METODY GRADACYJNE W ANALIZIE DANYCH WIELOWYMIAROWYCH

infrastruktura i implementacja

Plan prezentacji

Podstawowa infrastruktura gradacyjna – Rozwój infrastruktury – Implementacja metod – Przykłady zastosowań

Podstawowa infrastruktura gradacyjna

Dane wielowymiarowe – Metody gradacyjne – Modele lilipucie – Pomiar zależności – Regularność

Dane wielowymiarowe

- głównym celem ich eksploracji jest znalezienie zależności między zmiennymi (cechami, atrybutami) i obiektami
- ważna jest też ich klarowna wizualizacja
- rzadko można traktować je jako próbkę losową pochodzącą z dobrze zdefiniowanej populacji

Problemy obliczeniowe

(za: *3rd World Conference on Computational Statistics & Data Analysis, Cypr, 2005*)

- wielka liczba obiektów i/lub zmiennych
- nieoczyszczone dane – brakujące, odstające lub błędne
- niejednorodność danych
- dane, których typ trudno sklasyfikować (nienumeryczne bądź trudne do kwantyfikacji)
- niełatwe do sprecyzowania zapytania badawcze lub celowe eksploracyjne nastawienie wobec danych

Sedno analizy gradacyjnej

- porównywanie par rozkładów, gdzie para rozkładów jednowymiarowych jest reprezentowana przez pojedynczą zmienną lilipucią (tj. zmienną określoną na *odcinku jednostkowym*), a parę rozkładów wielowymiarowych reprezentuje pojedyncza para zmiennych lilipucich (określona na *kwadracie jednostkowym*)
- uciąglona dystrybuanta zmiennej bądź pary zmiennych lilipucich nazywa się *krzywą koncentracji* lub *powierzchnią koncentracji*, a wiodące do niej przekształcenia są nazywane *gradacyjnymi*
- termin *grade of x* wprowadzono w literaturze anglojęzycznej wiele lat temu do opisania wartości *x* zmiennej *X* przekształconej przez dystrybuantę; gradacja jest probabilistycznym odpowiednikiem rangowania ze względu na cechę *X*

Sedno analizy gradacyjnej

- rozkłady wielowymiarowe są tu reprezentowane przez *dwuwymiarowe macierze danych*, które przy pewnych założeniach można potraktować jako *dwudzielne tablice prawdopodobieństw*
- pomimo że w praktyce trudno jest idealnie spełnić wymagane założenia (zazwyczaj z powodu obecności zmiennych mierzonych na różnych skalach), metody gradacyjne okazywały się zwykle wystarczająco odporne, by wykryć strukturę (model) danych

Probabilistyczne modele lilipucie

- rozkłady określone na kwadracie jednostkowym z jednostajnymi rozkładami brzegowymi tworzą tzw. *kopuły*
- zbiór kopuł zawiera zagnieżdżone podzbiory ze stopniowo coraz bardziej *regularnymi monotonicznymi zależnościami* między zmiennymi brzegowymi (sumą wierszy i sumą kolumn)

Pomiar zależności gradacyjnej

- monotoniczna zależność w kopułach (a zatem i w odpowiadających im zbiorach danych) może być mierzona wskaźnikiem ρ^* zwanym **korelacją gradacyjną** pary zmiennych (X,Y) lub ρ-Spearmana
- innym (pokrewnym) gradacyjnym wskaźnikiem monotonicznej zależności w kopule jest τ *Kendalla*
- oba wskaźniki są funkcjami macierzy wskaźników koncentracji kolumny *t* do kolumny *s* (s < t) lub wiersza *j* do wiersza *i* (i < j), przy czym funkcje te są rosnące ze względu na każdy element macierzy

Wskaźniki ρ^* i τ

- umożliwiają określenie stopnia regularności kopuły, a także:
- ustalenie przynależności do wysoce regularnych kopuł oznaczanych TP_2 (*totally positive of order 2*)
- ustalenie przynależności do zbioru takich kopuł, dla których żadna permutacja wierszy i kolumn nie może zwiększyć wartości wskaźnika (tzw. kopuły optymalnie uporządkowane względem zależności monotonicznej)

Kopuły TP_2 vs. kopuły maksymalnie uporządkowane

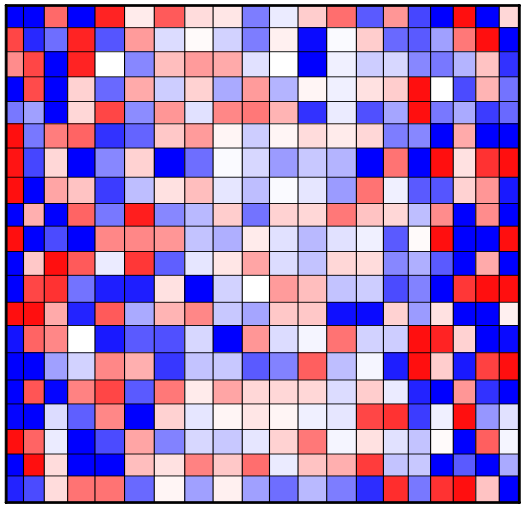
- w kopułach TP_2 wskaźniki koncentracji dla **wszystkich par** wierszy/kolumn, ustawione zgodnie z optymalnym uporządkowaniem **według zależności monotonicznej,** osiągają swoje maksymalne wartości
- w każdej optymalnie uporządkowanej kopule macierz wskaźników koncentracji zawiera „możliwie największą" liczbę maksymalnych wskaźników koncentracji, z czego wynika że:
  - każda optymalnie uporządkowana kopuła „zbliża się" do TP_2 „tak bardzo jak to możliwe"
  - każda kopuła TP_2 jest optymalnie uporządkowana

Wskaźniki ρ^*_{abs} i τ_{abs}

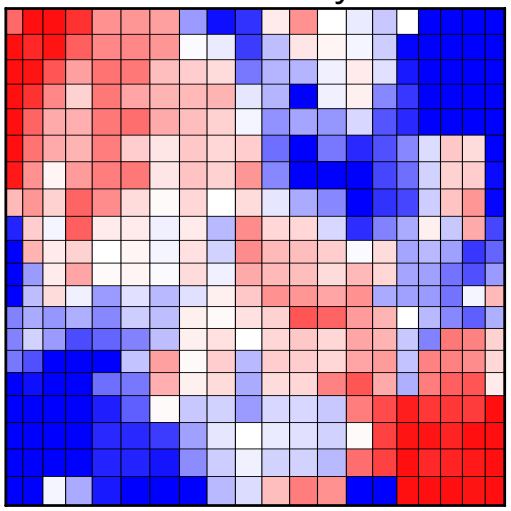
- wprowadźmy dwa parametry: ρ^*_{abs} i τ_{abs} , które zamiast być liczonymi z macierzy wskaźników koncentracji są liczone z macierzy **maksymalnych** wskaźników koncentracji; wtedy:
  - $\rho^* <= \rho^*_{abs}$ i $\tau <= \tau_{abs}$
- udowodniono, że optymalnie uporządkowana (względem zależności monotonicznej) kopuła jest TP_2 wtedy i tylko wtedy gdy:
  - $\rho^* = \rho^*_{abs}$ lub $\tau = \tau_{abs}$

Hierarchia uporządkowania - przykład

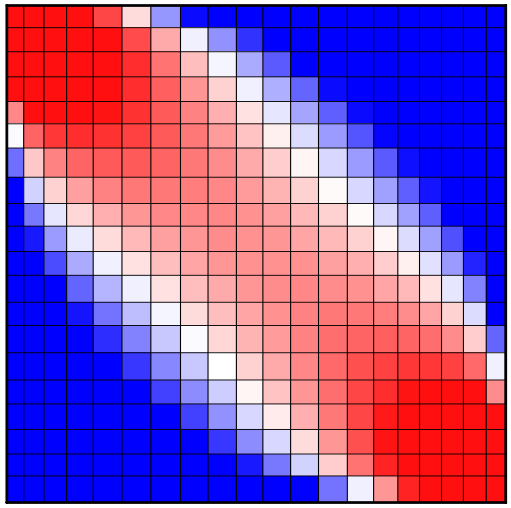
brak zależności



zależność maksymalna



zależność idealna



Regularność

- ilorazy $\rho^*_{max}/\rho^*_{abs}$ i τ_{max}/τ_{abs} wskazują, jak bliska jest **dodatnia zależność** optymalnie uporządkowanej kopuły (macierzy danych) do **zależności** w odpowiadającej jej macierzy typu TP_2
- pomiar regularności zbioru danych jest ważny, ponieważ im regularniejsze dane, tym czytelniejsza struktura zależności między zmiennymi i obiektami

Znaczenie regularności w danych

- silna zależność między zmiennymi i obiektami połączona z regularnością pomaga przy opisie i porównywaniu danych, przy podziale na skupienia, w problemach predykcyjnych itd.
- miary regularności są używane do podziału każdego zestawu danych na dalsze skupienia rekordów, oparte o modele o regularności monotonicznej innej niż w źródłowej macierzy

Rozwój infrastruktury

„Infrastruktura gradacyjna dopełnia oraz koryguje infrastrukturę tradycyjną; obie są równoważne w modelach bardzo regularnych”

Krzywe Kendalla i Spearmana

- Schweizer i Wolff (1981) zauważyli, że w ciągłym rozkładzie pary (X,Y) zachodzi:

$$\tau(X,Y) = 4\int_0^1\int_0^1 Cop\,(u,v)\,cop\,(u,v)\,dudv - 1$$

$$\rho^*(X,Y) = 12\int_0^1\int_0^1 Cop\,(u,v)\,dudv - 3$$

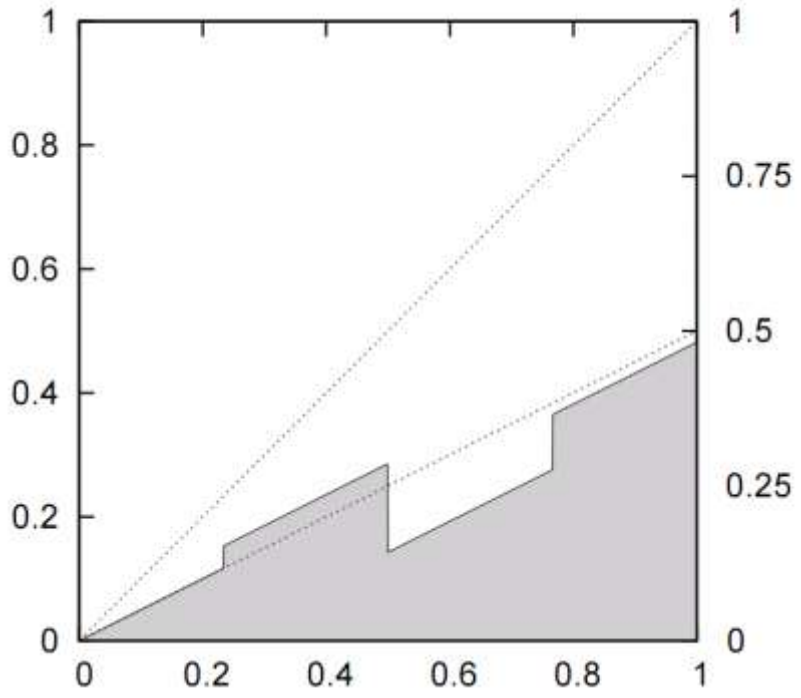
  - gdzie $Cop_{X,Y}$ oznacza kopułę utworzoną z rozkładu (X,Y), a $cop_{X,Y}$ oznacza gęstość prawdopodobieństwa w kopule

- W innym ujęciu:
  - τ (X,Y) = 4 (masa pod kopułą pary (X,Y))-1
  - ρ*(X,Y) = 12 (objętość pod kopułą pary (X,Y))-3

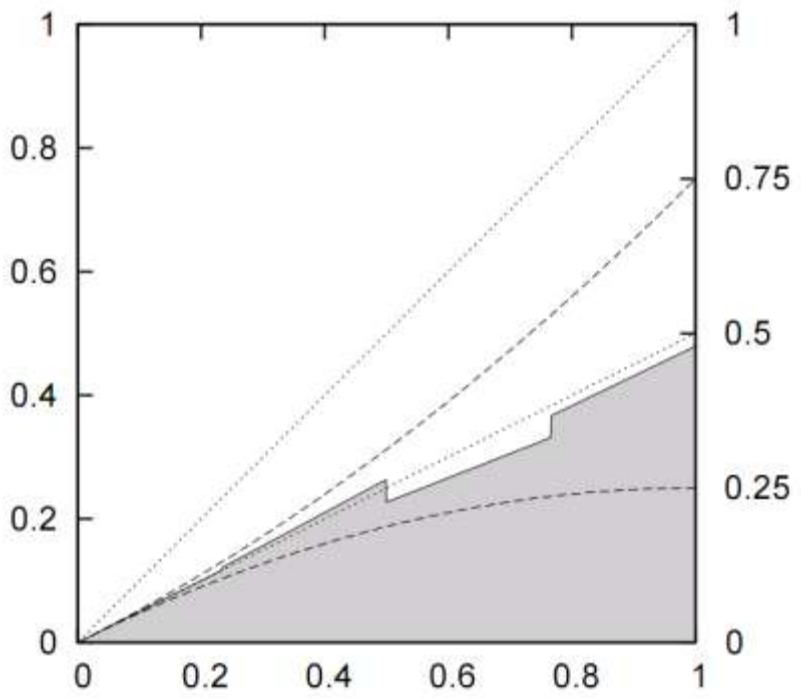
- Kolejni autorzy (prace z lat 1983-2007) wykazali, że te wzory są prawdziwe również dla par zmiennych dyskretnych

Krzywe Kendalla i Spearmana – przykład dla macierzy 7x4

krzywa Kendalla



krzywa Spearmana



Krzywe Kendalla i Spearmana we wzorach analogicznych do Schweizera i Wolffa

- $\tau(X,Y) = 4$ (pole pod krzywą Kendalla)-1
- $\rho^*(X,Y) = 12$ (pole pod krzywą Spearmana)-3

- ponieważ pole, w którym mogą występować krzywe Kendalla jest **trzykrotnie** większe niż pole, w którym mogą występować krzywe Spearmana,
  to współczynniki funkcji liniowej, które normalizują pola pod krzywymi do przedziału $\langle -1;1 \rangle$, są **trzykrotnie** większe dla ρ^* niż dla τ

Krzywe Kendalla i Spearmana

- przebieg krzywej Kendalla lub Spearmana rejestruje np. zmiany, jakie zachodzą przy zwiększaniu macierzy danych o później zmierzoną wartość cechy (np. gdy dodamy wynik badań grupy chorych w kolejnych dniach); przebieg rejestruje także zmiany regularności
- dla przykładu: przy badaniu dzieci kilka razy testem, jest to uzupełnienie wartości współczynnika T lub ρ^* (która daje nam tylko wartość sumaryczną o zależności dla uporządkowania dzieci i wartości wyników) o informację o zależności w kolejnych badaniach (w których odcinkach zależności były dodatnie, w których ujemnie)
- praca autorstwa T. Kowalczyk i W. Szczesnego pt. „Kendall and Spearman Curves" jest prawie gotowa do druku

Nierówność dla wielu zmiennych

- jest to dekompozycja wskaźnika τ, który wyznacza się z rozkładu dwuwymiarowego, utworzonego z wielowymiarowej macierzy danych uzupełnionej wektorem prawdopodobieństw rekordów (**Z**)
- w przypadku pojedynczej zmiennej nierówność przybiera wartość kierunkowego wskaźnika Giniego (zmiennej **Z**); dla wielu zmiennych nierówność jest wypukłą kombinacją kierunkowego wskaźnika Giniego dla **Z** i kierunkowego τ Kendalla dla macierzy
- korzysta się z tego, że kierunkowe τ Kendalla (czyli τ dla macierzy z permutowanymi wierszami i kolumnami) jest liniową kombinacją kierunkowych wskaźników koncentracji dla *par rozkładów warunkowych*

Nierówność dla wielu zmiennych

- proponowana gradacyjna miara nierówności maksymalizuje tę wypukłą kombinację po wszystkich parach uporządkowań rekordów i zmiennych
- stanowi to kompromis między zróżnicowaniem sumy zmiennych i zróżnicowaniem profili rekordów, co sugeruje dwudzielną analizę skupień dla rekordów
- praca T. Kowalczyk, E. Pleszczyńskiej, W. Szczesnego i M. Wiecha „Grade multivariate inequlity and diversity with implications for clustering” jest prawie gotowa do druku

Implementacja metod

GradeStat – Charakterystyka – Krótki schemat analizy danych – Mapy nadreprezentacji – GCA – Analiza skupień – Wizualizacje – Metody gradacyjne i klasyczne

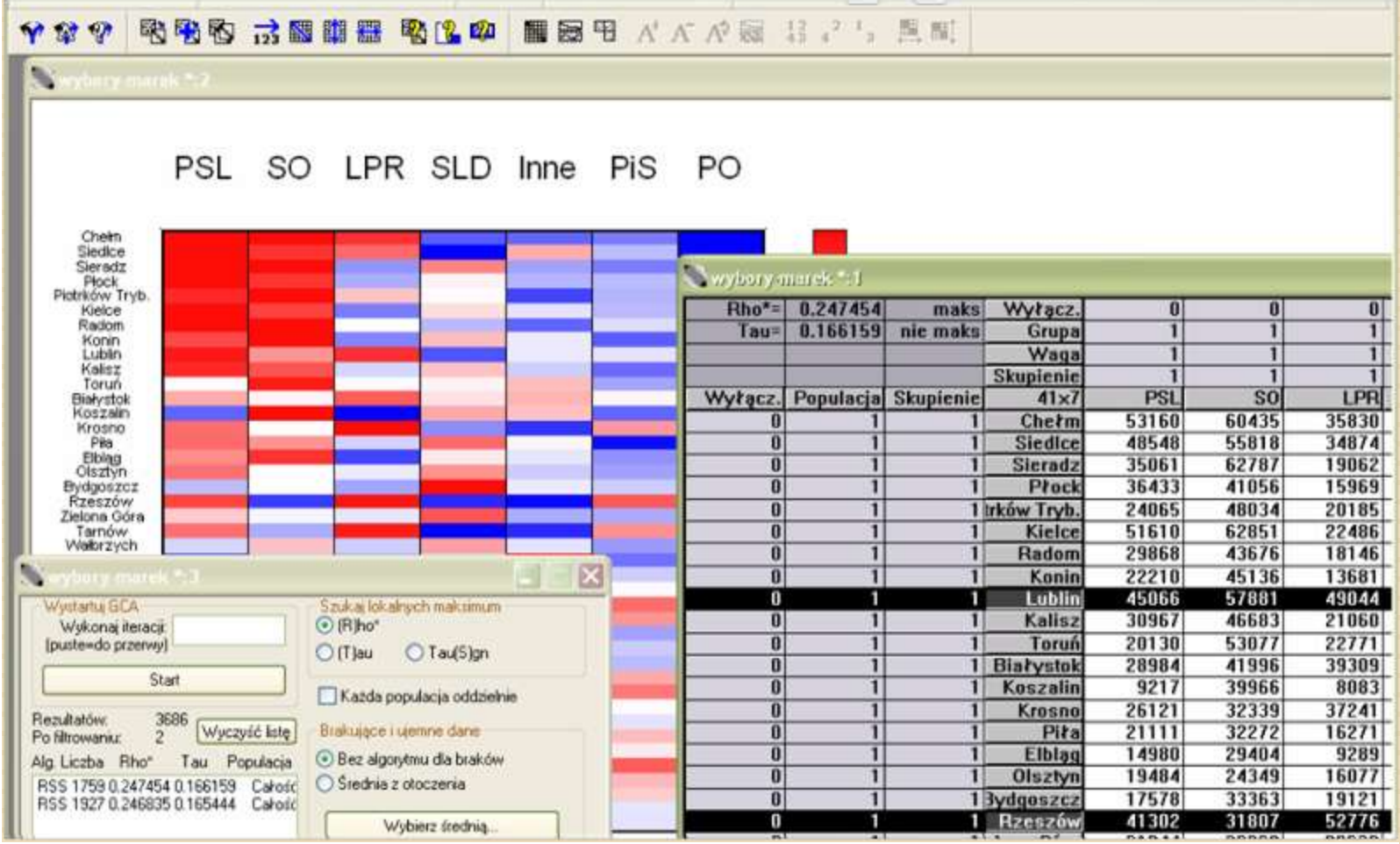
Oprogramowanie

- dobre algorytmy bez implementacji pozostają w sferze rozważań teoretycznych
- dobrą implementacją algorytmów analizy powinno być oprogramowanie o przyjaznym interfejsie, inaczej metodami zainteresuje się jedynie wąskie grono specjalistów
- przy zalewie informacji i łatwym operowaniu dużymi zbiorami kluczowa staje się czytelna wizualizacja macierzy danych i rezultatów ich analizy

GradeStat

- metody gradacyjne są zaimplementowane w programie GradeStat, autorstwa dr Olafa Matyji
- program ma dwujęzyczny interfejs (polski i angielski)
- napisany w Visual C++ pod Microsoft Windows (95/98/ME/XP/Vista), działa jednak również pod X-Windows
- wersja instalacyjna mieści się na... dyskietce

GradeStat 3.0



ilustracja typowej pracy w programie: w tle wizualizacja mapy danych, na pierwszym planie macierz danych

Główne cechy użytkowe GradeStatu

- import danych (tabeli) zapisanych w dowolnym formacie tekstowym lub przeniesionych przez Schowek Windows (np. z Excela)
- wielkość tabeli zależna tylko od wielkości pamięci RAM komputera
- wydzielona biblioteka obliczeniowa (GradeAPI.dll), z której można korzystać z zewnętrznych programów
- interfejs graficzny + wizualizacje

Typowy schemat podstawowej analizy danych w GradeStacie

1. macierz danych

- wartości nieujemne
- brakujące dane usunięte lub uzupełnione
- przezroczyste dla użytkownika przekształcanie na rozkład gradacyjny

2. GCA

- Gradacyjna Analiza Odpowiedniości (GCA) - uporządkowanie maksymalizujące kontrast między skrajnymi wierszami i kolumnami

3. regularność

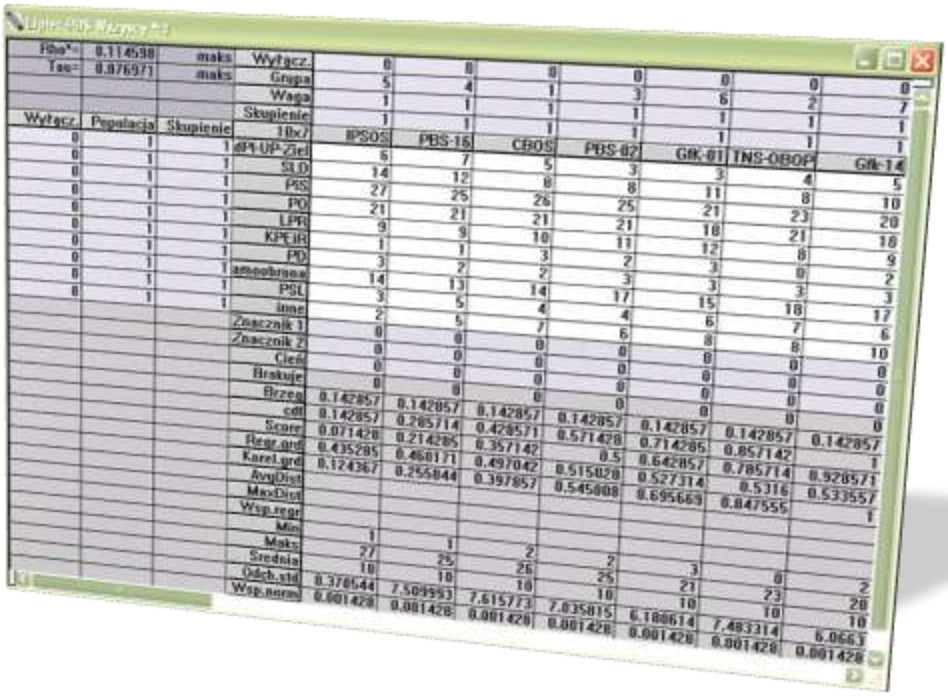
- ocena regularności macierzy (pod kątem zależności monotonicznej)
- elementy zgodne z trendem: podpopulacja FIT
- elementy odstające: podpopulacja OUT

4. analiza skupień

- elementy FIT rozdzielane do zadanej liczby skupień
- elementy OUT ponownie porządkowane przez GCA i analizowane

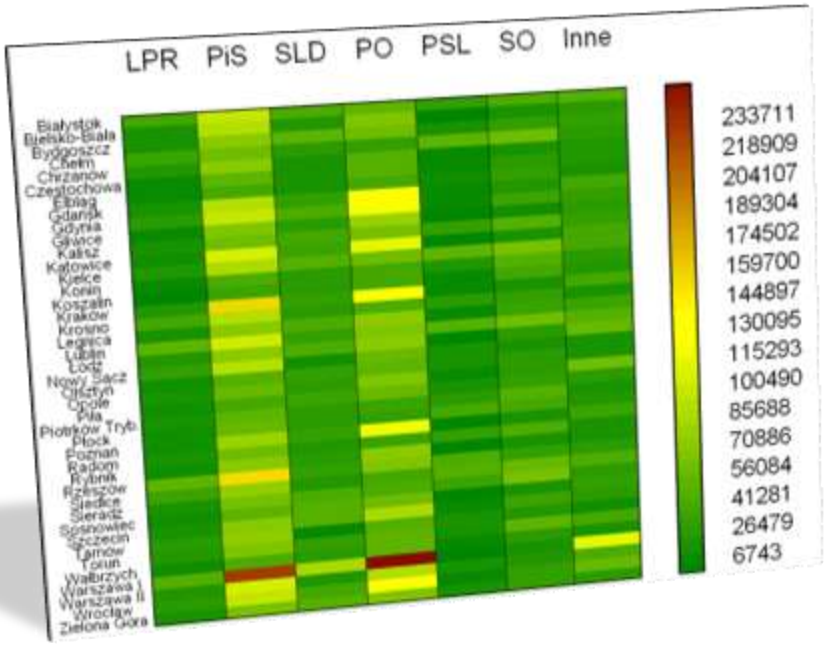
Macierz danych jako tabela

- dane są zaprezentowane jako edytowalny „arkusz danych”
- użytkownik widzi oryginalne wartości, ale przy obliczeniach wykonywana jest „niewidzialna” normalizacja danych

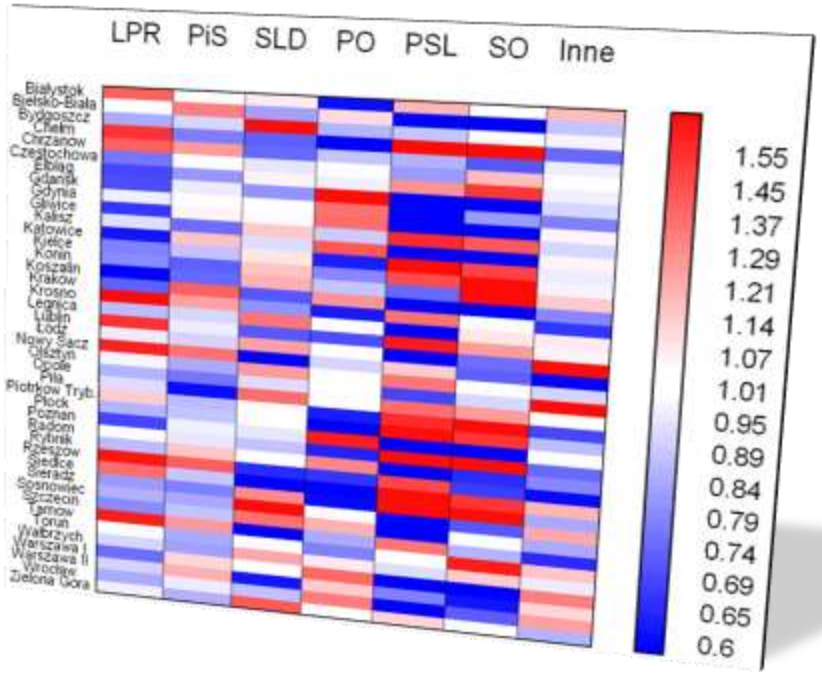


Mapy nadreprezentacji (gęstości gradacyjnej)

dane surowe – gdzie oddano ile głosów na każdą z partii



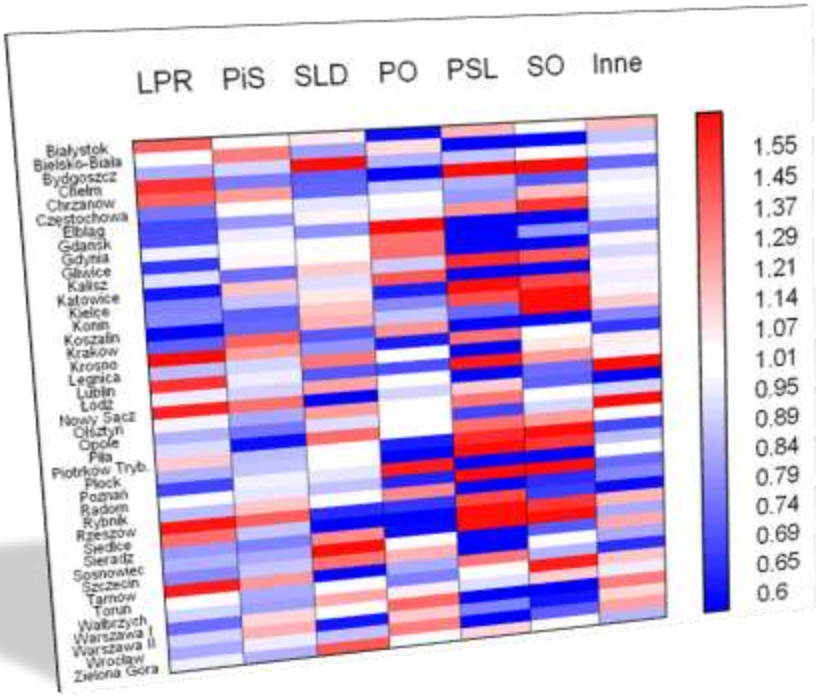
nadreprezentacje głosów oddanych w okręgach

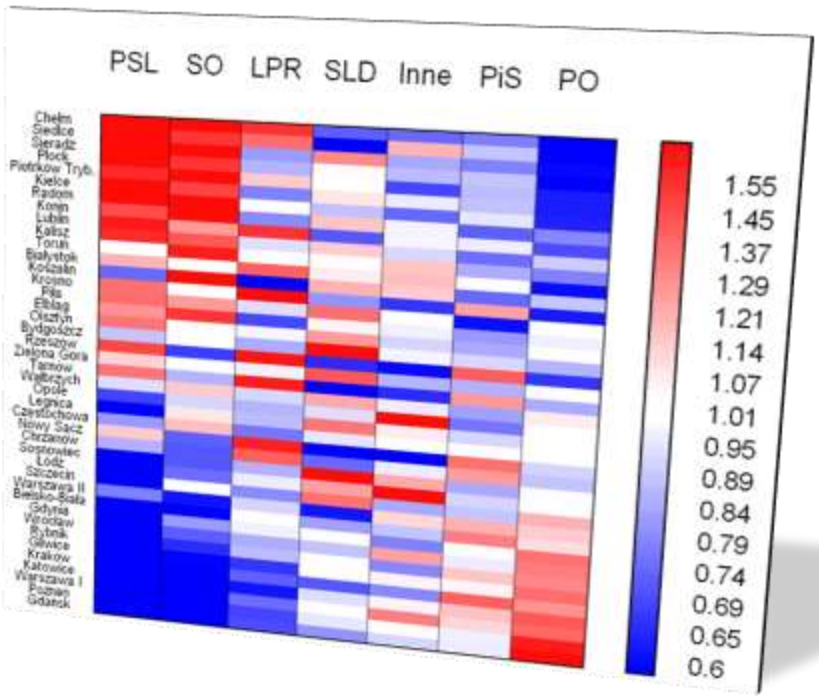


Gradacyjna Analiza Odpowiedniości (GCA)

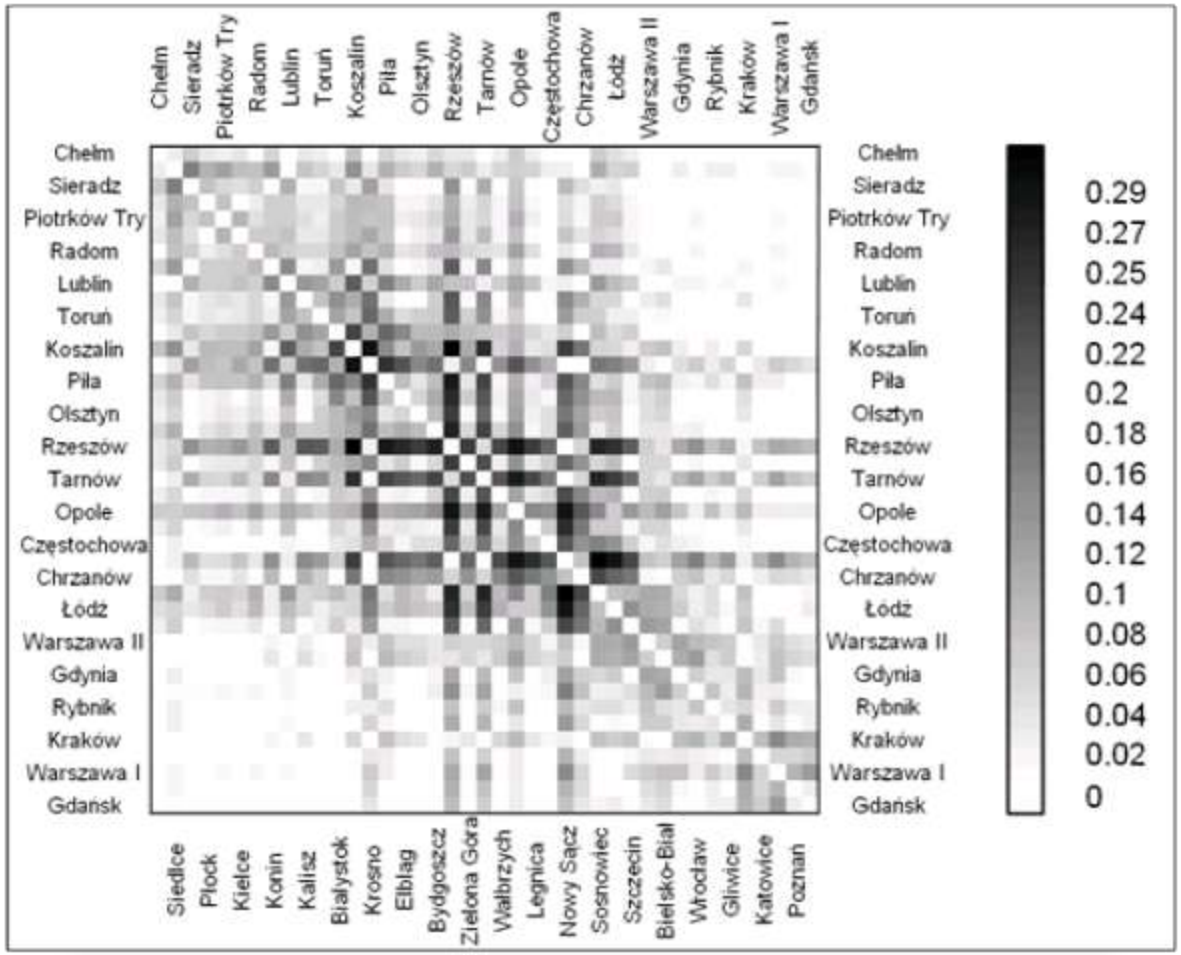
przed GCA: $\rho^* \approx -0{,}00004$

po GCA: $\rho^*_{max} = 0{,}247$



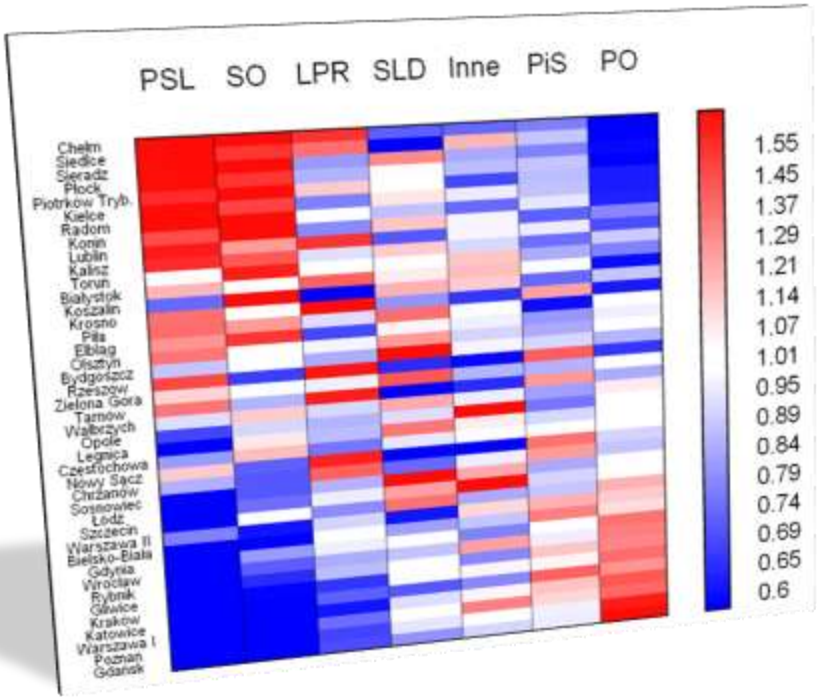


Regularność – odstępstwo od trendu

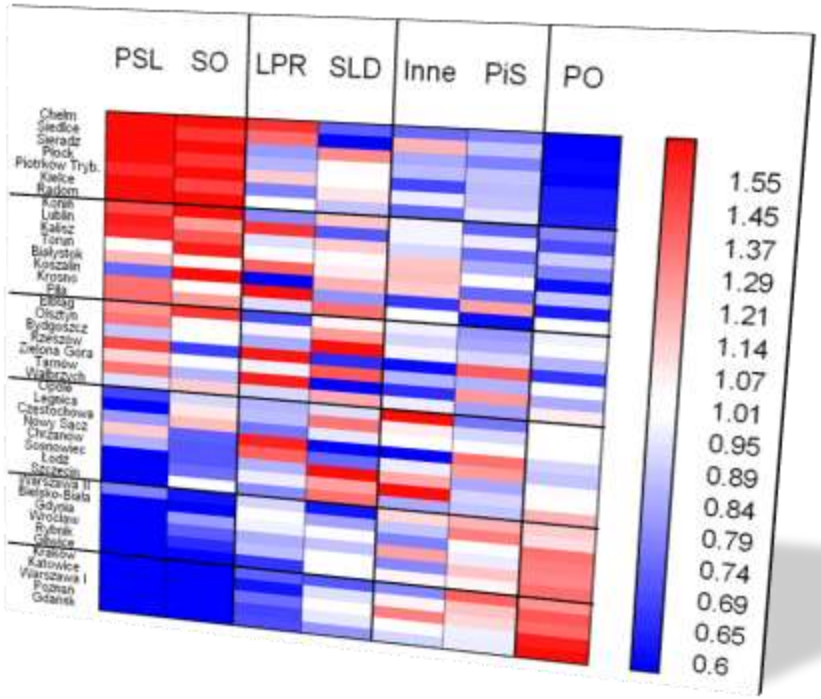


Analiza skupień

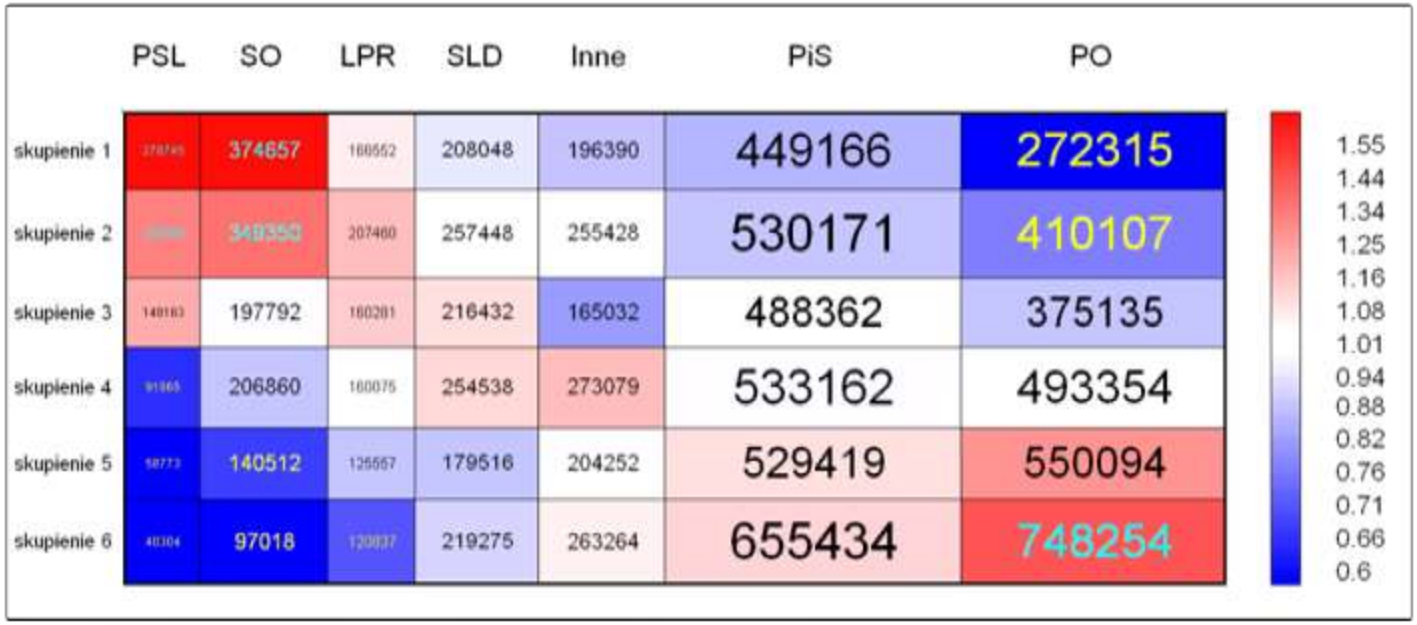
Przed podziałem na skupienia



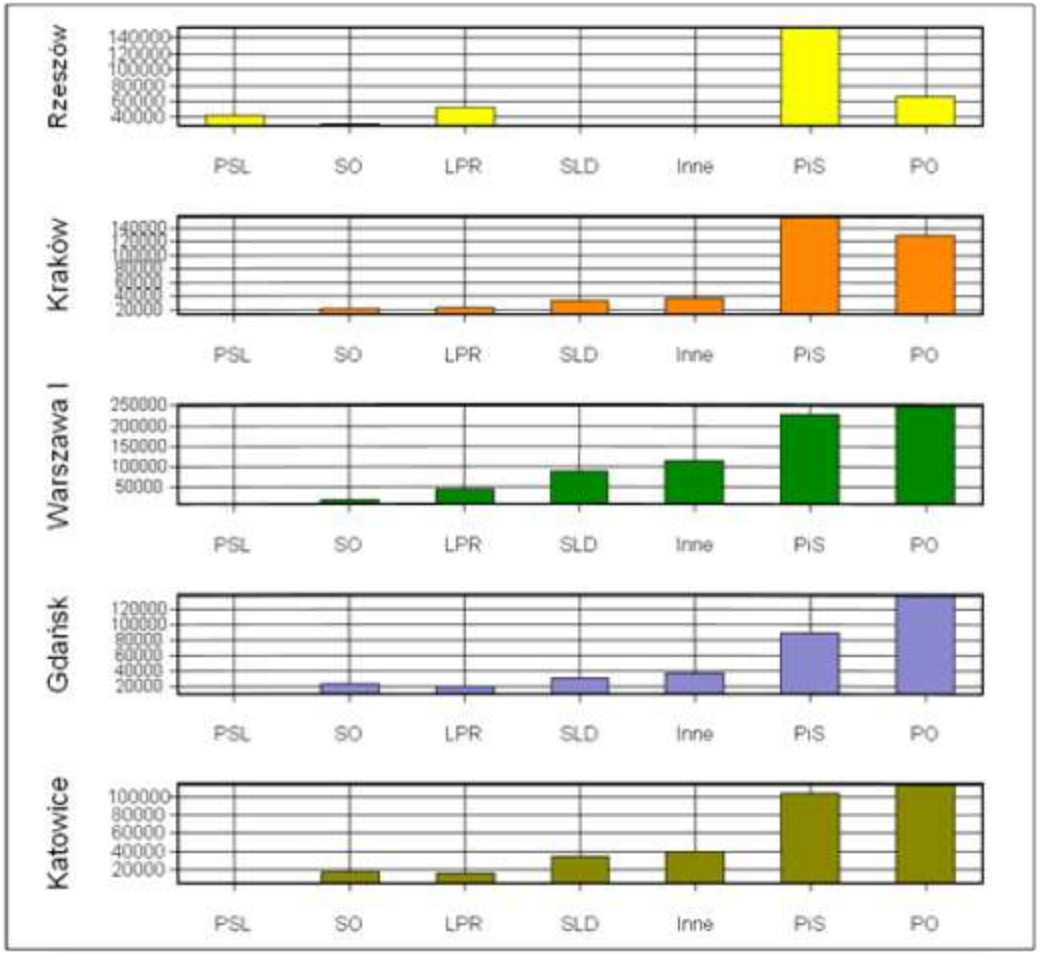
Podział na 6×3 skupień



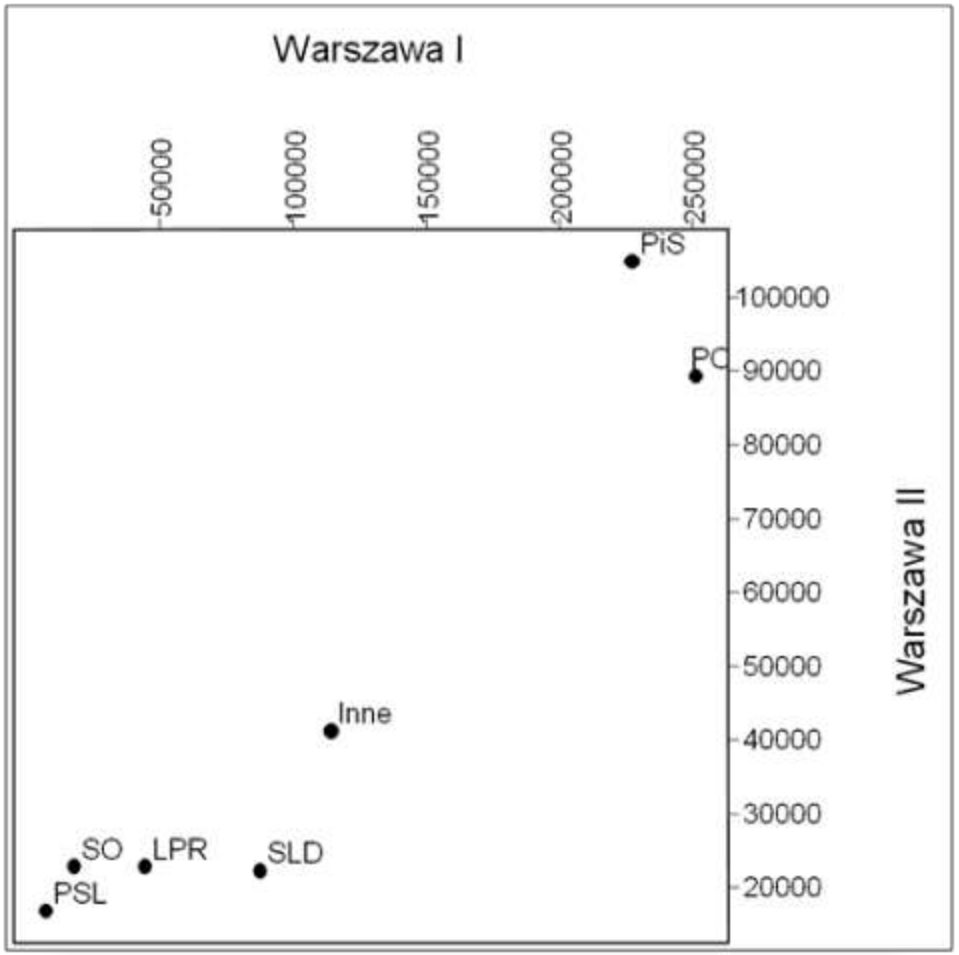
Analiza skupień – agregacja



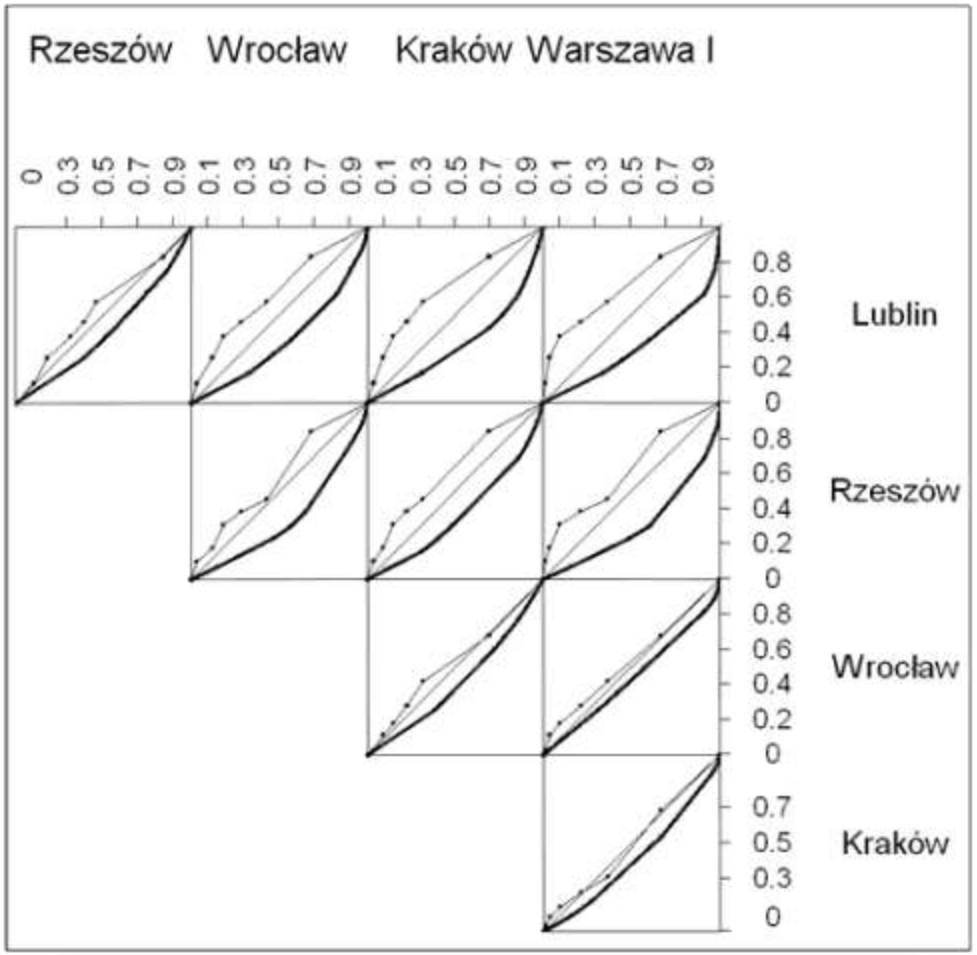
Wizualizacje: wykresy



Wizualizacje: mapy rozrzutu



Wizualizacje: krzywe koncentracji



Metody gradacyjne

- Gradacyjna Analiza Odpowiedniości (GCA)
- Wygładzana Gradacyjna Analiza Odpowiedniości (SGCA)
- uzupełnianie brakujących danych
- wyszukiwanie elementów brakujących i sprawdzanie ich jakości
- wyliczanie wskaźników opartych o ρ* Spearmana, τ Kendalla, Gini, regularność i innych
- znajdowanie wierszy i kolumn odstających od zadanej regularności
- wygładzanie danych po GCA
- gradacyjna analiza skupień i ich agregacja

Metody klasyczne

- wyliczanie współczynników korelacji Pearsona, Spearmana, τ Kendalla
- autokorelacje, korelacje krzyżowe
- odporna analiza składowych głównych (PCA)
- maksymalizowanie korelacji liniowej (pierwszy wymiar analizy odpowiedniości)
- test istotności różnic średnich dwóch populacji (wyliczany empirycznie)

Przegląd danych

- łączenie różnych tabel
- wyłączanie i włączanie do obliczeń wybranych kolumn lub wierszy
- wyszukiwanie wierszy identycznych
- automatyczne obliczanie statystyk pomocniczych dla każdego wiersza i kolumny: brzegów, skumulowanej funkcji rozkładu *(cdf)*, parametrów gradacyjnych, wartości minimalnej i maksymalnej, średniej, odchylenia standardowego
- … i jeszcze dużo więcej

GradeStat w Internecie

Zapraszamy na stronę http://gradestat.ipipan.waw.pl

Można na niej znaleźć:

- najświeższą wersję programu
- przykłady zastosowań
- literaturę
- kilka słów o współtwórcach metod gradacyjnych

Przykłady zastosowań

Przegląd zastosowań – Lingwistyka – Kryptografia – Analiza obrazów medycznych – Dane o żywieniu młodzieży

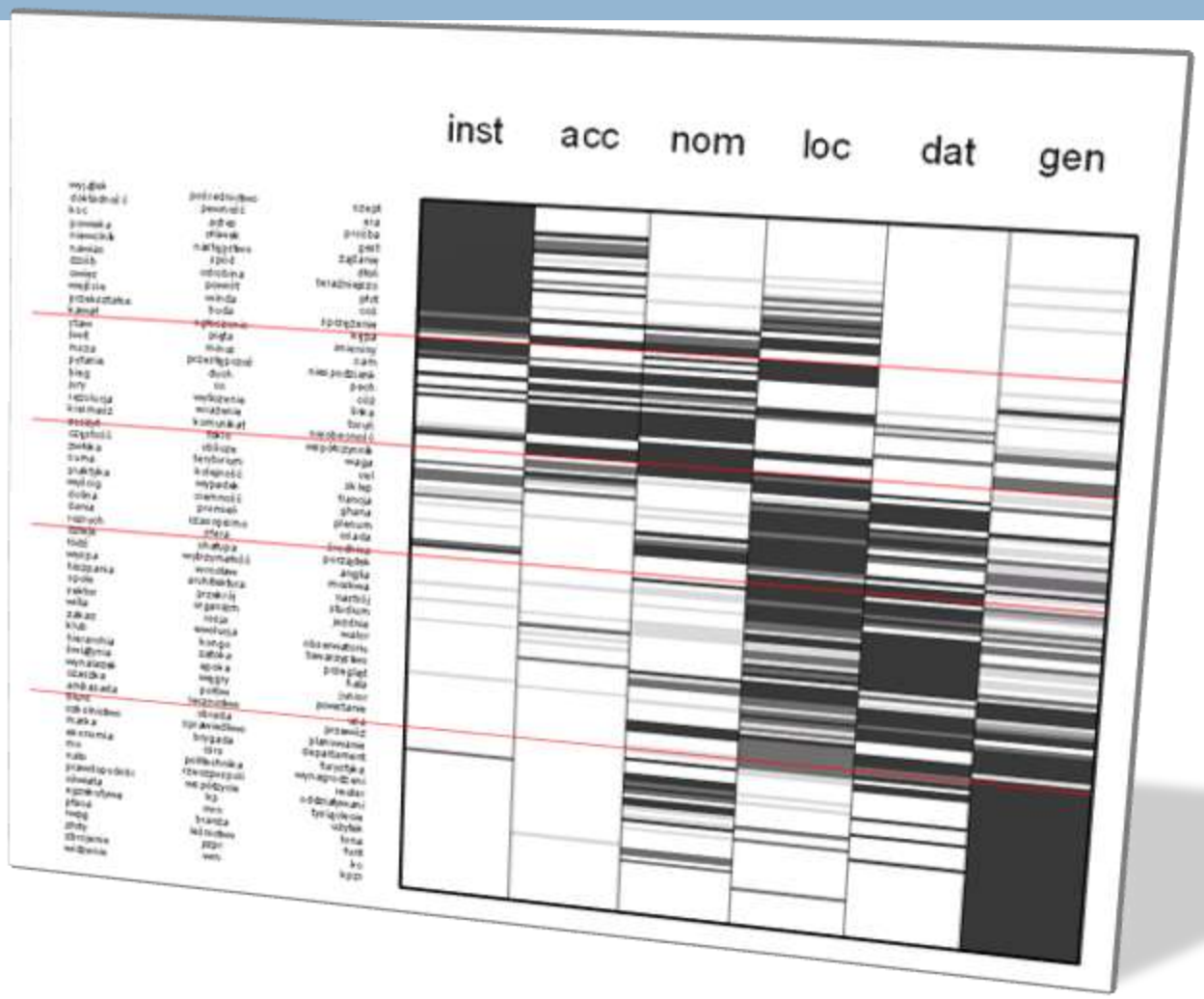
Zastosowania na dzień dzisiejszy

- inżynieria lingwistyczna – analiza danych z korpusów języka polskiego – Hajnicz, Dębowski
- kryptografia (analiza jakości zaszyfrowania) – Srebrny, Such
- analiza obrazów medycznych (rozpoznawanie obrazów) – Grzegorek
- dane o żywieniu młodzieży – Kołłątaj-Dołowy
- gradacyjna analiza skupień i seriacja słów w oparciu o ich współwystępowanie – Jarochowska, Ciesielski
- ankieta dotycząca telepracy osób niepełnosprawnych – Bąkała
- sondaż ekonomiczny EES'2005 – Grabowska, Wiech
- jakość zasobów e–learningowych – Stemposz, Stasiecka

Zastosowania na dzień dzisiejszy

- psychologiczne dane kwestionariuszowe związane z temperamentem i przesądnością – Wiech
- postrzeganie własnego stanu zdrowia (EuroStat) – Pleszczyńska, Jarochowska
- dane o żywieniu niemowląt – książka „Analiza danych medycznych i demograficznych przy użyciu programu GradeStat”
- zawartość biopierwiastków we włosach – Dunicz–Sokolowska
- refundacja leków dla Łódzkiego Oddziału Wojewódzkiego NFZ – Jarochowska
- psychologiczne i medyczne dane kwestionariuszowe związane z objawami nerwicowymi – Welcz
- występowanie określeń emotywnych w internetowych grupach dyskusyjnych (*Usenet*) - Matyja

Gradacyjna analiza danych z korpusów SFPW i IPI PAN



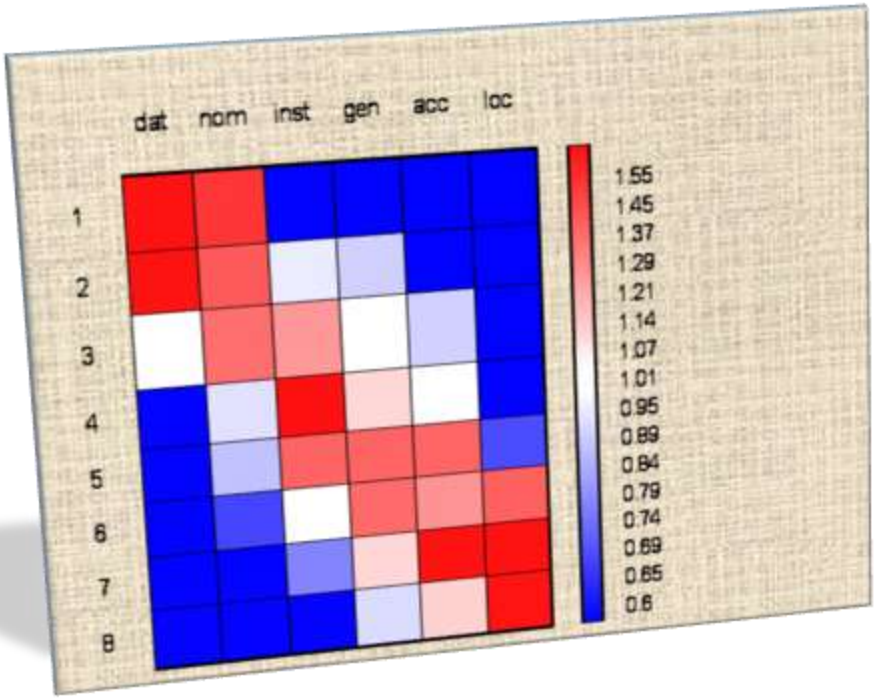
Rzeczowniki – podzbiór OUT

- za ułożeniem rzeczowników wydaje się stać opozycja **rzeczywista informacja - nowomowa**

- **skupienie 1 –** *wyjątek, pośrednictwo, dokładność, adres, prośba, żądanie, wejście, teraźniejszość, ogłoszenie, niewolnik, coś*; częściej występowały w narzędniku

- **skupienie 5 –** *biuro, obrada, sprawiedliwość, planowanie, departament, mo, csrs, turystyka, nato, rzepospolita, reuter, oświata, kp, współżycie, mrn, rwpg, pzpr, złoty, zbrojenie, kc, kpzr*... częściej występowały w dopełniaczu

Skupienia w podzbiorze FIT



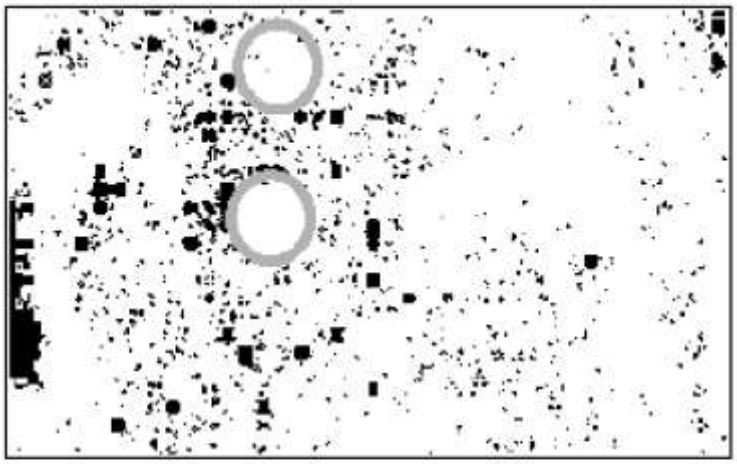
- 1 - pan, pani, siebie, nikt, ojciec, minister, naród
- 2 – człowiek, dziecko, państwo, kobieta, zmiana
- 3 – to, tysiąc, rada, problem, rząd, organizacja, liczba
- 4 – sprawa, wszystko, życie, pomoc, siła
- 5 – praca, nic, oko, woda, szkoła, rzecz
- 6 – kraj, związek, świat, miasto, ręka, warunek, głowa, ziemia
- 7 – raz, dzień, chwila, przykład, droga, dom, polska, sposób, strona
- 8 – rok, czas, miejsce, godzina, okres

Kryptografia

- badanie jakości szyfrów konkretnych tekstów polegało na sprawdzeniu, czy bez dodatkowych informacji możliwe jest **odróżnienie zaszyfrowanego tekstu** od realizacji ciągu niezależnych zmiennych losowych o rozkładzie jednostajnym na odcinku jednostkowym
- całością prac opisanych w raporcie kierował W. Szczesny po zapoznaniu się z literaturą dostarczoną przez M. Srebrnego; W. Szczesny i T. Kowalczyk zaproponowali sposób **przybliżania gradacyjnych miar zróżnicowania dwóch macierzy liczbowych**, co zostało następnie oprogramowane przez O. Matyję i włączone do programu GradeStat
- plan doświadczenia sporządzili W. Szczesny i T. Kowalczyk
- szyfrogramy zostały dostarczone przez P. Sucha i M. Srebrnego
- obliczenia za pomocą programu GradeStat wykonał P. Bielawski
- końcową fazę analizy przeprowadził W. Szczesny posługując się programem GradeStat i Excelem

Analiza obrazów - M. Grzegorek

sąsiedztwo min. 2 pikseli

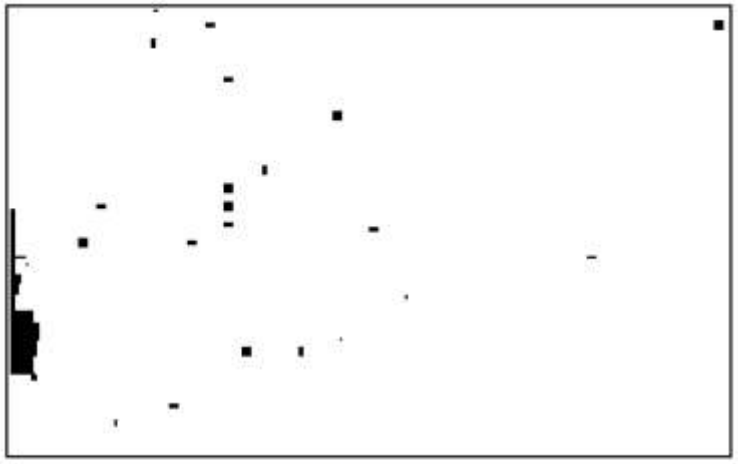


Subimage 3

$(1, 2, 2, 2, 2, 2, 2)$
$\preceq (n_1, n_2, \ldots, n_7) \preceq$
$(8, 8, 8, 8, 8, 8, 8,)$

number of pixels: 5845

sąsiedztwo 8 pikseli

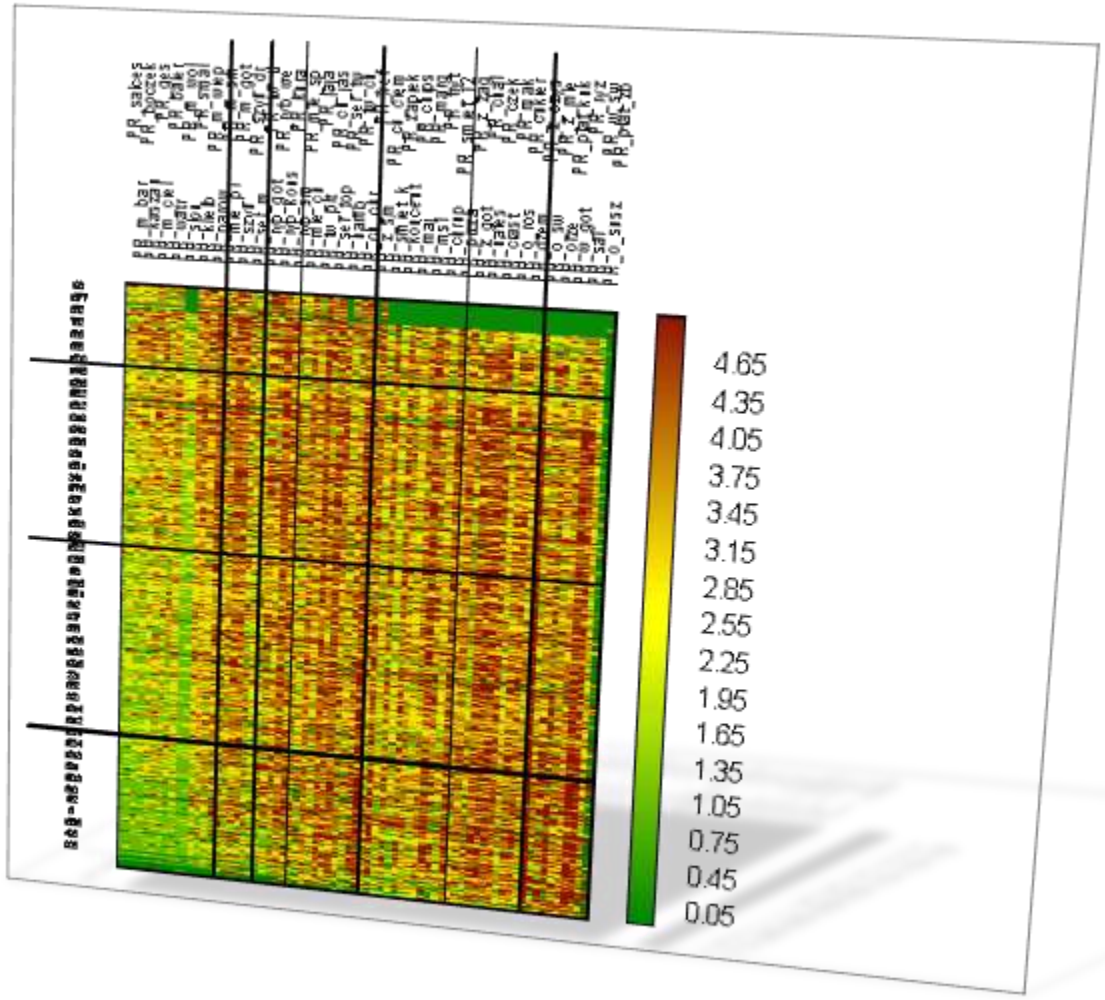


Subimage 4

$(n_1, n_2, \ldots, n_7) = (8, 8, 8, 8, 8, 8, 8)$
number of pixels: 635

Żywienie młodzieży

Preferencje żywnościowe młodzieży w wieku 15-19 lat



Preferencje żywieniowe młodzieży

- skupienie 1 – **mięso i tłuste**, osoby te lubią tradycyjną, tłustą i niezdrową kuchnię; przypuszczalnie spożywają duże objętości jedzenia, polane tłuszczem
- skupienie 2 – „**fast-food" i węglowodany**, osoby te preferują bardziej od innych zapiekanki, pizzę, chrupki, kurczaki, ziemniaki, margaryny i masła
- skupienie 3 – **fitness**, osoby lubiące bardziej chleb chrupki i ciemny, kefiry, bliższe wegetarianom, ale lubiące też ryby i niektóre wędliny, jednak preferujące również „fast-food"; jest to połączenie jedzenia „zdrowego" i „niezdrowego"; być może są to osoby starające się jeść zdrowo, ale niezbyt to lubiące
- skupienie 4 – **wegetarianie**, osoby bardzo nie lubiące mięs i tłuszczy, niezbyt ryby, mniej od innych lubią jajka, zaś znacznie bardziej płatki zbożowe i kukurydziane, także ryż, świeże warzywa, sałatki, owoce suszone

Dziękujemy za uwagę

- Zapraszamy raz jeszcze na stronę: http://gradestat.ipipan.waw.pl

- Zespół Analizy i Modelowania Statystycznego