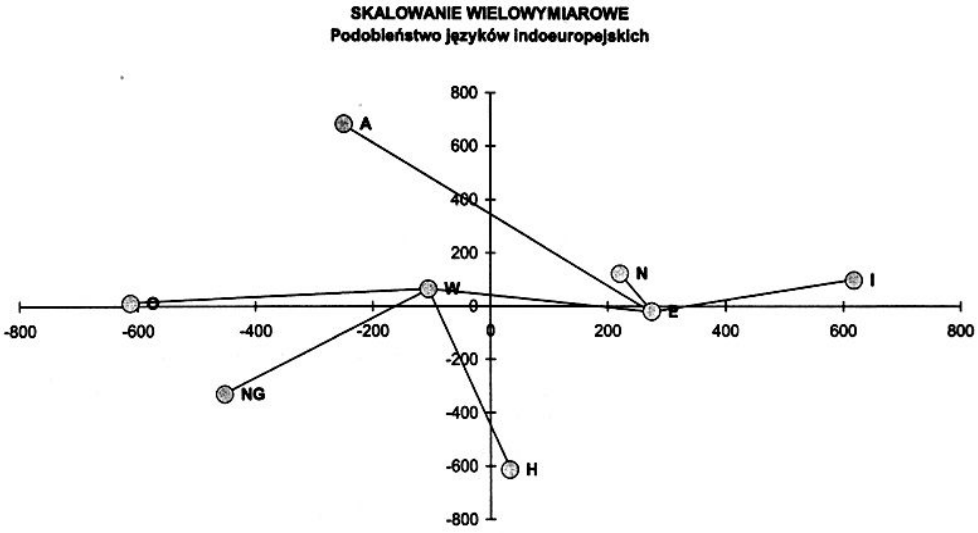
W książce [3, str. 138], *językoznawca, prof. Witold Mańczak opisał przeprowadzone przez siebie badania nad podobieństwem języków indoeuropejskich. Celem badań było odtworzenie drogi, jaką język praindoeuropejski wędrował po terytorium Europy. W tym celu porównał on te same fragmenty ewangelii w tłumaczeniu na języki albański (A), hindi (H), irlandzki (I), litewski (L), niemiecki (N), nowogrecki (NG), ormiański (O), polski (P) i włoski (P). Podobieństwo dwóch języków określone było jako liczba słów jednego języka, mających odpowiedniki etymologiczne w drugim języku.*

	P	L	N	W	I	NG	H	O	A
P		824	565	412	448	277	287	253	271
L	824		542	377	390	219	283	231	204
N	565	542		407	378	296	236	243	223
W	412	377	407		311	320	302	306	248
I	448	390	378	311		158	204	130	104
NG	277	219	296	320	158		220	302	215
H	287	283	236	302	204	220		274	146
O	253	231	243	306	130	302	274		164
A	271	204	223	248	104	215	146	164	

Aby zbudować tablicę rozbieżności należałoby znać liczbę słów jednego języka, nie mających odpowiedników etymologicznych w drugim języku. Nie dysponując tą liczbą, przyjąłem, że średnio tekst tłumaczenia ewangelii zawierał 1000 słów, więc tablica rozbieżności mogłaby mieć postać:

	P	L	N	W	I	NG	H	O	A
P	0	176	435	588	552	723	713	747	729
L	176	0	458	623	610	781	717	769	796
N	435	458	0	593	622	704	764	757	777
W	588	623	593	0	689	680	698	694	752
I	552	610	622	689	0	842	796	870	896
NG	723	781	704	680	842	0	780	698	785
H	713	717	764	698	796	780	0	726	854
O	747	769	757	694	870	698	726	0	836
A	729	796	777	752	896	785	854	836	0

Wyznaczymy dla tej tablicy płaską mapę języków. Każdy język reprezentuje punkt ma płaszczyźnie, której współrzędne są tak wybrane, aby odległości między punktami odpowiadały rozbieżnościom w tablicy. Na przykład, język polski ma współrzędne $[275, -19]$, włoski $[-106, 66]$ co daje odległość 389, gdy rzeczywista rozbieżność wynosi 588. Należy jednak pamiętać, że mamy tu do czynienia z rzutem z przestrzeni 8-wymiarowej na płaszczyznę.



Punkty na wykresie zostały połączone odcinkami, tworzącymi minimalny graf rozpinający. Jest to najkrótszy graf niecykliczny łączący te punkty.

Autor książki nie przygotował takiego wykresu, ale jego wnioski są doskonale na nim widoczne:

Litewski, niemiecki, włoski, irlandzki i albański są podobne przede wszystkim do polskiego. Fakt (ten)(...) idzie w parze z (...) wnioskiem, że praojczyznę indoeuropejską należy lokalizować w dorzeczu Wisły i Odry. Natomiast jak interpretować fakt, że nowogrecki, ormiański i hindi są spokrewnione przede wszystkim z włoskim? Wydaje się, że należy z tego wnosić, że plemiona indoeuropejskie, które się osiedliły w Italii, Grecji, Armenii i Indiach, wywędrowały z jakiegoś obszaru położonego gdzieś na Półwyspie Bałkańskim.

Na naszym wykresie punkt, reprezentujący język litewski praktycznie pokrywa się z punktem dla języka polskiego (moja uwaga).

Jak widać, wszystko da się narysować. Matematyk też to potrafi.

Literatura

[1] M. Kordos, *Wykłady z historii matematyki*, Warszawa 1994, WSiP.
[2] W.J. Krzanowski, *Principles of Multivariate Analysis*, Oxford 1996, Clarendon Press.
[3] W. Mańczak, *Wieża Babel*, Wrocław 1999, Ossolineum.