POLITECHNIKA ŁÓDZKA

WYDZIAŁ FIZYKI TECHNICZNEJ, INFORMATYKI I MATEMATYKI STOSOWANEJ

Kierunek: Matematyka
Specjalność: Matematyczne Metody Analizy Danych Biznesowych

WYBRANE ZASTOSOWANIE STATYSTYCZNYCH METOD PORZĄDKOWANIA DANYCH WIELOWYMIAROWYCH

Kamila Choja
Nr albumu: 204052

Praca licencjacka
napisana w Instytucie Matematyki Politechniki Łódzkiej

Promotor: dr, mgr inż. Piotr Kowalski

ŁÓDŹ, LIPIEC 2018

Spis treści

Rozdział 1

Wstęp

Wielowymiarowa analiza danych jest istotnym pojęciem we współczesnej analizie. Pośród wielu zadań z tej dziedziny, w tej pracy chcemy skupić się na zadaniu porządkowania danych, tj. wskazywaniu uporządkowania o biektów, reprezentowanych przez wielowymiarowe dane. Przedstawione w pracy rozwiązania mogą mieć zastosowanie do wielu problemów napotykanych w codziennej pracy analityków. Do poprawnego zrozumienia zagadnienia statystycznego porządkowania danych należy jednak zgromadzić wiedzę i teorię z wielu o bszarów matematyki i statystyki o raz zaprezentować ich zastosowanie w praktycznych przykładach. Takie zgrupowanie wyżej wskazanych zagadnień jest głównym zadaniem niniejszej pracy.

Praca została uporządkowana w 5 rozdziałach. Rozdział 1 stanowi bieżący wstęp. Z kolei rozdział 2 zawiera potrzebne teorie różnych działów matematyki i statystyki, które są wykorzystywane w kolejnych częściach pracy. o mówione są w nim także zagadnienia dotyczące podstaw rachunku prawdopodobieństwa o raz statystyki (sekcja 2.3), niezbędne do rozumienia wielowymiarowych danych jako losowej próby prostej pewnej wielowymiarowej zmiennej losowej o raz celem wprowadzenia jednolitych o znaczeń. Znaczną część tego rozdziału poświęcamy matematycznej teorii porządków - tak liniowych jak i częściowych, albowiem algorytmy prezentowane w dalszych rozdziałach nawiązują do teorii porządków, definiowanej w ramach współczesnej teorii mnogości. Z uwagi na fakt, iż nie wszystkie porządki wydobywane z danych są porządkami liniowymi, o mawiamy również podstawy o raz wybrane elementy z teorii grafów. Porządki częściowe mogą być bowiem prezentowane na strukturach grafowych ze znakomitą korzyścią dla przejrzystości. o prócz tego w rozdziale 2 zawarliśmy też elementy teorii przestrzeni metrycznych, gdyż o dległości pomiędzy wektorami danych są istotnym elementem w prawie każdym z omawianych algorytmów. Tutaj zawarte zostały również o pisy podstawowych przekształceń na zbiorach danych, wykorzystywanych na etapie wstępnego ich przetwarzania.

Kluczowe teorie dotyczące samego porządkowania danych statystycznych zgromadzone są w rozdziale 3. W tej części pracy o mawiane są również najistotniejsze sposoby wydobywania porządków z danych statystycznych, prezentowana jest ich systematyka o raz o mawiane są właściwości o raz o dmiany. Podstawowy podział na: metody porządkowania liniowego - pozwalające o kreślić ukryte porządki liniowe, o raz metody nieliniowe - służące do wskazywania grafowych reprezentacji o dkrytych porządków częściowych, rozdziela dwie główne sekcje tego rozdziału. Z kolei w rozdziale 4 przedstawiamy eksperymenty przeprowadzone celem lepszego zaprezentowania treści rozdziału 3. Dla potrzeb tej pracy został wytworzony zbiór wielowymiarowych danych - reprezentujący pewien podzbiór o głoszeń znaczącego portalu z ofertami sprzedaży pojazdów. Ponadto w rozdziale tym o pisujemy własne implementacje wybranych algorytmów z rozdziału 3 o pracowane w języku R i prezentujemy uzyskane porządki. W rozdziale 5 zawarte jest treściwe podsumowanie, zarówno z zakresu o pisu statystycznych algorytmów porządkowania danych jak i dla wyników przeprowadzonych eksperymentów.

Rozdział 2

Preliminaria

2.1 Notacja

Poniżej znajduje się lista pojęć powszechnie używanych w pracy wraz z symbolami, które się im przypisuje.

- $\mathbb{R}$ - zbiór liczb rzeczywistych,
- $\mathbb{N}$ - zbiór liczb naturalnych,
- K - o znaczenie dowolnego ciała zbioru,
- $O = \{\mathrm{O}_1, \mathrm{o}\ _2, \ldots, \mathrm{o}\ _n\}$ - zbiór o biektów przestrzennych, tj. o pisywanych przez wiele atrybutów, $n \in \mathbb{N}$,
- $X = [x_{ij}]$ - macierz surowych danych, gdzie x_{ij}-oznacza wartość j-tej zmiennej dla i-tego o biektu, gdzie: $i = 1, \ldots, n$, $j = 1, \ldots, m$,, $n, m \in \mathbb{N}$. W rozdziale drugim, przez X najczęściej będziemy o znaczać dowolny zbiór,
- $N = [n_{ij}]$ - macierz znormalizowanych danych, gdzie n_{ij} o znacza wartość j-tej cechy i-tego o biektu,
- x^S - o znaczenie zmiennej mającej charakter stymulanty,
- x^D - o znaczenie zmiennej mającej charakter destymulanty,
- x^N - o znaczenie zmiennej mającej charakter nominanty,
- $D = [d_{ik}]$ - o znaczenie macierzy o dległości, gdzie d_{ik} o znacza o dległość między i-tym i k-tym o biektem, gdzie: $i, k = 1, \ldots, n$, $n \in \mathbb{N}$,
- s_i - o znaczenie zmiennej syntetycznej i-tego o biektu,
- $P_0 = [n_{0j}]$ - o znaczenie o biektu wzorcowego, gdzie n_{0j}- znormalizowana j-ta współrzędna o biektu wzorcowego,
- Y - w rozdziale drugim używana jest najczęściej do o znaczenie zmiennej losowej,
- Ω - o znaczenie dowolnej przestrzeni zdarzeń elementarnych ω, z rodziną podzbiorów $\mathcal{F}$,
- $\mathcal{B}$ - rodzina zbiorów borelowskich,
- $\mathfrak{B}$ - rodzina wszystkich zbiorów o twartych,

- $\leqslant$ - o znaczenie relacji częściowego porządku, przy dołożeniu warunku spójności o znaczać będzie relację liniowego porządku,
- $\mathrm{med}\,(\cdot)$ - o znaczenie mediany zbioru,
- $\max\,(\cdot)$ - o znaczenie maksymalnej wartości zbioru,
- $\min\,(\cdot)$ - o znaczenie minimalnej wartości zbioru,
- $\overline{\overline{(\cdot)}}$ - o znaczenie mocy zbioru,
- G - o znaczenie o gólnego grafu prostego dla którego $V(G)$ jest zbiorem wierzchołków grafu, a $E(G)$ zbiorem jego krawędzi,

2.2 Słownik użytych pojęć

W pracy zostały wykorzystane następujące pojęcia:

- Statystyka matematyczna [6, w o parciu o rozdział 1]
  Statystyka matematyczna jest nauką zajmującą się o pisywaniem i analizą zjawisk przy użyciu metod rachunku prawdopodobieństwa.
- Cecha statystyczna [6, Rozdział 1]
  Cecha statystyczna jest to właściwość wspólna dla danego zbioru o bserwacji. Jej wartości pozwalają rozróżnić elementy zbioru między sobą. cechy statystyczne można podzielić na te mierzalne, tj. ilościowe (np. długość, ciężar), o raz niemierzalne tj. jakościowe (np. kolor, płeć, zawód, województwo).

W celu prezentacji dużych ilości danych, w analizie danych korzysta się z pojęcia macierzy. Poniżej zostanie przedstawiona formalna definicja macierzy o raz definicja macierzy o bserwacji, czyli zbiorze o biektów o pisywanych przez zmienne.

Definicja 2.2.1. *Macierz [1, Rozdział 1]*
Niech K będzie ciałem i $m, n \in \mathbb{N}$. Macierzą o m-wierszach, n-kolumnach i o wyrazach z K nazywamy każdą funkcję postaci $A : \{1, \ldots, m\} \times \{1, \ldots, n\} \to K$

Przykład 1. *Macierz A o m-wierszach i n-kolumnach najczęściej zapisuje się postaci $A = [a_{ij}]_{1\leqslant i\ \leqslant m, 1\leqslant j\leqslant n}$, tj.*

$$A = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{bmatrix}$$

Uwaga 1. *W statystyce koncepcja matematyczna macierzy jest rozszerzana, gdyż niektóre kolumny mają wartości z poza ciała zbioru liczb $\mathbb{R}$ (mogą być np. tekstem).*

Definicja 2.2.2. *Macierz o bserwacji [8, Rozdział 2]*
Niech $m > 1$ o raz $n > 1$ będą liczbami naturalnymi. Macierzą o bserwacji nazywamy macierz rozmiaru $n \times m$ postaci

$$X = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1m} \\ x_{21} & x_{22} & \ldots & x_{2m} \\ \ldots & \ldots & \ldots & \ldots \\ x_{n1} & x_{n2} & \ldots & x_{nm} \end{bmatrix}$$

gdzie: x_{ij} - zaobserwowana wartość j-tej cechy dla i-tego o biektu.

Definicja 2.2.3. *Macierz o dległości zmiennych [9, Rozdział 1.6]*
Macierzą o dległości cech zmiennych nazywamy macierz, której elementami są o dległości między parami badanych o biektów:

$$D = [d_{ik}].$$

gdzie:
d_{ik}*-odległość między i-tym a k-tym o biektem, dla* $i, k = 1, 2, \ldots, n$

Uwaga 2. *Zauważmy, że macierz o bserwacji nie musi być symetryczna, z kolei w przypadku macierzy o dległości jest to wymagane.*

W statystyce posługujemy się pojęciami skal do o pisu różnych typów danych, które przyjmowane przez nas mogą podlegać analizie. W związku z powyższym zdefiniujmy następujące rodzaje skal:

- Skala porządkowa [9, Rozdział 1.2]
  Zmienna o pisana jest na skali porządkowej jeśli jej zbiór wartości jest zbiorem, w którym wprowadzony jest porządek np. porządek liczb. Nie zawsze porządek ten jest ustalony w sposób matematyczny. W przypadku rozpatrywania zmiennych jakościowych, porządek ustalamy na podstawie o pinii ekspertów lub o gólnie przyjętych poglądów np. poziom wykształcenia, o ceny w systemie szkolnym.

- Skala przedziałowa [9, Rozdział 1.2]
  Zmienna jest o pisana na skali przedziałowej gdy podobnie jak na skali porządkowej jej zbiór wartości jest zbiorem uporządkowanym z wprowadzoną funkcją o dległości. Dodatkowo na skali tej możliwe jest wyznaczenie umownego punktu - zera. Przykład zmiennych przedstawianych na skali przedziałowej to: temperatura, czas.

- Skala ilorazowa [9, Rozdział 1.2]
  Zmienna jest o pisana na skali ilorazowej, jeśli jej zbiór wartości jest zbiorem postaci $[0, \infty)$ będący podzbiorem zbioru liczb $\mathbb{R}$, lub też taki zbiór wartości który można utożsamić z podzbiorem liczb $\mathbb{R}$. Przykłady zmiennych o pisywanych na skali ilorazowej: napięcie elektryczne, bezrobocie.

Uwaga 3. *Jeżeli zmiennej o pisującej dany o biekt nie da się o dnieść do żadnej z powyższych skal, to zmienna ta nazywana jest nominalną.*

Ze względu na to, że zmienne o pisujące o biekty mogą mieć różny charakter, poniżej zostały wprowadzone definicje trzech różnych typów zmiennych, którymi posługujemy się w statystyce.

Definicja 2.2.4. *Stymulanta [9, Rozdział 1.5]*
Stymulantami nazywane są te zmienne (cechy), dla których pożądane są wysokie wysokie wartości w badanych o biektach, ze względu na rozpatrywane zjawisko.

Definicja 2.2.5. *Destymulanta [9, Rozdział 1.5]*
Destymulantami nazywane są te zmienne, dla których niepożądane są wysokie wysokie wartości w badanych o biektach, ze względu na rozpatrywane zjawisko.

Definicja 2.2.6. *Nominanta [9, Rozdział 1.5]*
Nominantami nazywane są te zmienne, które mają o kresloną najkorzystniejszą wartość. o dchylenia o d tej wartości są niepożądane, ze względu na rozpatrywane zjawisko.

2.3 Podstawowe pojęcia rachunku prawdopodobieństwa o raz statystyki

Na potrzeby pracy, zostały wykorzystane pojęcia rachunku prawdopodobieństwa o raz statystyki, konieczne do zrozumienia danych jako próby losowej. W tym celu niezbędne było wprowadzenie definicji prawdopodobieństwa, zmiennej losowej, a także pojęć powiązanych z tymi definicjami tj. ciała zbiorów, σ-ciała zbiorów, przestrzeni zdarzeń elementarnych, zdarzenia losowego.

Definicja 2.3.1. *Ciało zbiorów [10, Rozdział 8.1]*
Rodzinę $\mathcal{F}$ podzbiorów, niepustego zbioru X nazywamy ciałem zbiorów, jeżeli spełnia o na następujące warunku:

1. $\emptyset \in \mathcal{F}$,
2. *jeżeli* $A \in \mathcal{F}$, *to* $X \setminus A \in \mathcal{F}$,
3. *jeżeli* $A \in \mathcal{F}$, *to* $A \cup B \in \mathcal{F}$.

Definicja 2.3.2. *σ-algebra/ciało zbiorów[10, Rozdział 8.1]*
Ciało zbiorów $\mathcal{F}$ nazywamy σ-ciałem zbiorów, jeżeli spełnia o na warunek dla dowolnych zbiorów $A_n \in \mathcal{F}, n \in \mathbb{N}$, mamy $\bigcup\limits_{i=1}^{\infty} A_n \in \mathcal{F}$.

Najważniejszym σ-ciałem zbiorów w matematyce są σ-ciała zbiorów Borelowskich, dlatego też wprowadzimy definicję zbiorów borelowskich.

Definicja 2.3.3. *Zbiory borelowskie [4, w o praciu o rozdział 2]*
Zbiorami borelowskimi względem danej przestrzeni X, nazywamy zbiory należące do σ-ciała X generowanego przez rodzinę $\mathfrak{B}(X)$ - wszystkich zbiorów o twartych w X. Rodzinę wszystkich zbiorów borelowskich względem X, o znaczamy $\mathcal{B}(X)$.

σ-ciała zbiorów często mogą być po prostu utożsamiane ze zbiorami, które można zmierzyć, w związku z tym wprowadzimy definicję miary zbioru.

Definicja 2.3.4. *Miara zbioru [4, Rozdział 2.10]*
Funkcję μ o kreśloną na ciele $\mathcal{F}$ podzbiorów zbioru Ω nazywamy miarą, jeśli spełnia następujące warunki:

1. $\forall A \in \mathcal{F} \quad \mu(A) \in [0, \infty]$,
2. $\mu(\emptyset) = 0$,
3. *jeśli* $A_1, A_2, ...$ *jest ciągiem rozłącznych zbiorów* $\mathcal{F}$*-mierzalnych takich, że* $\bigcup\limits_{k=1}^{\infty} A_k \in \mathcal{F}$, *to*

$$\mu\Big(\bigcup_{k=1}^{\infty} A_k\Big) = \sum_{k=1}^{\infty} \mu(A_k).$$

Definicja 2.3.5. *Przestrzeń mierzalna [4, Rozdział 2.10]*
Przestrzenią mierzalną nazywamy parę $(X, \mathcal{F})$, gdzie $\mathcal{F}$ jest σ-ciałem podzbiorów zbioru X

Definicja 2.3.6. *Funkcja mierzalna [10, w o parciu o rozdział 8.2]*
Niech X będzie niepustym zbiorem, $\mathcal{F}$ σ-ciałem na X i $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$. Funkcję $f : X \to \overline{\mathbb{R}}$ nazywamy mierzalną, jeżeli zbiór $\{x \in X : f(x) > a\}$ jest mierzalny przy dowolnym $a \in \mathbb{R}$.

Mając powyższe definicje, wprowadźmy możemy wprowadzić poszukiwane definicje rachunku prawdopodobieństwa.

Definicja 2.3.7. *Przestrzeń zdarzeń elementarnych [6, w o parciu o rozdział 1.1]*
Zbiór wszyskich możliwych wyników doświadczenia losowego nazywamy przestrzenią zdarzeń elementarnych i oznaczamy przez Ω. Elementy zbioru Ω nazywamy zdarzeniami elementarnymi i oznaczamy ω.

Definicja 2.3.8. *Zdarzenie losowe [6, w o parciu o rozdział 1.1]*
Zdarzeniem losowym (zdarzeniem) nazywamy każdy podzbiór A zbioru Ω, taki że $A \in \mathcal{F}$, gdzie $\mathcal{F}$ jest rodziną podzbiorów Ω spełniającą następujące warunki:

1. *$\Omega \in \mathcal{F}$,*
2. *Jeśli $A \in \mathcal{F}$, to $A' \in \mathcal{F}$, gdzie $A' = \Omega \setminus A$ jest zdarzeniem przeciwnym do zdarzenia A,*
3. *Jeśli $A_i \in \mathcal{F}$, $i = 1, 2, ...$,to $\bigcup\limits_{i=1}^{\infty} A_i \in \mathcal{F}$.*

Rodzinę $\mathcal{F}$ spełniającą warunki 1 - 3 nazywamy σ-ciałem podzbiorów zbioru Ω

Definicja 2.3.9. *Prawdopodobieństwo [6, w o parciu o rozdział 1.1]*
Prawdopodobieństwem nazywamy dowolną funkcję P o wartościach rzeczywistych, o kreśloną na σ-ciele zdarzeń $\mathcal{F} \subset 2^{\Omega}$, spełniającą warunki:

1. *$\forall A \in \mathcal{F} \quad P(A) \geqslant 0$,*
2. *$P(\Omega) = 1$,*
3. *Jeśli $A_i \in \mathcal{F}$, $i = 1, 2, ...$ o raz $A_i \cap A_j = \emptyset$ dla $i \neq j$, to*

$$P\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} P(A_i).$$

Definicja 2.3.10. *Przestrzeń probabilistyczna [6, w o parciu o rozdział 1.2]*
Przestrzenią probabilistyczną nazywamy uporządkowaną trójkę $(\Omega, \mathcal{F}, P)$, gdzie Ω jest zbiorem zdarzeń elementarnych, $\mathcal{F}$ jest σ-ciałem podzbiorów Ω, zaś P jest prawdopodobieństwem o kreślonym na $\mathcal{F}$.

Dla tak podanej definicji prawdopodobieństwa, definiujemy:

Definicja 2.3.11. *Zmienna losowa [6, Rozdział 2.1]*
Niech $(\Omega, \mathcal{F}, P)$ będzie dowolną przestrzenią probabilistyczną. Dowolną funkcję $Y : \Omega \to \mathbb{R}$ nazywamy zmienną losową jednowymiarową, jeśli dla dowolnej liczby rzeczywistej y zbiór zdarzeń elementarnych ω, dla których spełniona jest nierówność $Y(\omega) < y$ jest zdarzeniem, czyli

$$\{\omega : Y(\omega) < y\} \in \mathcal{F} \quad \forall y \in \mathbb{R}.$$

Definicja 2.3.12. *Wektor losowy [5, Rozdział 5.1]*
Wektorem losowym nazywamy o dwzorowanie $Y : \Omega \to \mathbb{R}^n$, spełniające następujący warunek: dla każdego układu liczb $t_1, t_2, \ldots, t_n \in \mathbb{R}^n$ zbiór $Y^{-1}((-\infty, t_1] \times \ldots \times (-\infty, t_n])$ należy do $\mathcal{F}$.

Definicja 2.3.13. *Rozkład prawdopodobieństwa zmiennej losowej [5, Rozdział 5.1]*
Rozkładem prawdopodobieństwa zmiennej losowej Y o wartościach w $\mathbb{R}$ nazywamy funkcję μ_Y o kreśloną na $\mathcal{B}(\mathbb{R})$ zależnością:

$$\mu_Y(B) = P_Y(B) = P(Y^{-1}(B)), \quad B \in \mathcal{B}(\mathbb{R}).$$

Definicja 2.3.14. *Rozkład dyskretny [5, Rozdział 5.1]*
Mówimy, że zmienna losowa jednowymiarowa Y ma rozkład dyskretny, jeśli istnieje przeliczany zbiór $S \subset \mathbb{R}$, taki że $\mu_Y(S) = 1$.

Definicja 2.3.15. *Gęstość i rozkład ciągły [5, Rozdział 5.1]*
Jeśli μ jest rozkładem prawdopodobieństwa na $\mathbb{R}$ i istnieje całkowalna funkcja $f : \mathbb{R} \to \mathbb{R}$ taka, że:

$$\mu(A) = \int_A f(y)dy, A \in \mathcal{B}(\mathbb{R})$$

to funkcję f nazywamy gęstością rozkładu μ. Rozkład, który ma gęstość, nazywamy rozkładem ciągłym.

Definicja 2.3.16. *Wartość o czekiwana [6, Rozdział 2.6]*
Niech X będzie zmienną losową typu dyskretnego lub ciągłego. Wartością o czekiwaną zmiennej losowej Y nazywamy:

$$\mathrm{E}(Y) = \begin{cases} \sum\limits_{i=1}^{n} y_i p_i, & \textit{jeśli zmienna ma rozkład dyskretny i przyjmuje dokładnie } n \textit{ wartości,} \\ \sum\limits_{i=1}^{\infty} y_i p_i, & \textit{jeśli zmienna przyjmuje nieskończenie, ale przeliczalnie wiele wartości,} \\ \int\limits_{-\infty}^{\infty} y f(y) dy, & \textit{jeśli zmienna ma rozkład ciągły.} \end{cases}$$

Definicja 2.3.17. *Wariancja [5, Rozdział 5.6]*
Niech Y będzie zmienną losową. Jeśli $\mathrm{E}(Y - \mathrm{E}Y)^2 < \infty$, to liczbę tę nazywamy wariancją zmiennej losowej Y o wartościach rzeczywistych i oznaczamy:

$$\mathrm{Var}\, Y = \mathrm{E}(Y - \mathrm{E}Y)^2.$$

Definicja 2.3.18. *Odchylenie standardowe [5, Rozdział 5.6]*
Niech Y będzie zmienną losową. o dchyleniem standardowym zmiennej losowej Y nazywamy pierwiastek z wariancji:

$$\sigma_Y = \sqrt{\mathrm{Var}\, Y}.$$

Definicja 2.3.19. *Próba losowa n-elementowa [2, Rozdział 2]*
Niech $(\Omega, \mathcal{F}, P)$ będzie przestrzenią probabilistyczną, zaś $Y_i : \Omega^n \to \mathbb{R}$ będzie zmienną losową. Próbą losowa n-elementową nazywamy n-elementowy ciąg niezależnych zmiennych losowych o jednakowym prawdopodobieństwie, tzn. ciąg postaci $(\omega_1, \omega_2, \ldots, \omega_n) \in \Omega^n$, taki że $Y_i(\omega_1, \omega_2, \ldots, \omega_n) = Y(\omega_i)$.

Uwaga 4. *Powyższa definicja próby losowej jest wyrażona w języku rachunku prawdopodobieństwa. W języku statystyki próbę losową rozumiemy jako wartość pochodząca z realizacji takiego doświadczenia.*

Definicja 2.3.20. *Rozkład normalny (Gaussa) [5, Rozdział 5.10]*
Jeśli zmienna losowa Y ma gęstość postaci

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(y-\mu_Y)^2}{2\sigma^2}}$$

dla $y \in \mathbb{R}$ i pewnych $\mu_Y \in \mathbb{R}$ i $\sigma^2 > 0$. To mówimy, że zmienna losowa ma rozkład normalny z parametrami μ i σ^2, co zapisujemy $\mathcal{N}(\mu, \sigma^2)$.
W przypadku, gdy $\mu = 0$ i $\sigma^2 = 1$, to rozkład ten nazywamy standardowym rozkładem normalnym i oznaczamy $\mathcal{N}(0, 1)$, a gęstość jest postaci

$$f(y) = \frac{1}{\sqrt{2\pi}} e^{\frac{-(y)^2}{2}}.$$

2.4 Wybrane o peracje statystyczne dla zmiennych

Normalizacja zmiennych

Ważnym krokiem przed rozpoczęciem pracy na zbiorze danych jest ujednolicenie ich charakteru, tj. przekształcenie zmiennych (mierzonych na skali przedziałowej lub ilorazowej) o pisujących o biekty w zbiorze, w celu pozbycia się dysproporcji między nimi czy też dominacji jednych zmiennych nad drugimi. W tym celu stosuje się transformację normalizacyjną. Wyróżniamy trzy podstawowe typy przekształceń normalizacyjnych:

- standaryzacja,
- unitaryzacja,
- przekształcenie ilorazowe,
- rangowanie zmiennych.

W dalszej części pracy j-ta zmienna znormalizowana, i-tego o biektu jest o znaczona jako n_{ij}. Dodatkowo przyjmujemy o znaczenia: $\overline{x}$ - średnia z próby, $\sigma(x)$ - o dchylenie standardowe z próby.

1. W wyniku standaryzacji zmienne uzyskują o dchylenie standardowe równe 1 i wartości o czekiwanej równą 0. W tym celu dla każdej zmiennej, będącej cechą o biektu o blicza się o dchylenie standardowe na podstawie wartości tej zmiennej dla wszystkich o biektów, a także wartość o czekiwaną na tej samej zasadzie. W kolejnym kroku dla każdego o biektu liczymy jego znormalizowaną wartość tj. o d wartości zmiennej o dejmujemy średnią wartość danej cechy, a o trzymaną różnicę dzielimy przez o dchylenie standardowe dla tej cechy. Powyższy o pis można zapisać w postaci:

$$n_{ij} = \frac{x_{ij} - \overline{x_j}}{\sigma(x_j)}, \quad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, m.$$

2. W pracy korzystamy również z unitaryzacji, stosowanej w celu uzyskania zmiennych o ujednoliconym zakresie zmienności, najczęściej jest to przedział $[0, 1]$. W tym celu o d wartości zmiennej w obiekcie, o dejmowana jest minimalna wartość występująca dla tej cechy, a następnie różnica ta dzielona jest przez różnicę między maksymalną a minimalną wartością zmiennej, która została poddana normalizacji. Znormalizowana zmienna jest postaci:

$$n_{ij} = \frac{x_{ij} - \min\limits_i(x_{ij})}{\max\limits_i(x_{ij}) - \min\limits_i(x_{ij})}, \quad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, m.$$

3. Kolejna metoda normalizacyjna, wykorzystana w pracy to przekształcenie ilorazowe. Stosuje się je aby o dnieść wartości zmiennej do ustalonej wartości - może to być wartość o czekiwana danej zmiennej na tle analizowanych o biektów, wartość minimalna lub maksymalna tej cechy. W pracy za tę wartość przyjęliśmy średnią wartość zmiennej. Każda wartość zmiennej dla danego o biektu jest dzielona przez wartość o czekiwaną tej zmiennej, a postać znormalizowanej zmiennej to:

$$n_{ij} = \frac{x_{ij}}{\overline{x_j}}, \quad \overline{x_j} \neq 0 \quad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, m.$$

4. Przy zastosowaniu metody rang wykorzystuje się normalizację rangową. Przekształcenie to, najczęściej stosowane jest, gdy zmienne o pisujące o biekty są wyrażone na skali porządkowej. W pierwszym kroku wartości zmiennych o pisujących o biekty, zostają uporządkowane ze względu na ich wartości po procesie normalizacji. W kolejnym kroku wartościom zmiennej przyporządkowywane są rangi - czyli wartości liczbowe, będące najczęściej numerami miejsc zajmowanych przez o biekty w uporządkowanym zbiorze. Postać zmiennej znormalizowanej rangowo:

$$n_{ij} = r, \quad dla \quad x_{hj} = x_{ij}, \quad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, m.$$

gdzie:
r-ranga nadana i-temu o biektowi znajdującemu się na r-tym miejscu w uporządkowanym zbiorze, ze względu na wartość j-tej zmiennej.

Stymulacja zmiennych

W celu ujednolicenia charakteru zmiennych należy poddać je pewnym przekształceniom, polegającym na zamianie destymulant i nominant na stymulanty. Tego typu transformacje nazywamy stymulacją. Można wyróżnić dwie najczęściej stosowane metody, tj. przekształcenie ilorazowe o raz przekształcenie różnicowe. W zależności o d skali, na której mierzone są zmienne, należy stosować o dpowiednie przekształcenie stymulacyjne.

Przekształcenie ilorazowe można stosować tylko dla zmiennych mierzonych na skali ilorazowej. Poniżej zaprezentujemy jego postać dla zmiennych o charakterze destymulant o raz nominant.

- Niech x^D o znacza dane o charakterze destymulant, wtedy dane x^S wyznaczone wg. poniższego wzoru, mają charakter stymulant:

$$x_{ij}^S = [x_{ij}^D]^{-1}, \quad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, m,$$

- Analogicznie do powyższego, niech x^N o znaczają dane o charakterze nominant, wtedy dane x^S wyznaczone wg. poniższego wzoru, mają charakter stymulant:

$$x_{ij}^S = \frac{\min\{x_j^N, x_{ij}^N\}}{\max\{x_j^N, x_{ij}^N\}}, \quad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, m.$$

gdzie:
x_j^N - o znacza pożądaną wartość j-tej zmiennej,
x_{ij}^N - o znacza wartość j-tej zmiennej w i-tym o biekcie.

Dla zmiennych mierzonych na skali ilorazowej czy też przedziałowej stosuje się przekształcenie różnicowe. Postać tego przekształcenia dla zmiennych o charakterze destymulant o raz nominant jest następująca.

- Ponownie niech x^D o znacza dane o charakterze destymulant, wtedy dane x^S wyznaczone wg. poniższego wzoru, mają charakter stymulant:

$$x_{ij}^S = \max_i\{x_{ij}^D\} - x_{ij}^D, \quad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, m,$$

- Niech teraz x^N o znaczają dane o charakterze nominant, wtedy dane x^S wyznaczone wg. poniższego wzoru, mają charakter stymulant:

$$x_{ij}^S = -|x_{ij}^N - x_j^N|, \quad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, m.$$

2.5 Podstawowe pojęcia teorii grafów

W pracy zostaną o pisane zarówno metody porządkowania liniowego jak i nieliniowego. W tym celu należy wprowadzić definicje związane z teorią grafów, niezbędne przy o pisywaniu metod porządkowania nieliniowego.

W celu wprowadzenia kluczowych definicji, należy wcześniej podać podstawowe pojęcia dotyczące grafów. Zacznijmy o d wprowadzenia definicji pary uporządkowanej o raz nieuporządkowanej, gdyż pojęcia te zostały wykorzystane w definicji grafu.

Definicja 2.5.1. *Para uporządkowana [7, w o parciu o rozdział 3]*
Niech X będzie dowolnym, niepustym zbiorem o raz niech dane będą dwa elementy $a, b \in X$. Parą uporządkowaną nazywamy parę postaci $< a, b >$, gdzie element a jest poprzednikiem, zaś element b jest następnikiem: $< a, b >= \{\{a\}, \{a, b\}\}$.

Definicja 2.5.2. *Para nieuporządkowana [7, w o parciu o rozdział 3]*
Niech X będzie dowolnym, niepustym zbiorem o raz niech dane będą dwa elementy $a, b \in X$. Parą nieuporządkowaną nazywamy zbiór postaci $\{a, b\}$, tj. zawierający elementy a i b i nie zawierający żadnego innego elementu. W przypadku, gdy $a = b$, to para nieuporządkowana $\{a, b\}$, składa się dokładnie z jednego elementu.

W analogiczny sposób jak parę dwóch punktów można wprowadzić parę dwóch zbiorów. Istotne jest aby podkreślić różnicę pomiędzy parą dwóch wierzchołków, które tworzą krawędź, a parą zbiorów definiującą graf.

Definicja 2.5.3. *Graf [12, w o parciu o rozdział 2]*
Grafem nazywamy parę $G = (V, E) = (V(G), E(G))$, gdzie V jest niepustym, skończonym zbiorem wierzchołków grafu G, zaś E jest skończonym podzbiorem zbioru nieuporządkowanych par elementów zbioru V.

Definicja 2.5.4. *Pętle [12, Rozdział 2]*
Niech $G = (V(G), E(G))$ będzie grafem o raz niech $a \in V(G)$. Pętlami w grafie nazywamy krawędzie reprezentowane przez a, tj. łączące wierzchołek z samym sobą. Innymi słowy jest to para nieuporządkowana składająca się z jednego elementu.

Definicja 2.5.5. *Wierzchołki sąsiednie [12, Rozdział 2]*
Mówimy, że dwa wierzchołki v i w grafu G są sąsiednie, jeśli istnieje krawędź vw która je łączy.

$$v \text{ ——— } w$$

Analogicznie definiuje się krawędzie sąsiednie.

Definicja 2.5.6. *Krawędzie sąsiednie [12, Rozdział 2]*
Dwie krawędzie e i f grafu G są sąsiednie, jeśli mają wspólny wierzchołek, tj

$$\exists d \in V(G) \quad d \in e \wedge d \in f$$

$$\overset{e}{\text{———}}\, d \,\overset{f}{\text{———}}$$

Aby dowiedzieć się o połączeniu dwóch wierzchołków w grafie, wprowadzimy pojęcie trasy.

Definicja 2.5.7. *Trasa/marszruta [12, Rozdział 3]*
Trasą (lub marszrutą) w danym grafie G nazywamy skończony ciąg krawędzi postaci $v_0v_1, v_1v_2, \ldots, v_{m-1}v_m$, zapisywany również w postaci $v_0 \to v_1 \to v_2 \to \ \ldots \to v_m$, w którym każde dwie kolejne krawędzie są albo sąsiednie, albo identyczne. Taka trasa wyznacza ciąg wierzchołków $v_0, v_1, \ldots, v_m$. Wierzchołek v_0 nazywamy wierzchołkiem początkowym, a wierzchołek v_m wierzchołkiem końcowym trasy, mówimy też wtedy, o trasie o d wierzchołka v_0 do wierzchołka v_m. Liczbę krawędzi na trasie nazywamy długością trasy.

Definicja 2.5.8. *Ścieżka [12, Rozdział 3]*
Trasą, w której wszystkie krawędzie są różne, nazywamy ścieżką.

Definicja 2.5.9. *Droga [12, Rozdział 3]*
Ścieżkę, w której wierzchołki o znaczane kolejno: $v_0, v_1, \ldots, v_m$ są różne (z wyjątkiem, być może, równości $v_0 = v_m$, nazywamy drogą.

Definicja 2.5.10. *Droga zamknięta/ścieżka zamknięta [12, Rozdział 3]*
Droga jest zamknięta gdy rozpoczyna się i kończy w tym samym punkcie, tj. według przyjętej notacji $v_0 = v_m$.

Definicja 2.5.11. *Cykl [12, Rozdział 3]*
Cyklem nazywamy drogę zamkniętą.

Definicja 2.5.12. *Graf spójny [12, Rozdział 3]*
Graf jest spójny wtedy i tylko wtedy, gdy każda para wierzchołków jest połączona drogą.

Definicja 2.5.13. *Drzewo [12, Rozdział 4]*
Drzewem nazywamy graf spójny, nie zawierający cykli.

Definicja 2.5.14. *Graf skierowany(digraf albo graf zorientowany) [12, Rozdział 7]*
Graf skierowany lub digraf D, składa się z niepustego zbioru skończonego $V(D)$ elementów nazywanych wierzchołkami i skończonej rodziny $E(D)$ par uporządkowanych elementów zbioru $V(D)$, nazywanych łukami. Zbiór $V(D)$ nazywamy zbiorem wierzchołków, a rodzinę $E(D)$ rodziną łuków digrafu D (krawędzi grafu skierowanego). Łuk (v, w) zwykle zapisujemy jako vw. Graf skierowany o znaczamy zwykle w postaci pary uporządkowanej $G =< V, E >$.

Uwaga 5. *Każdy graf jednoznacznie wyznacza pewną relację dwuargumentową (binarną) w skończonym zbiorze V. Można również powiedzieć o dwrotnie, że każda relacja dwuargumentowa (binarna) r w skończonym zbiorze V, wyznacza jednoznacznie graf zorientowany, którego węzłami są elementy skończonego zbioru V, z kolei krawędziami są uporządkowane pary $< v, v^{'} >$, należące do r.*

Uwaga 6. *Niech dany będzie digraf D składający się z niepustego zbioru skończonego wierzchołków $V(D)$ i skończonej rodziny krawędzi $E(D)$. W momencie gdy*

$$\forall a, b \in V(D) \quad < a, b > \in E(D) \quad \Rightarrow \quad < b, a > \in E(D)$$

to taki graf skierowany, może być utożsamiony z grafem niezorientowanym.

$$\begin{array}{c} a \\ \updownarrow \\ b \end{array}$$

2.6 Wybrane pojęcia z teorii mnogości, topologii i algebry liniowej

2.6.1 Relacja porządkująca

W niniejszej pracy skupiamy się na zagadnieniu porządkowania danych wielowymiarowych. Konieczne jest zatem przywołanie o dpowiednich sformułowań dotyczących matematycznej definicji porządku. Najbardziej podstawowym pojęciem jest relacja porządku, która zostanie zdefiniowana poniżej. W sekcji tej zostaną również podane pojęcia równoliczności zbiorów, mocy zbioru, które są potrzebne przy definiowaniu własności porządkowania liniowego zbioru o biektów. Dodatkowo przytoczona zostanie definicja zbioru skończonego, ze względu na zastosowanie metod porządkowania na zbiorze skończonym.

Definicja 2.6.1. *Relacja [7, Rozdział 3]*
Niech dane będą zbiory X i Y. Relacją (dwuargumentową) między elementami zbiorów X i Y nazywamy dowolny podzbiór $\rho \subset X \times Y$. Jeśli $X = Y$ to mówimy, że ρ jest relacją na zbiorze X.

Definicja 2.6.2. *Relacja porządkująca (częściowego porządku) [3, Rozdział 2]*
Niech dana relacja ρ, którą o znaczać będziemy przez $\leqslant$, będzie o kreślona dla elementów ustalonego zbioru X. Mówimy, że relacja $\leqslant$ jest relacją częściowego porządku, jeśli spełnione są warunki:

1. *$x \leqslant x$ dla każdego x (zwrotność),*
2. *jeśli $x \leqslant y$ i $y \leqslant x$, to $x = y$ (słaba antysymetryczność),*
3. *jeśli $x \leqslant y$ i $y \leqslant z$, to $x \leqslant z$ (przechodniość).*

Przykład 2. *Częściowego porządku na zbiorze*
Wykorzystanie częściowego porządku, o brazuje diagram Hassego, będący grafem skierowanym, którego wierzchołki zostały poddane relacji porządkowania i reprezentują elementy skończonego zbioru X. Aby go skonstruować, należy postępować według poniższych kroków:

- *Punkty o brazujące elementy zbioru X, umieszcza się na płaszczyźnie.*
- *Punkt $x \in X$ łączony jest o dcinkiem z punktem $y \in X$ (gdzie $y \neq x$), jeśli x jest następnikiem y, czyli gdy $y \leqslant x$ o raz nie istnieje taki punkt $z \in X$, że $y \leqslant z \leqslant x$ o raz $x \neq z \neq y$.*

Zobrazowanie relacji częściowego porządku, dla punktów $A, B, C, D, E, F \in X$.

$$\begin{tikzcd}
 & B \arrow[r] & D & \\
 & & C \arrow[u] \arrow[r] & E \\
A \arrow[uur] \arrow[ur] \arrow[rrr] & & & F \arrow[u]
\end{tikzcd}$$

Definicja 2.6.3. *Relacja liniowo porządkująca (liniowy porządek) [3, Rozdział 2]*
Niech dany będzie niepusty zbiór X. Relację $\leqslant$ porządkującą zbiór X, nazywamy relacją liniowo porządkującą lub porządkiem liniowym, gdy dla dowolnych x, $y \in X$ spełnia o na następujący warunek spójności tzn. $x \leqslant y$ lub $y \leqslant x$. Parę $(X, \leqslant)$ nazywamy zbiorem liniowo uporządkowanym lub łańcuchem.

Definicja 2.6.4. *Elementy wyróżnione [3, Rozdział 2]*
Niech X będzie zbiorem częściowo uporządkowanym przez relacje $\leqslant$ o raz niech $a \in X$. Mówimy, że:

1. *a jest elementem najmniejszym w X, gdy dla każdego$x \in X \quad a \leqslant x$*
2. *a jest elementem minimalnym w X, gdy jest jedynym elementem najmniejszym w X*
3. *a jest elementem największym w X, gdy dla każdego $x \in X \quad x \leqslant a$*
4. *a jest elementem maksymalnym w X, gdy jest jedynym elementem największym w X*

Definicja 2.6.5. *Ograniczenie górne [3, Rozdział 2]*
Niech $A \subseteq X$, gdzie $(X, \leqslant)$ jest zbiorem uporządkowanym. Element $x \in X$ nazywamy o graniczeniem górnym zbioru A względem relacji $\leqslant$, gdy dla każdego $a \in A, \quad a \leqslant x$.

Definicja 2.6.6. *Ograniczenie dolne [3, Rozdział 2]*
Niech $A \subseteq X$, gdzie $(X, \leqslant)$ jest zbiorem uporządkowanym. Element $y \in X$ nazywamy o graniczeniem dolnym zbioru A względem relacji $\leqslant$, gdy dla każdego $a \in A, \quad y \leqslant a$.

Definicja 2.6.7. *Zbiór o graniczony [3, Rozdział 2]*
Niech $A \subseteq X$, gdzie $(X, \leqslant)$ jest zbiorem uporządkowanym. Zbiór nazywamy o graniczonym z góry (ograniczonym z dołu), jeśli ma o n o graniczenie górne (dolne). Zbiór o graniczony z dołu i z góry nazywamy o graniczonym.

Definicja 2.6.8. *Kres górny [3, Rozdział 2]*
Niech $A \subseteq X$, gdzie $(X, \leqslant)$ jest zbiorem uporządkowanym. Jeśli zbiór A jest o graniczony z góry i wśród o graniczeń górnych zbioru A istnienie element najmniejszy x_0, to element ten nazywamy kresem górnym zbioru A i oznaczamy symbolem $supA$. Tak więc $x_0 = supA$, gdy spełnione są następujące warunki:

1. *dla każdego $a \in A \quad a \leqslant x_0$,*
2. $\forall x \in X \quad (\forall a \in A \quad a \leqslant x) \Rightarrow x_0 \leqslant x$

Definicja 2.6.9. *Kres dolnym [3, Rozdział 2]*
Niech $A \subseteq X$, gdzie $(X, \leqslant)$ jest zbiorem uporządkowanym. Jeśli zbiór A jest o graniczony z dołu i wśród o graniczeń dolnych zbioru A istnienie element największy x_0, to element ten nazywamy kresem dolnym zbioru A i oznaczamy symbolem $infA$. Tak więc $y_0 = infA$, gdy spełnione są następujące warunki:

1. *$y_0 \leqslant a$ dla każdego $a \in A$,*
2. $\forall y \in X \quad (\forall a \in A \quad y \leqslant a) \Rightarrow y \leqslant y_0$

Definicja 2.6.10. *Zbiory równoliczne [3, Rozdział 5]*
Mówimy, że zbiory A i B są równoliczne (tej samej mocy), gdy istnieje bijekcja, tj. funkcja f różnowartościowa, przekształcająca zbiór A na zbiór B, tzn. $f : A \to B$. Piszemy wtedy: $\overline{\overline{A}} = \overline{\overline{B}}$.

Definicja 2.6.11. *Zbiór skończony [3, Rozdział 5]*
Mówimy, że zbiór A jest skończony, gdy jest pusty lub równoliczny ze zbiorem $\{1, \ldots, n\}$, dla pewnego $n \in \mathbb{N}$. Gdy zbiór jest równoliczny ze zbiorem $\{1, \ldots, n\}$, to mówimy że jest o n n-elementowy, tj. mocy równej n.

Definicja 2.6.12. *Zbiór przeliczalny [3, Rozdział 5]*
Mówimy, że zbiór X jest przeliczalny, gdy jest skończony lub jest równoliczny z $\mathbb{N}$.

2.6.2 Przestrzenie metryczne, miary o dległości

Niezbędnym jest również wprowadzenie podstawowych pojęć z topologii, ze względu na stosowanie funkcji o dległości w celu uporządkowania o biektów.

Definicja 2.6.13. *Metryka [7, Rozdzial 9]*
Niech X będzie niepustym zbiorem, wtedy funkcję $\mathrm{d} : X \times X \to [0, \infty)$*, nazywamy metryką jeśli spełnione są warunki:*

1. $\forall x, y \in X \quad \Big(\mathrm{d}(x, y) = 0 \Longleftrightarrow x = y\Big),$
2. $\forall x, y \in X \quad \mathrm{d}(x, y) = \mathrm{d}(y, x),$
3. $\forall x, y, z \ \in X \quad \mathrm{d}(x, y) \leqslant \mathrm{d}(x, z) + \mathrm{d}(z, y).$

Definicja 2.6.14. *Przestrzeń metryczna [7, Rozdział 9]*
Niech X będzie niepustym zbiorem, d *metryką, wówczas parę* (X, d) *nazywamy przestrzenią metryczną.*

Przykład 3. *Metryka euklidesowa w* $\mathbb{R}^2$
Niech $\mathrm{d}_e : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ *bedzie metryką euklidesową, wówczas*

$$\forall (x_1, x_2), (y_1, y_2) \in \mathbb{R}^2 \quad \mathrm{d}_e((x_1, y_1), (x_2, y_2)) := \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}.$$

Przykład 4. *Metryka miejska (Manhattan) w* $\mathbb{R}^2$
Niech $\mathrm{d}_m : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ *będzie metryką miejską, wówczas*

$$\forall (x_1, x_2), (y_1, y_2) \in \mathbb{R}^2 \quad \mathrm{d}_m((x_1, y_1), (x_2, y_2)) := |x_1 - x_2| + |y_1 - y_2|.$$

Przykład 5. *Przestrzeń euklidesowa n-wymiarowa* $\mathbb{R}^n$
Niech $\mathrm{d}_e : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ *będzie metryką euklidesową, wówczas*

$$\forall (x_1, x_2, \ldots, x_n), (y_1, y_2, \ldots, y_n) \in \mathbb{R}^n \quad \mathrm{d}_e(x, y) := \sqrt{\sum_{i=1}^{n} |x_i - y_i|^2}.$$

Rozdział 3

Metody porządkowania

Rozdział ten został o pracowany w oparciu o pozycję [9, Rozdział 2]. o mówimy w nim wybrane metody porządkowania zarówno liniowego jak i nieliniowego. W kolejnym rozdziale szczegółowo przyjrzymy się wybranym metodom wraz z przedstawieniem ich algorytmów o raz dokładnych o pisów matematycznych.

Metody porządkowania liniowego umożliwiają utworzenie uporządkowanej listy o biektów na podstawie o kreślonego kryterium (np. wartości zmiennych). Z kolei metody porządkowania nieliniowego zwracają graf połączeń podobnych o biektów, ze względu na opisujące je zmienne.

3.1 Metody porządkowania liniowego

W wielowymiarowej przestrzeni zmiennych, porządkowanie liniowe o biektów sprowadza się do rzutowania na prostą punktów, które reprezentują o biekty poddane porządkowaniu. Taka o peracja pozwala na ustalenie hierarchii o biektów.

Poniżej zostaną przedstawione własności uporządkowania liniowego o biektów, wraz z podaniem ich matematycznej interpretacji.

Obiekty uporządkowane liniowo charakteryzują się tym, że:

- każdy o biekt ma przynajmniej jednego sąsiada i nie więcej niż dwóch sąsiadów,
- jeżeli sąsiadem i-tego o biektu jest k-ty o biekt, to jednocześnie sąsiadem k-tego o biektu jest i-ty o biekt,
- dokładnie dwa o biekty mają tylko jednego sąsiada.

Powyżej wymienione własności są wynikiem posiadania jedynie skończonej ilości o biektów, które poddane są uporządkowaniu. W następnej części chcielibyśmy:

- sformalizować rozumienie powyższych własności,
- udowodnić ich poprawność,
- rozważyć dostateczność tych własności w zbiorach o skończonej ilości o biektów.

Na początku zaczniemy o d sprecyzowania takich pojęć jak sąsiad względem relacji.

Definicja 3.1.1. *Sąsiad względem relacji* $\leqslant$
Niech X będzie niepustym zbiorem, a x, y będą dwoma różnymi elementami należącymi do tego zbioru. Mówimy, że $y \in X$ jest sąsiadem $x \in X$,co zapisujemy ySx, jeśli
$$(y \leqslant x \vee x \leqslant y) \quad \wedge \quad (\neg \exists_{z \in X} \quad x \neq z \ \neq y \Rightarrow y \leqslant z \ \leqslant x \vee x \leqslant z \ \leqslant y).$$

Twierdzenie 3.1.2. *Własności porządku liniowego w zbiorach skończonych*
Niech $\leqslant$ będzie relacją porządku liniowego zdefiniowaną w X, gdzie X jest zbiorem ze skończoną liczbą o biektów, złożonym co najmniej z dwóch elementów. Wtedy:

1. $\forall_{x\in X} \quad \overline{\overline{\{y \in X, ySx\}}} \in \{1, 2\}$,

2. $\forall_{x,y\in X} \quad ySx \Rightarrow xSy$,

3. $\overline{\overline{\{x \in X, \quad \overline{\overline{\{y \in X, \quad ySx\}}} = 1\}}} = 2$,

gdzie S o znacza sąsiada względem relacji $\leqslant$.

Dowód. Poniżej zostaną udowodnione powyższe własności.

1. Niech $x \in X$. Przypuśćmy na początek, że $\overline{\overline{\{y \in X, ySx\}}} = 0$, tzn. że o biekt x nie posiada sąsiadów w tej relacji. Nasz zbiór X jest jednak co najmniej dwuelementowy, zatem istnieje element $y \in X$ i $x \neq y$. Wobec spójności linowego porządku z Definicji 2.6.3 zachodzi wtedy
$$x \leqslant y \vee y \leqslant x.$$
Jednak wiemy, że y nie może być sąsiadem x gdyż ten nie posiada sąsiadów. Zatem z definicji sąsiada musi istnieć $z \in X$ różny o d o bu $x \neq z \neq y$, spełniający warunek
$$z \leqslant x \vee x \leqslant z.$$
Powyższe rozumowanie dla y można by dalej zastosować do z, uzyskując kolejne z_1 a później $z_2, z_3, \ldots$ dowolną ilość różnych elementów, z których każdy występuje w relacji liniowego porządku z x, ale żaden z nich nie jest sąsiadem. Jednak nasz zbiór X jest zbiorem skończonym, więc nigdy nie uda nam się utworzyć dowolnej ilości różnych elementów ze zbioru X (elementy się wyczerpią). Zatem nasze przypuszczenie, że $\overline{\overline{\{y \in X, ySx\}}} = 0$ jest fałszywe.

   Przypuśćmy dalej, że $\overline{\overline{\{y \in X, ySx\}}} \geqslant 3$. Niech a, b, c będą trzema różnymi elementami z X będącymi sąsiadami dla x. Wtedy bez straty o gólności możemy przyjąć, że $a \leqslant x, b \leqslant x$ lub $x \leqslant a, x \leqslant b$. Istotnie mając 3 elementy w relacji wtedy co najmniej dwa muszą znajdować się po zgodnej stronie, a z dokładnością do o znaczeń możemy przyjąć, że będą nimi a o raz b. Ustalmy zatem, że $a \leqslant x, b \leqslant x$. Wobec definicji 2.6.3 wiemy, że $a \leqslant b$ lub $b \leqslant a$. Jeśli $a \leqslant b$ to $a \leqslant b \leqslant x$. Co przeczy temu, że a jest sąsiadem x. Jeśli $b \leqslant a$ to $b \leqslant a \leqslant x$ co przeczy temu, że b jest sąsiadem. Analogicznie postępujemy dla przypadku $x \leqslant a, x \leqslant b$. Uzyskujemy zatem sprzeczność, będącą efektem przypuszczenia, że mogą istnieć takie 3 elementy a, b, c. Zatem o statecznie $\overline{\overline{\{y \in X, ySx\}}} \in \{1, 2\}$.

2. Niech $x, y \in X$ o raz niech ySx. Korzystając z definicji sąsiada 3.1.1 mamy, że skoro ySx to
$$(y \leqslant x \vee x \leqslant y) \quad \wedge \quad (\neg\exists_{z\in X} \quad x \neq z \neq y \Rightarrow y \leqslant z \leqslant x \vee x \leqslant z \leqslant y).$$
Natomiast xSy o znacza, że
$$(x \leqslant y \vee y \leqslant x) \quad \wedge \quad (\neg\exists_{z\in X} \quad y \neq z \neq x \Rightarrow x \leqslant z \leqslant y \vee y \leqslant z \leqslant x).$$
Wobec powyższego widać, że te dwa zdania znaczą to samo, stąd widać że $ySx \Rightarrow xSy$.

3. Intuicyjnie te dwa elementy posiadające po jednym sąsiedzie są elementami maksymalnym i minimalnym w tym zbiorze. Udowodnimy kolejno:

- Element minimalny w zbiorze ma pojedynczego sąsiada. Zauważmy, że zbiór musi posiadać dokładnie 1 element minimalny, tzn. $x_m \in X$ takie, że

$$\forall x \in X \quad x_m \leqslant x.$$

Istotnie przypuśćmy, że nie istnieje element minimalny. Niech x_1 będzie dowolnym elementem z X. Skoro nie istnieje element minimalny, to istnieje $x_2 \in X$ takie, że $x_2 \leqslant x_1$ i $x_2 \neq x_1$. Dla x_2 z braku elementu minimalnemu, musi istnieć z kolei $x_3 \leqslant x_2$ takie, że $x_2 \neq x_3$. Itd. Co nie jest możliwe, gdyż zbiór X jest przecież skończonym zbiorem. Rozważmy dalej przypuszczenie gdyby były dwa lub więcej takich elementów. Wtedy to, z antysymetryczności, o czywiście musiałyby być sobie równe. Jeśli x_m, y_m są jednocześnie minimalne to

$$\forall x \in X \quad x_m \leqslant x,$$

oraz

$$\forall x \in X \quad y_m \leqslant x.$$

Skąd natychmiast mamy, że $x_m \leqslant y_m$ o raz $y_m \leqslant x_m$. Wobec antysymetryczności z definicji 2.6.2 mamy, że $x_m = y_m$ wbrew naszemu przypuszczeniu, że są o d siebie różne. Pozostaje pokazać, że element minimalny ma pojedynczego sąsiada. Przypuśćmy, że $y, z \in X$ są dwoma różnymi sąsiadami dla x_m. Wtedy $x_m \leqslant y \vee y \leqslant x_m$ o raz $x_m \leqslant z \ \vee z \ \leqslant x_m$. Skoro x_m jest minimalny to musi to o znaczać, że

$$x_m \leqslant y \wedge x_m \leqslant z.$$

Wobec spójności z definicji 2.6.3 zachodzi $y \leqslant z$ lub $z \leqslant y$. Sprzeczność, gdyż wtedy któryś z nich nie mógłby być sąsiadem dla x_m.

- Element maksymalny x_M w zbiorze ma pojedynczego sąsiada. Analogicznie do powyższego punktu, zbiór musi posiadać dokładnie 1 element maksymalny, tzn. $x_M \in X$ takie, że

$$\forall x \in X \quad x \leqslant x_M.$$

Istotnie przypuśćmy, że nie istnieje element maksymalny. Niech x_1 będzie dowolnym elementem z X. Skoro nie istnieje element maksymalny, to istnieje $x_2 \in X$ takie, że $x_1 \leqslant x_2$ i $x_1 \neq x_2$. Dla x_2 z braku elementu maksymalnego, musi istnieć taki element $x_3 \in X$ i $x_3 \neq x_2$, że $x_2 \leqslant x_3$. Itd. Co nie jest możliwe, gdyż z założenia zbiór X jest skończonym zbiorem. Rozważmy dalej przypuszczenie gdyby były dwa lub więcej takich elementów. Wtedy to z antysymetryczności, musiałyby być sobie równe. Jeśli x_M, y_M są jednocześnie maksymalne, to

$$\forall x \in X \quad x \leqslant x_M,$$

oraz

$$\forall x \in X \quad x \leqslant y_M.$$

Stąd natychmiast mamy, że $x_M \leqslant y_M$ o raz $y_M \leqslant x_M$. Wobec antysymetryczności z definicji 2.6.2, mamy że $x_M = y_M$, co wbrew naszemu przypuszczeniu daje, że elementy te nie są o d siebie różne. Pozostaje pokazać, że element maksymalny ma pojedynczego sąsiada. Przypuśćmy, że $y, z \ \in X$ są dwoma różnymi sąsiadami dla x_M. Wtedy $x_M \leqslant y \vee y \leqslant x_M$ o raz $x_M \leqslant z \ \vee z \ \leqslant x_M$. Skoro x_M jest elementem maksymalny to musi to zatem o znaczać

$$y \leqslant x_M \wedge z \ \leqslant x_M.$$

Wobec spójności z definicji 2.6.3 zachodzi $y \leqslant z$ lub $z \leqslant y$. Sprzeczność, gdyż wtedy któryś z nich nie mógłby być sąsiadem dla x_M.

- Żaden inny element nie może mieć pojedynczego sąsiada. Przypuśćmy, że $x \in X$ nie będąc ani elementem minimalnym ani maksymalnym ma pojedynczego sąsiada. Wobec definicji elementu minimalnego i maksymalnego o raz spójności zachodzi

$$x_m \leqslant x \leqslant x_M.$$

  Zatem albo x_m jest sąsiadem x albo istnieje $y_1 \in X$ taki, że $y_1 \leqslant x$. Tworzy to kilka możliwych przypadków. W pierwszym x_m będzie tym jedynym sąsiadem, w drugim x_M nim będzie, w ostatnim natomiast, ani x_m, ani x_M nie będą sąsiadami.

  Zajmiemy się najpierw pierwszym z nich, tj. x_m jest sąsiadem x. Zauważmy teraz, że z faktu, iż x_M jest elementem maksymalnym zbioru X, wynika że $x \leqslant x_M$. Nie jest jednak sąsiadem elementu x. Zatem istnieje takie $x_1 \in X$, że $x \leqslant x_1 \leqslant x_M$. Jednak x_1 również nie może być sąsiadem X co powoduje, że istnieje taki element $x_2 \in X$, że $x \leqslant x_2 \leqslant x_1 \leqslant x_M$. Itd. Jednakże, skoro zbiór X jest zbiorem skończonym, to musi istnieć taki element $x_j \in X$, że $x \leqslant x_j$ i $x_j \neq x_m$, który będzie sąsiadem z x, zatem xSx_j. Zatem o statecznie xSx_m i xSx_j, a to przeczy założeniu, że x ma pojedynczego sąsiada.

  Przejdźmy teraz do drugiego przypadku, tj. gdy x_M jest sąsiadem dla x. Wtedy x_m nie jest sąsiadem dla x jednak wiedząc, że $x_m \leqslant X$ musi istnieć $y_1 \in X$, taki że $y_1 \leqslant x$. Jednak wiedząc iż y_1 nie jest sąsiadem dla X wnioskujemy, że istnieje taki $y_2 \in X$, że $x_m \leqslant y_1 \leqslant y_2 \leqslant x$. Itd. Jednakże, skoro zbiór X jest zbiorem skończonym, to musi istnieć taki element $y_i \in X$, że $y_i \leqslant x$, który będzie sąsiadem z x. Zatem o statecznie y_iSx i xSx_M, co przeczy założeniu że x ma pojedynczego sąsiada.

  Zajmijmy się teraz trzecim przypadkiem, tj. gdy ani x_m o raz x_M nie są sąsiadami elementu x. Z faktu, iż zbiór X posiada element minimalny x_m, który nie jest sąsiadem elementu x, wynika że istnieje taki element $x_1 \in X$, że $x_m \leqslant x_1 \leqslant x$. Co więcej istnieje taki $x_2 \in X$, że $x_m \leqslant x_1 \leqslant x_2 \leqslant x$. Itd. I znów skoro zbiór X jest skończony, to istnieje taki element $x_j \in X$, że $x_j \leqslant x$ i x_jSx. Z drugiej strony, skoro zbiór X posiada element maksymalny x_M, który nie jest sąsiadem elementu x, wynika że istnieje taki element $y_1 \in X$, że $x \leqslant y_1 \leqslant x_M$. Analogicznie do wcześniejszych kroków, istnieje takie element $y_2 \in X$, że $x \leqslant y_2 \leqslant y_1 \leqslant x_M$. Itd. Zbiór X jest zbiorem skończonym, zatem musi istnieć taki element $y_i \in X$, że $x \leqslant y_i$ i xSy_1. Łącząc te dwa warunki, wynika że x musi mieć dwóch sąsiadów. Co kończy dowód własności.

□

Własności, które wykazaliśmy powyżej są często podawane niemal na równi z definicją takiego uporządkowania. Poniżej zostaną jednak podane przykłady takich relacji, które mimo, że posiadają powyższy zestaw własności, to nie o pisują relacji będących porządkami liniowymi.

Przykład 6. *Rozważmy zbiór dwuelementowy $X = \{a, b\}$ gdzie $a \leqslant b$ jest jedynym punktem tej relacji. Tak zdefiniowana relacja spełnia wszystkie własności, ale nie spełnia założenia o zwrotności - zatem relacja ta nie jest liniowym porządkiem. Diagram Hassego prezentujący tę relację, jest postaci:*

$$a \longrightarrow b$$

Przykład 7. *Rozważmy zbiór $X = \{a, b, c, d, e\}$ o raz relację definiującą następujące sąsiedztwa (wypisaną bez par symetrycznych) aSb, bSc, aSc, dSe. Ponadto dołóżmy warunek zwrotności, tj. $a \leqslant a$, $b \leqslant b$, $c \leqslant c$, $d \leqslant d$, $e \leqslant e$. Diagram Hassego prezentujący relację porządku tego zbioru,* *jest postaci:*

$$\begin{tikzcd} d \arrow[r] & e \\ & c \\ a \arrow[r] \arrow[ur] & b \arrow[u] \end{tikzcd}$$

Z diagramu widać, że taka relacja spełnia wszystkie o mawiane wcześniej własności - jednak nie jest spójna. I tak np. nie możemy porównać elementów a i d, bowiem nie możemy o kreślić czy $d \leqslant a$ lub $a \leqslant d$.

W podsumowaniu tej sekcji należy podkreślić, że by uporządkować liniowo o biekty z macierzy o bserwacji, charakteryzujące je zmienne muszą być mierzone przynajmniej na skali porządkowej. Istotne jest również aby miały jednakowy charakter. Na potrzeby pracy zakładamy, że zmienne o pisujące o biekty powinny być stymulantami. Dlatego też gdy nimi nie są, należy poddać je np. stymulacji. o peracja ta umożliwia w dalszym kroku przejście do transformacji normalizacyjnej, która konieczna jest gdy zmienne o pisujące o biekty mierzone są na skali przedziałowej lub ilorazowej, a chcemy uzyskać ich porównywalność.

Metody porządkowania liniowego można podzielić na metody diagramowe, procedury o parte na zmiennej syntetycznej o raz procedury iteracyjne bazujące na funkcji kryterium dobroci uporządkowania. W kolejnej sekcji zostaną pokrótce przedstawione różne metody, z wyszczególnieniem najważniejszych założeń o każdej z nich.

3.1.1 Metody diagramowe

W metodach diagramowych stosuje się graficzną reprezentację macierzy o dległości zwanej diagramem. Macierz konstruowana jest w oparciu o o dległości między o biektami, wyznaczone za pomocą dowolnej metryki. Porządkowanie o biektów polega na porządkowaniu diagramu, tzn. przestawieniu wierszy i odpowiadających im kolumn, aby wzdłuż przekątnej skupiały się najmniejsze o dległości zaś im dalej o d głównej przekątnej tym większe o dległości między zmiennymi o pisującymi porządkowane o biekty. Narzędzie pomocnicze w porządkowaniu danych, może stanowić kryterium postaci:

$$F^1 = \sum_{i=1}^{n} \sum_{k>1}^{n} d_{ik} w_{ik}$$

gdzie:
$D = [d_{ik}]$ - macierz o dległości między i-tym i k-tym o biektem,
$W = [w_{ik}] \quad i, k = 1, 2, \ldots, n$ - macierz wag elementów macierzy o dległości, wymiaru równego wymiarowi macierzy o dległości.

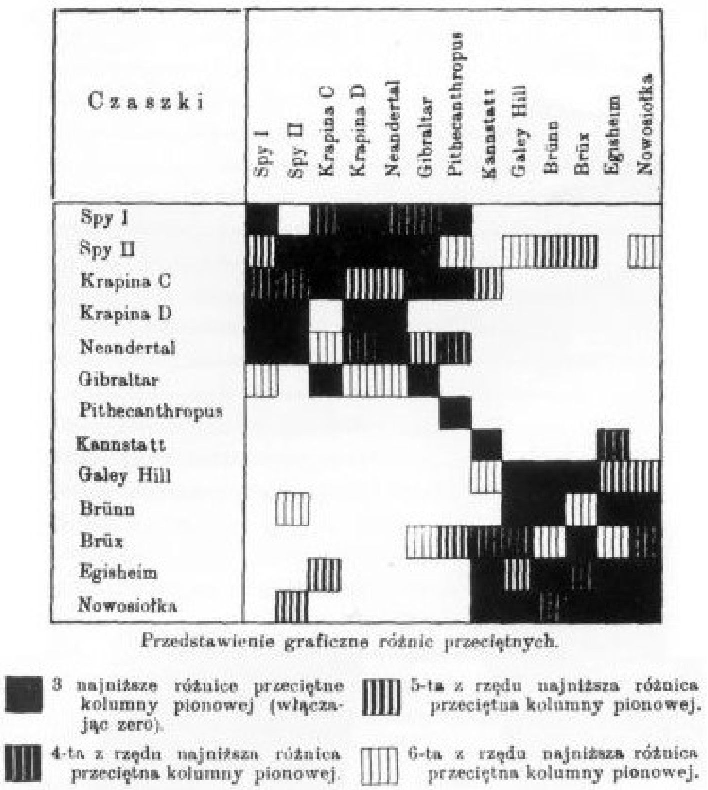
Elementy macierzy wag wyznaczone są za pomocą jednego z poniższych wzorów:

(a) $w_{ik} = \frac{|i-k|}{n-1}$,

(b) $w_{ik} = \frac{1}{n(n-1)}[2n|i-k-1| + i + k - (i-i)^2]$,

(c) $w_{ik} = \frac{1}{n(n-1)}[2n|i-k| + 2 - i - k - (i-i)^2]$.

Zaprezentujmy teraz przykład uporządkowanego diagramu, przedstawiającego wynik badań Jana Czekanowskiego, dotyczących metod badania różnic między kopalnymi czaszkami ludzkimi. Analizując diagram zauważamy że wzdłuż głównej przekątnej skupiają się najmniejsze o dległości między zmiennymi, a im dalej o d niej tym o dległości zwiększają się. Przykład został znaleziony na stronie [11].



Rysunek 3.1: Diagram o publikowany w podręczniku „Statystyka dla antropologów” Jana Czekanowskiego, 1913r. [11]

Metody diagramowe nie leżą w zakresie zainteresowań tej pracy, z uwagi na mały formalizm matematyczny. Są jednak istotnym narzędziem do wizualizacji porządku.

3.1.2 Metody o parte na zmiennych syntetycznych

W tym podrozdziale zostaną o pisane metody porządkowania o parte na zmiennych syntetycznych, tj. funkcji wyznaczonej na podstawie wartości zmiennych o pisujących o biekty, której wartości będą służyć do porządkowania zbioru. Metody o parte na zmiennych syntetycznych dzielimy na wzorcowe i bezwzorcowe. Poniżej zostaną o ne o pisane szczegółowo, jednak wcześniej zostaną przedstawione wzory wyznaczające zmienną syntetyczną.

Sposoby wyznaczania zmiennej syntetycznej

W pracy dla zachowania o gólności, przyjmujemy że wszystkie zmienne o pisujące o biekty mają jednakowe wagi, w związku z tym wzory służące do wyznaczenia zmiennej syntetycznej są postaci:

1. średniej arytmetycznej:

$$s_i = \frac{1}{m}\sum_{j=1}^{m} n_{ij}, \quad i = 1, 2, \ldots, n,$$

2. średniej geometrycznej:

$$s_i = \prod_{j=1}^{m} (n_{ij})^{\frac{1}{m}}, \quad i = 1, 2, \ldots, n,$$

3. średniej harmonicznej

$$s_i = \Big[\sum_{j=1}^{m} \frac{1}{n_{ij}}\Big]^{-1} \cdot m, \quad i = 1, 2, \ldots, n,$$

gdzie: s_i - wartość zmiennej syntetycznej w i-tym o biekcie,

Metoda wzorcowe

W metodach tych zakłada się istnienie o biektu wzorcowego $P_0 = [n_{0j}], \quad j = 1, 2, \ldots, m$. Zmienne tego o biektu są znormalizowane. Przyjmują o ne o ptymalne wartości, które są ustalane na podstawie o gólnie przyjętych norm, subiektywnej o pinii dotyczącej o bserwowanego o biektu, lub też o pinii ekspertów. Wtedy o biekty, porządkowane są na podstawie o dległości o d o biektu wzorcowego. Poszczególne metody mogą różnić się sposobem wyznaczania o biektu wzorcowego, o dległości o raz miary syntetycznej, na podstawie której dokonywane jest porządkowanie.

Metoda Hellwiga

Metoda Hellwiga jest jedną z najstarszych metod wzorcowych. W metodzie tej, o biekt wzorcowy wyznaczony jest na podstawie wystandaryzowanych zmiennych wejściowych. Współrzędnym o biektu wzorcowego przyporządkowuje się maksimum, gdy zmienne wejściowe są stymulantami lub minimum gdy zmienne są destymulantami. o biekty są uporządkowywane na podstawie o dległości o d o biektu wzorcowego, przy wykorzystaniu o dległości euklidesowej. Do porządkowania używana jest natomiast miara syntetyczna, postaci:

$$s_i = 1 - \frac{d_{i0}}{d_0}, \quad i = 1, 2, \ldots, n,$$

gdzie:
d_{i0} - o dległość i-tego o biektu, o d o biektu wzorcowego
d_0 - wartość wyznaczana dla każdej zmiennej, będąca sumą średniej o dległości o d o biektu o raz podwojonej wartości o dchylenia standardowego dla tej zmiennej. Doprecyzujmy
Współrzędne o biektu wzorcowego są o bliczane na podstawie wzoru:

$$n_{0j} = \begin{cases} \max\limits_i(n_{ij}) & \text{gdy } n_j \text{ jest stymulantą}, \quad j = 1, 2, \ldots, m, \quad i = 1, 2, \ldots, n, \\ \min\limits_i(n_{ij}) & \text{gdy } n_j \text{ jest destymulantą}, \quad j = 1, 2, \ldots, m, \quad i = 1, 2, \ldots, n. \end{cases}$$

Dalej wyznaczamy miarę syntetyczną, licząc kolejno:

- Dla każdego o biektu, wyznaczana jest o dległość o d o biektu wzorcowego:

$$d_{i0} = \left[\sum_{j=1}^{m}(n_{ij} - n_{0j})^2\right]^{\frac{1}{2}}$$

- Następnie dla każdej zmiennej wyznaczana jest średnia o dległość o d o biektu wzorcowego wg. wzoru:

$$\overline{d_0} = \frac{1}{n}\sum_{i=1}^{n} d_{i0}$$

- W kolejnym kroku wyznaczamy o dchylenie standardowe dla każdej zmiennej za pomocą wzoru:

$$\sigma(d_0) = \left[\frac{1}{n}\sum_{i=1}^{n}(d_{i0} - \overline{d_0})^2\right]^{\frac{1}{2}}$$

- Mając powyższe, możemy wyznaczyć wartość d_0 jako sumę średniej o dległości o raz podwojonej wartości o dchylenia standardowego:

$$d_0 = \overline{d_0} + 2\sigma(d_0)$$

Wartości miary s_i zazwyczaj są z przedziału $[0, 1]$, jednakże mogą zdarzyć się wartości ujemne. Należy tu zaznaczyć, że wartości miary są tym wyższe, im mniej jest o ddalony o biekt o d o biektu wzorcowego.

Metoda dystansowa

Podobnie jak we wcześniejsze metodach, na początku zmienne należy poddać stymulacji, o raz przekształceniu normalizacyjnemu, wybranemu na podstawie skal do których należą zmienne o pisujące o biekty. W kolejnym kroku wyznaczane są współrzędne o biektu wzorcowego, a następnie macierz o dległości każdego o biektu o d o biektu wzorcowego. o dległość o d o biektu wzorca jest wyznaczana przy zastosowaniu dowolnej metryki, np. metryki euklidesowej. Dla metody tej, miara syntetyczna jest wyznaczana za pomocą przekształcenia unitaryzacyjnego postaci:

$$s_i = \left(\frac{d_{i0} - \min\limits_i(d_{i0})}{\max\limits_i(d_{i0}) - \min\limits_i(d_{i0})} \right)^p, \quad i = 1, 2, \ldots, n, \quad p \in \mathbb{N}.$$

Miara syntetyczna uzyskana tą metodą jest unormowana i przyjmuje wartości z przedziału: $[0, 1]$. Czym niższa wartość miary, tym bliżej o biektu wzorcowego leży dany o biekt.

Metody bezwzorcowe

W metodach tych zakładamy, że nie istnieje o biekt wzorcowy. Porządkowanie dokonywane jest na podstawie wartości zmiennej syntetycznej, wyznaczonej dla każdego o biektu. Poniżej zostaną o mówione wybrane metody porządkowania bezwzorcowego.

Metoda rang

Definicja 3.1.3. *Ranga [9, Rozdział 1.5]*
Rangą nazywamy zmienną o wartościach ze zbioru $\mathbb{R}_+ \setminus \{0\}$, będącą najczęściej liczbą całkowitą. Wartości reprezentują numery miejsc o biektów po uporządkowaniu malejąco.

Metoda ta o piera się na normalizacji rangowej, w związku z tym zmienne poddane porządkowaniu, powinny być mierzone na skali porządkowej. Dla każdego o biektu wyznacza się sumę przyporządkowanych mu rang ze względu na wszystkie zmienne. Na końcu o bliczana jest wartość zmiennej syntetycznej, jako średniej wartości rang. W oparciu o tę wartość następuje porządkowanie o biektów, tj. im wartość zmiennej syntetycznej jest mniejsza tym wyżej w hierarchii znajduje się uporządkowany o biekt. Wzór na obliczenie wartości zmiennej syntetycznej:

$$s_i = \frac{1}{m} \sum_{j=1}^{m} n_{ij}, \quad i = 1, 2, \ldots, n,$$

gdzie:
n_{ij}-zmienna znormalizowana rangowo, tj. $n_{ij} = r$ dla $x_{rj} = x_{ij}$, $r, i = 1, 2, \ldots, n$,
r-ranga nadana i-temu o biektowi znajdującemu się na r-tym miejscu w uporządkowanym szeregu o biektów ze względu na j-tą zmienną.

Metoda sum

Metoda ta używana jest w momencie, gdy zmienne mierzone są na skali ilorazowej lub przedziałowej. W związku z tym tuż po stymulacji zmiennych, należy dokonać przekształcenia normalizacyjnego, za pomocą unitaryzacji. W kolejnym kroku, dla każdego o biektu wyznaczana jest zmienna syntetyczna, jako średnia arytmetyczna wartości zmiennych przy przyjęciu jednakowych wag dla każdej zmiennej. Następnie muszą zostać wyeliminowane ujemne wartości zmiennej syntetycznej, do czego służy poniższe przekształcenie:

$$s'_i = s_i - \min_i(s_i), \quad i = 1, 2, \ldots, n.$$

Końcowa postać zmiennej syntetycznej o trzymywana jest, przy wykorzystaniu normalizacji postaci:

$$s''_i = \frac{s'_i}{\max_i(s'_i)}, \quad i = 1, 2, \ldots, n.$$

Powyższe przekształcenia ujednolicają zakres miary syntetycznej do przedziału [0, 1]. Im wyższa wartość zmiennej syntetycznej, tym wyżej w hierarchii znajduje się o biekt.

3.1.3 Metody iteracyjne

W metodach tych przyjmowania jest funkcja kryterium dobroci porządkowania, dla której w kolejnych iteracjach poszukiwane jest takie uporządkowanie liniowe o biektów, które o ptymalizuje wartość funkcji kryterium, aż do o siągnięcia przez nią wartości o ptymalnej tj. maksymalnej lub minimalnej.

Metoda Szczotki

Metoda ta polega na znalezieniu takiego uporządkowania liniowego o biektów, dla którego funkcja kryterium dobroci uporządkowania o siąga maksimum:

$$F^2 = \sum_{k=1}^{n-1} k \sum_{i=1}^{n-k} d_{ik} \rightarrow \max$$

gdzie:

d_{ik} - o dległość euklidesowa miedzy i-tym i k-tym o biektem.

W pierwszym kroku działania tej metody, przeprowadzane jest dowolne liniowe uporządkowanie o biektów. Następnie, dla tego uporządkowania o bliczana jest wartość funkcji kryterium dobroci uporządkowania, według powyższego wzoru. W kolejnych etapach wyznaczana jest wartość tej funkcji, dla każdej transpozycji pary o biektów. Powyższe kroki wykonywane są do momentu, gdy dowolna transpozycja pary o biektów nie spowoduje zwiększenia wartości funkcji kryterium dobroci uporządkowania.

3.2 Metody porządkowania nieliniowego

Metody porządkowania nieliniowego w odróżnieniu o d metod porządkowania liniowego, polegają nie na uporządkowaniu o biektów w sposób hierarchiczny, a na określeniu dla każdego z nich, stopnia podobieństwa z innymi o biektami, na podstawie o pisujących je zmiennych.

Aby zastosować metody porządkowania nieliniowego, zmienne o pisujące o biekty, powinny być mierzone na skali przedziałowej lub ilorazowej. Gdy zmienne te mierzone są na skali przedziałowej lub ilorazowej, należy dokonać ich normalizacji.

Metody porządkowania nieliniowego dzielimy na metody dendrytowe o raz aglomeracyjne. Metody dendrytowe prowadzą do powstania dendrytu, prezentującego położenie o biektów ze względu na ich podobieństwo między sobą. Z kolei metody aglomeracyjne sprowadzają się do powstania dendrogramu lub łańcucha połączeń, prezentując sposób łączenia o biektów do siebie podobnych.

3.2.1 Metody dendrytowe

Definicja 3.2.1. *Dendryt [9, Rozdział 2.3]*
Dendrytem nazywamy acykliczny graf spójny, bez pętli.

Metody dendrytowe o pierają się na pojęciach teorii grafów. Metody te polegają na stworzeniu dendrytu, którego wierzchołki o dpowiadają o dpowiednim o biektom poddanym porządkowaniu. Krawędzie łączące wierzchołki o dpowiadają zaś o dległością między o biektami. Przykładem metod dendrytowych jest taksonomia wrocławska, metoda Prima. Poniżej zostanie jednak o pisana jedynie metoda taksonomii wrocławskiej.

Taksonomia wrocławska

W metodzie tej o biekty dzielone są na grupy o biektów najbardziej do siebie podobnych, tj. takich, dla których o dległość między sobą jest jak najmniejsza. W związku z tym w pierwszej kolejności należy wyznaczyć macierz o dległości o biektów D np. przy użyciu metryki euklidesowej, a następnie w każdym wierszu (kolumnie) macierzy, wyznaczamy jest element najmniejszy:

$$d_{ik} = \min_k d_{ik}, \quad i, k = 1, 2, \ldots, n, i \neq k.$$

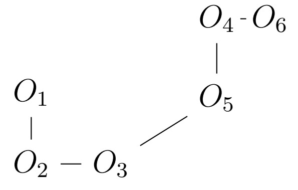
Za pomocą grafu prezentowane są pary o biektów, najbardziej do siebie podobnych. W grafie tym, długość krawędzi łączących wierzchołki (czyli o biekty poddane porządkowaniu) o dpowiadają o dległości między parą o biektów. Jeżeli wśród połączonych par o biektów, pojawią się krawędzie dwukrotne, należy je wyeliminować, ze względu na to, że kolejność połączeń w dendrycie nie jest istotna. o biekty w dendrycie nie mogą się powtarzać, jeżeli natomiast niektóre o biekty w łączeniu wystąpią wielokrotnie, to o biekty te zostaną połączone w zespoły zwane skupieniami. Metoda ta kończy swoje działanie, w momencie uzyskania grafu spójnego.

Zaprezentujemy teraz tę metodę na przykładzie.

Przykład 8. *Niech dana będzie macierz o dległości* $D = [d_{ik}]_{i \leqslant 6, k \leqslant 6}$ *sześciu różnych o biektów* $\{O_1, O_2, ..., o\ _6\}$*, w której to w kolejnych wierszach znajdują o dległości i-tego o biektu o d o biektu k-tego:*

$$D = \begin{bmatrix} 0.00 & \underline{0.35} & 0.70 & 0.95 & 2.36 & 2.99 \\ \underline{0.35} & 0.00 & 0.45 & 1.45 & 2.00 & 0.36 \\ 0.70 & \underline{0.45} & 0.00 & 1.05 & 0.90 & 0.75 \\ 0.95 & 1.45 & 1.05 & 0.00 & 0.50 & \underline{0.16} \\ 2.36 & 2.00 & 0.90 & \underline{0.50} & 0.00 & 0.52 \\ 2.99 & 0.36 & 0.75 & \underline{0.16} & 0.52 & 0.00 \end{bmatrix}$$

Mając wyznaczoną macierz o dległości, postępujemy według wyżej wskazanej reguły, która prowadzi do uzyskania dendrytu postaci:



W powstałym dendrycie wierzchołkami są o biekty przechodzące zbioru C, z kolei krawędzie łączące te wierzchołki są współrzędne wektora c, które powstały przez wybór najmniejszej o dległości miedzy dołączanymi o biektami, w kolejnych etapach dołączania o biektów do dendrytu.

3.2.2 Metody aglomeracyjne

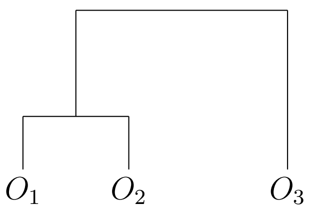
Przed o pisem działania metod aglomeracyjnych, wprowadzimy definicję łańcucha połączeń.

Definicja 3.2.2. *Łańcuch połączeń*
Niech $O = \{O_1, O_2, ..., o_{\ n}\}$, $\quad n \in \mathbb{N}$ o znacza zbiór o biektów. Łańcuchem połączeń nazwiemy skończony ciąg $(C_i)_{i=1}^{N}, N \in \mathbb{N}, n \leqslant N \leqslant 2^n$, taki że:

1. $C_i = \{O_i\}$ *dla* $i = 1, 2, \ldots, n$
2. $\forall_{i>n} \quad \exists_{j,k \in \mathbb{N}, j \leqslant k < i} \quad C_i = C_j \cup C_k$
3. $C_N = O = \{O_1, O_2, ..., o_{\ n}\}$
4. $\forall_{i<j, i,j \in \mathbb{N}} \quad C_i \cap C_j \neq \emptyset \Rightarrow C_i \subset C_j$

Uwaga 7. *Graficzną reprezentacje łańcucha połączeń nazywamy dendrogramem.*

Przykład 9. *Zaprezentujemy teraz przykład tworzenia łańcucha połączeń. Niech dany będzie zbiór o biektów $O = \{O_1, o_{\ 2}, o_{\ 3}\}$. Wtedy przykładowy łańcuch połączeń jest postaci $C = \{\{O_1\}, \{O_2\}, \{O_3\}, \{O_1, o_{\ 2}\}, \{O_1, o_{\ 2}, o_{\ 3}\}\}$. Zaprezentujemy go teraz graficznie, w postaci dendrogramu:*



Opis działania metod aglomeracyjnych:

Istotą metod aglomeracyjnych jest utworzenie łańcucha połączeń i dendrogramu. W ten sposób zobrazowana jest kolejność łączenia o biektów, na podstawie zmniejszającego się podobieństwa między o biektami włączonymi do dendrogramu, a tymi wcześniej do niego należącymi. Położenie o biektów o raz grup o biektów, które powstały w kolejnych etapach tworzenia łańcucha, jest przeprowadzone na podstawie kolejności połączeń tych o biektów i grup. Każde o gniwo w łańcuchu o znacza grupy o biektów podobnych do siebie. Wyjściowym założeniem metod aglomeracyjnych jest to, że każdy o biekt stanowi o drębną, jednoelementową grupę ($\mathrm{C_i}, i = 1, 2, \ldots, n$).

W kolejnych krokach następuje łączenie ze sobą grup o biektów najbardziej podobnych do siebie ze względu na ich zmienne. Podobieństwo weryfikowane jest na podstawie o dległości między grupami.

Na początku o dległości między jednoelementowymi grupami $\mathrm{C_1}, \ldots, \mathrm{C_n}$ wyznacza wyjściowa macierz o dległości D. W macierzy D poszukiwane są najmniejsze o dległości pomiędzy grupami o biektów:

$$d_{ii'} = \min_{ik} d_{ik}, \quad i = 1, 2, \ldots, n_i, \quad k = 1, 2, \ldots, n_{i'}, \quad i, i^{'} = 1, 2, \ldots, n, i \neq i^{'}.$$

gdzie:
$d_{ii'}$ - o dległość i-tej o d $i^{'}$-tej grupy.

W kolejnym kroku, o biekty o najmniejszej o dległości między sobą łączone są w jedną grupę, dzięki czemu liczba grup zmniejsza się o jeden. Zostaje rozpoczęty proces tworzenia łańcucha połączeń. Ponownie badane są o dległości między nowo stworzoną grupą, a pozostałymi grupami. Proces trwa do momentu stworzenia pełnego łańcucha połączeń, tj. jednej grupy.

Ogólna postać wzoru służącego do wyznaczenia o dległości nowo powstałej grupy $\mathrm{C_{i''}}$, (powstałej dzięki połączeniu grup $\mathrm{C_i}$ i $\mathrm{C_{i'}}$), o d grup które zostały C_i''' to:

$$d_{i'''i''} = \alpha_i d_{i'''i} + \alpha_{i'} d_{i'''i'} + \beta d_{ii'} + \gamma |d_{i'''i} - d_{i'''i'}|$$

gdzie: $\alpha_i, \alpha_{i'}, \beta, \gamma$ - współczynniki przekształceń, różne dla poszczególnych metod aglomeracyjnych

Możemy wyróżnić sześć różnych metod aglomeracyjnych, różniących się sposobem wyznaczenia o dległości między grupami o biektów. Poniżej zostaną podane współczynniki przekształceń dla każdej z nich, a w dalszej części pracy wybrane z nich zostaną szczegółowo o mówione.

- metoda najbliższego sąsiedztwa (metoda pojedynczego wiązania):

  parametry przekształceń $\alpha_i = 0,5 \quad \alpha_{i'} = 0,5 \quad \beta = 0 \quad \gamma = 0,5.$

- metoda najdalszego sąsiedztwa (metoda pełnego wiązania):

  parametry przekształceń $\alpha_i = 0,5 \quad \alpha_{i'} = 0,5 \quad \beta = 0 \quad \gamma = -0,5.$

- metoda średniej międzygrupowej (metoda średnich połączeń):

  parametry przekształceń $\alpha_i = \frac{n_i}{n_i+n_{i'}} \quad \alpha_{i'} = \frac{n_{i'}}{n_i+n_{i'}} \quad \beta = 0 \quad \gamma = 0).$

- metoda mediany:

  parametry przekształceń $\alpha_i = 0,5 \quad \alpha_{i'} = 0,5 \quad \beta = -025 \quad \gamma = 0.$

- metoda środka ciężkości:

  parametry przekształceń $\alpha_i = \frac{n_i}{n_i+n_{i'}}; \alpha_{i'} = \frac{n_{i'}}{n_i+n_{i'}} \quad \beta = \frac{-n_i n_{i'}}{(n_i+n_{i'})^2} \quad \gamma = 0.$

- metoda Warda:

  parametry przekształceń $\alpha_i = \frac{n_i+n_{i'''}}{n_i+n_{i'}+n_{i'''}} \quad \alpha_{i'} = \frac{n_{i'}+n_{i'''}}{n_i+n_{i'}+n_{i'''}} \quad \beta = \frac{-n_{i'''}}{n_i+n_{i'}+n_{i'''}} \quad \gamma = 0.$

Metoda najbliższego sąsiedztwa

W metodzie tej o dległość między dwoma grupami o biektów jest równa o dległości pomiędzy najbliższymi o biektami (sąsiadami), które należą do dwóch różnych grup. o dległość ta o pisana jest wzorem:

$$d_{ii'} = \min_{ik} d_{ik}(\mathbf{O_i} \in \mathbf{C_i}, \mathbf{o}\ _\mathbf{k} \in \mathbf{C_{i'}}),$$

$$i = 1, 2, \ldots, n_i, \quad k = 1, 2, \ldots, n_{i'}, \quad i, i^{'} = 1, 2, \ldots, n, \quad i \neq i^{'},$$

gdzie:

$$\mathbf{O_i} = [n_{ij}], \quad j = 1, 2, \ldots, m.$$

Metoda najdalszego sąsiedztwa

W metodzie tej o dległość między dwoma grupami o biektów jest równa o dległości pomiędzy najdalszymi o biektami (sąsiadami), które należą do dwóch różnych grup. o dległość ta o pisana jest wzorem:

$$d_{ii'} = \max_{ik} d_{ik}(\mathbf{O_i} \in \mathbf{C_i}, \mathbf{o\ _k} \in \mathbf{C_{i'}}),$$

$$i = 1, 2, \ldots, n_i, \quad k = 1, 2, \ldots, n_{i'}, \quad i, i^{'} = 1, 2, \ldots, n, \quad i \neq i^{'},$$

Metoda średniej międzygrupowej

W metodzie tej o dległość między dwoma grupami o biektów równa jest średniej arytmetycznej o dległości między wszystkimi parami o biektów należących do dwóch różnych grup. o dległość ta o pisana jest wzorem:

$$d_{ii'} = \frac{1}{n_i n_{i'}} \sum_{k=1}^{n_{i'}} \sum_{i=1}^{n_i} d_{ik}(\mathbf{O_i} \in \mathbf{C_i}, \mathbf{o\ _k} \in \mathbf{C_{i'}})$$

$$i, i^{'} = 1, 2, \ldots, n, \quad i \neq i^{'}.$$

Metoda mediany

W metodzie tej o dległość między grupami o biektów jest równa medianie o dległości pomiędzy wszystkimi parami o biektów należących do dwóch grup. o dległość ta o pisana jest wzorem:

$$d_{ii'} = \mathrm{med}_{i,k}\{d_{ik}(\mathbf{O_i} \in \mathbf{C_i}, \mathbf{o\ _k} \in \mathbf{C_{i'}})\},$$

$$i = 1, 2, \ldots, n_i, \quad k = 1, 2, \ldots, n_{i'}, \quad i, i^{'} = 1, 2, \ldots, n, \quad i \neq i^{'}.$$

Rozdział 4

Zastosowanie wybranych metod porządkowania danych wielowymiarowych

4.1 Opis zbioru

Zbiór danych jest o pracowaniem własnym, na podstawie o fert sprzedaży samochodów o sobowych, zamieszczonych na portalu *www.otomoto.pl* w okresie listopad - grudzień 2017 roku. Zebrane dane dotyczą szczegółowych informacji o dnośnie samochodu, tj. jego marki, modelu, wersji, typu, koloru lakieru, pojemności silnika, roku produkcji, przebiegu, liczby drzwi, rodzaju skrzyni biegu, rodzaju paliwa, rodzaju napędu, wyposażenia w: ABS, komputer pokładowy, ESP, klimatyzację. o prócz danych ściśle związanych z budową i wyposażeniem samochodu, pojawiły się również atrybuty, tj. cechy umieszczone w kolumnach, związane z informacją o tym czy auto jest uszkodzone o raz bezwypadkowe,czy jest sprowadzane, jaki jest kraj aktualnej rejestracji, czy było serwisowane, czy sprzedający jest pierwszym właścicielem. Dodatkowo o prócz powyższych, został dodany atrybut najbardziej interesujący kupującego - czyli cena o raz województwo tj. miejsce skąd wystawiana została o ferta. Zbiór został dołączony do pracy na płycie.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	MARKA	MODEL	WERSJA	TYP	WOJEWC	CENA.NET	CENA.BRU	MOC[km]	POJEMNC	ROK.PRO	PRZEBIEG	KOLOR	L.DZRZWI	RODZAJ.F	SKRZYNIA	NAPED	KRAJ.AKT
2	Hyundai	i20	II	kompakt	malopolskie		46500	90	1396	2016	18300	bialy	3	diesel	manualna	na przedni	Polska
3	Hyundai	i20	I	kompakt	mazowieckie		22900	86	1248	2013	319000	niebieski	5	benzyna+L	manualna	na przedni	Polska
4	Subaru	Legacy	V	kombi	mazowieckie		36900	150	1998	2010	149000	srebrny-m	5	benzyna	manualna	4x4(stały)	Polska
5	Ford	Mondeo	Mk4	sedan	dolnoslaskie		30000	146	1999	2008	166290	czarny	5	benzyna+L	manualna	na przedni	Polska
6	Opel	Astra	G	kompakt	slaskie		3990	136	1998	1998	230000	czarny	5	benzyna	manualna	na przedni	Polska
7	Mazda	Premacy		minivan	dolnoslaskie		7750	100	1998	2004	210563	srebry	5	diesel	manualna	na przedni	Niemcy
8	Seat	Leon	II	kompakt	dolnoslaskie		19900	200	1984	2006	198000	czarny	5	benzyna	manualna	na przedni	Szwajcaria
9	Volkswage	Passat	B6	sedan	lodzkie		13999	105	1900	2005	202749	czarny	5	diesel	manualna	na przedni	Polska
10	Opel	Zafira	A	minivan	lodzkie		7900	125	1800	2001	196000	srebrny	5	benzyna	manualna	4x4(stały)	Niemcy
11	Mercedes-	Klasa A	W168	kompakt	lodzkie		6500	102	1598	2002	200000	niebieski	5	diesel	manualna	na przedni	Polska
12	Skoda	Octavia	II	kombi	malopolskie		19500	140	1986	2008	220000	czarny	5	diesel	manualna	na przedni	Polska
13	Seat	Toledo	II	kompakt	wielkopolskie		11300	105	1598	2003	174600	szary	4	benzyna	manualna	na przedni	Niemcy
14	Peugeot	508		kombi	slaskie		42500	115	1560	2014	156800	bialy	5	diesel	manualna	na przedni	Polska
15	Skoda	Octavia	III	kombi	wielkopolskie		50900	105	1598	2015	91800	bialy	5	diesel	manualna	na przedni	Polska
16	Peugeot	Partner	I	kombi	lodzkie		7990	90	2000	2004	275763	zloty	5	diesel	manualna	na przedni	Polska
17	Porsche	Cayman		coupe	wielkopolsl	95900	117957	300	3400	2006	83000	srebrny	3	benzyna	automatyc.	na tylne ko	Polska
18	Toyota	Auris	II	kompakt	mazowieckie		46000	132	1598	2014	46000	bialy-metal	5	benzyna	manualna	na przedni	Polska
19	Toyota	Auris	II	kompakt	dolnoslaskie		30300	99	1329	2014	151783	bialy	5	benzyna	manualna	na przedni	Polska
20	Mazda	3	II	kompakt	zachodniopomorskie		25200	105	1598	2009	135800	czarny	5	benzyna	manualna	na przedni	Austria
21	Mazda	3	II	kombi	malopolskie		30600	150	1999	2009	116000	brazowy	5	benzyna	manualna	na przedni	Niemcy
22	Mazda	6	I	kombi	dolnoslaskie		15700	146	2000	2007	190000	czarny	5	diesel	manualna	na przedni	Polska
23	Mazda	6	I	kompakt	lodzkie		17900	147	2000	2006	238482	szary	5	benzyna+L	manualna	na przedni	Polska
24	Mazda	6	I	kombi	kujawsko-pomorskie		11000	121	1998	2005	204000	srebrny	5	diesel	manualna	na przedni	Polska
25	Citroen	C4	II	kompakt	wielkopolskie		36200	92	1560	2014	86561	bialy	5	diesel	manualna	na przedni	Polska
26	Citroen	C4	II	kompakt	malopolskie		36997	90	1397	2014	33000	bialy	5	benzyna	manualna	na przedni	Polska
27	Citroen	C4	II	kompakt	lodzkie		16900	90	1397	2014	35000	szary	5	benzyna	manualna	na przedni	Francja

Rysunek 4.1: Podgląd stworzonego zbioru

Użyte zmienne

W stworzonym zbiorze danych znajduje się 29 atrybutów, o pisujących 61 różnych rekordów, tj. o biektów reprezentowanych przez wiersze, którym przypisano pewne wartości atrybutów. Wśród zebranych danych można wyróżnić zarówno zmienne jakościowe, jak i ilościowe.

Zmiennymi jakościowymi są atrybuty (pisownia zgodna z plikiem danych):

- marka,
- model,
- typ,
- wojewodztwo,
- kolor,
- rodzaj.paliwa,
- skrzynia.biegow,
- naped,
- kraj.aktualnej.rejestracji,
- kraj.pochodzenia,
- stan,
- ABS,
- uszkodzony,
- pierwszy.wlasciciel,
- kto.sprzedaje,
- serwisowany,
- komputer.pokladowy,
- ESP,
- klimatyzacja,
- bezwypadkowy,
- status.pojazdu.sprowadzanego

Zmiennymi ilościowymi są atrybuty:

- cena.netto,
- cena.brutto,
- moc,
- pojemnosc.skokowa,
- rok.produkcji,
- przebieg,
- l.drzwi.

4.2 Użyte programy

Zbiór danych został umieszczony w pliku arkusza programu Excel, z kolei implementacje zostały stworzone w języku R, który ma zastosowanie w statystyce i ekonometrii.

4.3 Implementacje wybranych metod

W sekcji tej zaprezentujemy implementacje wybranych metod porządkowania liniowego o raz nieliniowego. W części dotyczącej porządkowania liniowego porównamy również wyniki porządkowania par metod: metody rang i metody sum, metody sum i metody Hellwiga, metody rang i metody Hellwiga. Zostanie to przeprowadzone na całym zbiorze o biektów. Zaczniemy jednak o d przedstawienia o gólnych funkcji o dpowiedzialnych za stymulacje o raz transformacje normalizacyjne. Po to by móc następnie je wykorzystać przy prezentacji wybranych metod porządkowania nieliniowego o raz liniowego.

4.3.1 Stymulacja zmiennych

Poniżej zostaną przedstawione metody stymulacji zmiennych o pisane wcześniej w sekcji 2.4. o graniczamy się tu do przypadku, że dana zmienna jest destymulantą i należy ją poddać stymulacji. W tym celu stworzyliśmy dwie funkcje:

- stymulacja_przeksztalcenie_ilorazowe(x,y),
- stymulacja_przeksztalcenie_roznicowe(x,y).

W miejscu argumentu 'x' należy wpisać nazwę zbioru na którym dokonywane jest porządkowanie, z kolei w miejscu argumentu 'y' należy podać nazwę kolumny poddanej stymulacji, z tym że nazwa kolumny musi zostać podana w cudzysłowie.

```
stymulacja_przeksztalcenie_ilorazowe<-function(x,y)
{
  for (i in 1:nrow(x))
  {
    x[i,which(colnames(x)==y)]=1/x[i,which(colnames(x)==y)]
  }
  return(x)
}
```

```
stymulacja_przeksztalcenie_roznicowe<-function(x,y)
{
  max_wartosc=max(x[which(colnames(x)==y)])
  for (i in 1:nrow(x))
  {
    x[i,which(colnames(x)==y)]=max_wartosc-x[i,which(colnames(x)==y)]
  }
  return(x)
}
```

Uwaga 8. *Stymulacja zmiennych dokonywana jest pojedynczo, tj. jeżeli w naszym zbiorze jest wiele zmiennych mających charakter destymulanty, dla każdej z nich musimy użyć funkcji, na sam koniec nadpisać nasz zbiór, tym nowym wystymulowanym, dzięki czemu przekształcenia zostaną zapisane.*

4.3.2 Transformacje normalizacyjne

Poniżej zostaną przedstawione o gólne funkcje transformacji normalizacyjnej - standaryzacja, unitaryzacja, przekształcenie ilorazowe. Podobnie jak metody stymulacji zmiennych, zostały o ne wcześniej o mówione w sekcji 2.4. Aby wywołać którąś z poniższych funkcji, należy w miejsce 'x' wpisać nazwę zbioru.

```
standaryzacja<-function(x)
{
  suma=0
  srednia=0
  odchylenie=0
  for (j in 2:ncol(x))
  {
    suma[j]=sum(x[j])
    srednia[j]=suma[j]/nrow(x)
    suma_kwadratow=0
    kwadrat=0
    for(i in 1:nrow(x))
    {
      kwadrat=(x[i,j]-srednia[j])^2
      suma_kwadratow=suma_kwadratow+kwadrat
```

```
    }
    odchylenie[j]=sqrt(suma_kwadratow/nrow(x))
    for (i in 1:nrow(x))
    {
      x[i,j]=(x[i,j]-srednia[j])/odchylenie[j]
    }
  }
  return(x)
}
```

```
unitaryzacja<-function(x)
{
  maksi=0
  minim=0
  for (j in 2:ncol(x))
  {
    maksi[j]=max(x[j])
    minim[j]=min(x[j])
    for (i in 1:nrow(x))
    {
    x[i,j]=(x[i,j]-minim[j])/(maksi[j]-minim[j])
    }
  }
return(x)
}
```

```
przeksztalcenie_ilorazowe<-function(x)
{
  suma=0
  srednia=0
  for (j in 2:ncol(x))
  {
    suma[j]=sum(x[j])
    srednia[j]=suma[j]/nrow(x)
    for(i in 1:nrow(x))
    {
      x[i,j]=x[i,j]/srednia[j]
    }
  }
return(x)
}
```

4.3.3 Metody porządkowania nieliniowego

Zostanie tutaj zaprezentowane zastosowanie nieliniowego porządkowania danych przy pomocy istniejących funkcji biblioteki cluster, w której to funkcja agnes umożliwia uporządkowanie zbioru po wyborze odpowiedniej metody aglomeracyjnej. Mamy tu do wyboru metody: single - metoda najbliższego sąsiedztwa, complete - metoda najdalszego sąsiedztwa, ward - metoda

Warda, average - metoda średniej między grupowej. Poniżej zostanie zaprezentowane zastosowanie metod single o raz complete wraz z porównaniem wyników porządkowania.

Import danych

Na początku należy jednak zaimportować dane, które chcemy poddać porządkowaniu. W tym celu należy zaimportować biblotekę readxl - gdyż dane pobieramy z excela, a w kolejnym kroku wywołujemy plik, podając jego ścieżkę z rozszerzeniem xlsx. My użyjemy tutaj zbioru wszystkich o biektów, będących o fertami sprzedaży aut.

```
library(readxl)
zbior_danych <- read_excel("datasets/zbior_danych.xlsx",
                           sheet = "DANE_INNA_WERSJA")
```

Podgląd danych:

```
head(zbior_danych)
```

```
## # A tibble: 6 x 30
##      Nr MARKA   MODEL   WERSJA TYP     WOJEWODZTWO  `CENA.NETTO_[pln]`
##   <dbl> <chr>   <chr>   <chr>  <chr>   <chr>                     <dbl>
## 1  1.00 Hyundai i20     II     kompakt malopolskie                  NA
## 2  2.00 Hyundai i20     I      kompakt mazowieckie                  NA
## 3  3.00 Subaru  Legacy  V      kombi   mazowieckie                  NA
## 4  4.00 Ford    Mondeo  Mk4    sedan   dolnoslaskie                 NA
## 5  5.00 o~pel    Astra   G      kompakt slaskie                      NA
## 6  6.00 Mazda   Premacy <NA>   minivan dolnoslaskie                 NA
## # ... with 23 more variables: `CENA.BRUTTO_[pln]` <dbl>, `MOC_[km]` <dbl>,
## #   `POJEMNOSC.SKOKOWA_[cm3]` <dbl>, ROK.PRODUKCJI <dbl>, `PRZEBIEG_[km]`
## #   <dbl>, KOLOR <chr>, L.DZRZWI <dbl>, RODZAJ.PALIWA <chr>,
## #   SKRZYNIA.BIEGOW <chr>, NAPED <chr>, KRAJ.AKTUALNEJ.REJESTRACJI <chr>,
## #   KRAJ.POCHODZENIA <chr>, STATUS.POJAZDU.SPROWADZONEGO <chr>,
## #   PIERWSZY.WLASCICIEL <dbl>, KTO.SPRZEDAJE <chr>, STAN <chr>,
## #   SERWISOWANY <dbl>, ABS <dbl>, KOMPUTER.POKLADOWY <dbl>, ESP <dbl>,
## #   KLIMATYZAJCA <dbl>, BEZWYPADKOWY <dbl>, USZKODZONY <dbl>
```

Podzbiór danych

W kolejnym, kroku po przyjrzeniu się zbiorowi danych, użytkownik musi zadecydować na których danych ilościowych chce pracować - ważna jest znajomość danych. Dodatkowo pierwszą kolumną musi być kolumna zawierająca numery indeksów o biektów, ze względu na to, że jako wynik zastosowania funkcji o dpowiedzialnej za porządkowanie, zostaną zwrócone w kolejności malejącej numery indeksów, mówiące o kolejności uporządkowania. W związku z tym, za pomocą poniższej procedury użytkownik tworzy podzbiór zaimportowanego zbioru, gdzie w cudzysłowie wpisuje nazwy kolumn zawierających zmienne ilościowe, wybrane do porządkowania (zakładamy, że podzbiór będzie nazywał się dane_porzadkowanie)). Wybranymi kolumnami są: cena, moc, pojemność, rok produkcji, przebieg.

```
dane_porzadkowanie<-zbior_danych[c("Nr","CENA.BRUTTO_[pln]","MOC_[km]",
                                   "POJEMNOSC.SKOKOWA_[cm3]",
                                  "ROK.PRODUKCJI","PRZEBIEG_[km]")]
```

Transformacje danych

Przed samym porządkowaniem, wymagane jest aby zmienne miały charakter stymulant o raz by zostały poddane transformacji normalizacyjnej. Aby funkcja dokonująca porządkowania dawała poprawny wynik, użytkownik musi zająć się transformacją przed jej zastosowaniem. Poniżej podaliśmy tego przykład. Dla zmiennych, które stymulantami nie są, należy dokonać stymulacji. Wśród zmiennych poddanych porządkowaniu, do stymulant nie należy zmienna zmienna: przebieg - jest destymulantą, w związku z tym, została przekształcona na stymulantę, za pomocą przekształcenia ilorazowego.

```
dane_porzadkowanie<-stymulacja_przeksztalcenie_ilorazowe(dane_porzadkowanie,
"PRZEBIEG_[km]")
```

W celu uzyskania porównywalności między zmiennymi, zostały o ne poddane transformacji normalizacyjnej - unitaryzacji.

```
dane_porzadkowanie<-unitaryzacja(dane_porzadkowanie)
```

Zastosowanie metod aglomeracyjnych

Chcąc zastosować funkcję agnes, należy w pierwszej kolejności wyznaczyć macierz o dległości pomiędzy wszystkimi parami o biektów. Do wyznaczenia odległości zostanie użyta metryka euklidesowa - w tym celu zostanie wykorzystania funkcja dist(x, method="") - w miejsce 'x' należy wpisać nazwę tabeli zawierających dane do uporządkowania, a w nawiasie[,] na miejscu drugiej współrzędnej należy podać wektor kolumn, na podstawie którego wartości zostanie wyznaczona macierz o dległości. W miejsce argumentu method należy wpisać nazwę metryki na podstawie której zostanie obliczona o dległość - w naszym przypadku będzie to euclidean - euklidesowa.

```
odleglosci <- dist(dane_porzadkowanie[,c("CENA.BRUTTO_[pln]","MOC_[km]",
            "POJEMNOSC.SKOKOWA_[cm3]","ROK.PRODUKCJI","PRZEBIEG_[km]")],
```

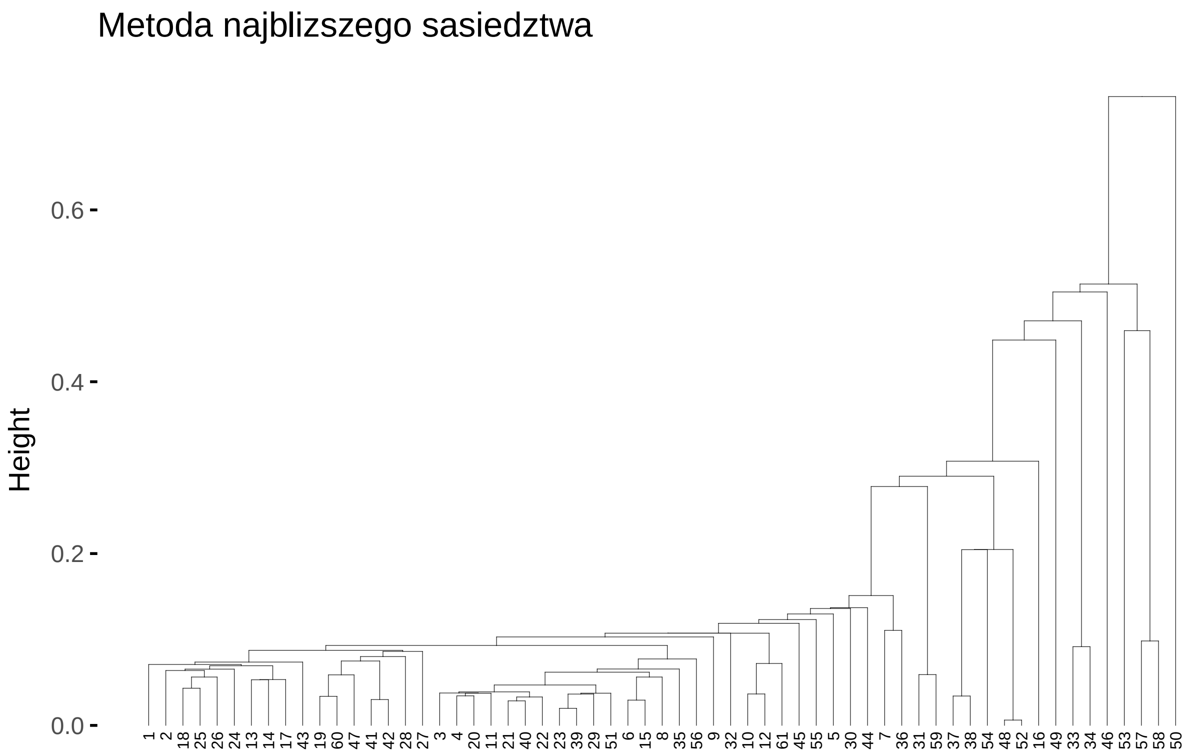
Następnie należy zaimportować bibliotekę cluster, by móc skorzystać z metod aglomeracyjnych. Jak już zostało wspomnanie we wstępnie, wywołanie metod aglomeracyjnych o dbywa się dzięki funkcji agnes(x, method="") . W miejsce argumentu 'x' - zostanie podana wyznaczona macierz o dległości z kolei, w kolejnym argumencie - method zostanie podana reguła wyznaczania odległości pomiędzy nową grupą a pozostałymi o biektów. Regułą tą może być metoda najbliższego sąsiedztwa, najdalszego, Warda lub średniej między grupowej. My wykorzystamy metodę najbliższego sąsiedztwa o raz najdalszego.

```
library(cluster)
metoda_najblizszego <- agnes(odleglosci, method = "single")
metoda_najdalszego<- agnes(odleglosci, method = "complete")
```

Ostatni etap to graficzne zaprezentowanie wyniku w postaci dendrogramu. W tym celu należy zaimportować bibliotekę factoextra, w której to jest funkcja fviz_dend(x, main = ""). W miejsce argumentu x należy wpisać nazwę o biektu powstałego przy pomocy funkcji agnes, main to tytuł wykresu. Dodatkowo na osi pionowej zaprezentowane są o dległości między obiektami, z kolei na osi poziomej znajdują się numery indeksów obiektów.

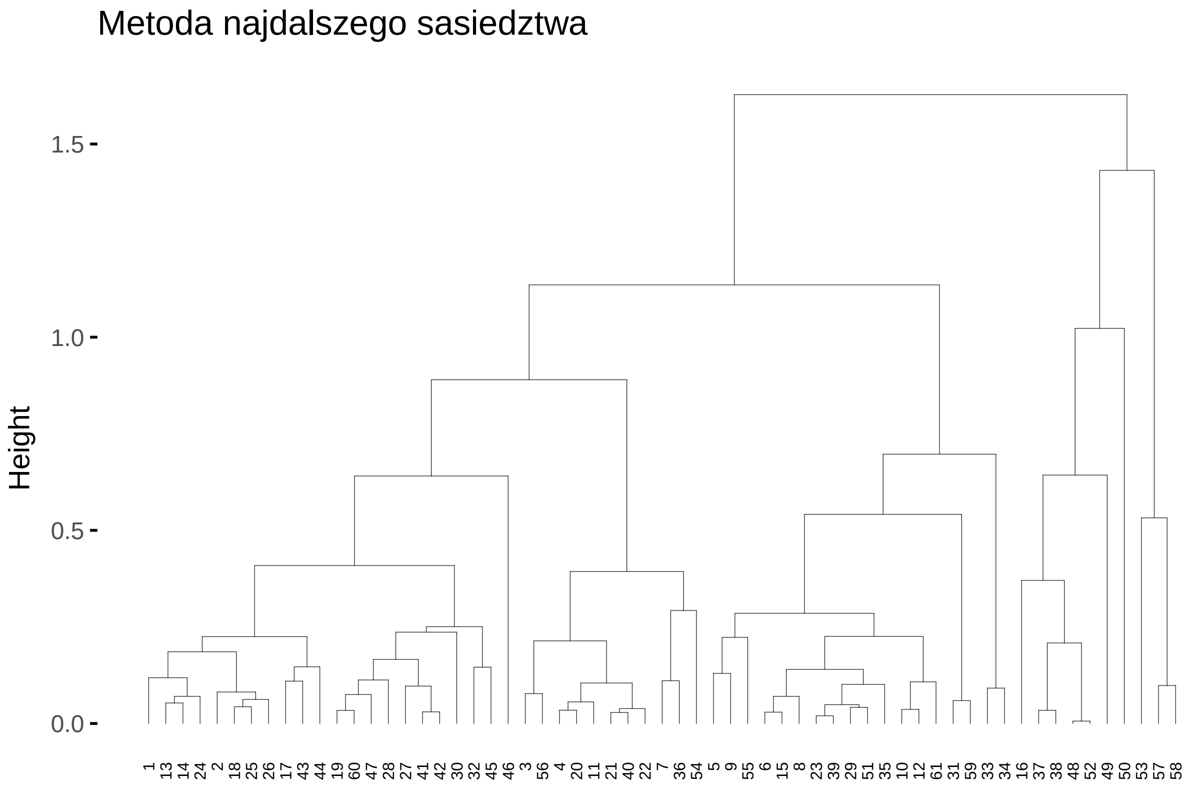
Dendrogram dla metody najbliższego sąsiada, prezentuje się następująco:

```r
library(factoextra)
fviz_dend(metoda_najblizszego, lwd=0.1, cex=0.45,
main = "Metoda najbliższego sąsiedztwa")
```



Dendrogram dla metody najdalszego sąsiada, wygląda natomiast:

```r
fviz_dend(metoda_najdalszego, lwd=0.1, cex=0.45,
main = "Metoda najdalszego sąsiedztwa")
```



Porównanie wyników uporządkowania

W przypadku dendrogramu uzyskanego metodą najdalszego sąsiedztwa można zauważyć większą o dległość wiązań niż dla metody najbliższego sąsiedztwa, które to o brazują o dległość między grupami. Dodatkowo obiekty charakteryzujące się najbardziej korzystnym wartości zmiennych, zostały pogrupowane w jedną grupę (tu mowa o o biektach o numerach indeksów 49, 50, 53, 57, 58), natomiast w przypadku metody najbliższego sąsiedztwa o biekty te zostały rozdzielone na pojedyncze grupy. Można również zauważyć, że w przypadku uporządkowania metodą najdalszego sąsiedztwa o biekty tworzą pewnego rodzaju podgrupy - skupiska, z kolei w przypadku uporządkowania metodą najbliższego sąsiada, wynik porządkowania można porównać do wyglądu lawiny, lub góry, u której podnóża znajdują się o biekty o najniższych wartościach o pisywanych je zmiennych, które to następnie łączą się i dochodzą do szczytu, który stanowią obiekty o najwyższych wartościach o pisywanych je zmiennych.

4.3.4 Metody porządkowania liniowego

Zostaną tu zaprezentowane metody liniowe: metoda sum, rang o raz Hellwiga. Dla wszystkich metod będziemy korzystać z tego samego zbioru danych, który jest podzbiorem wyjściowego zbioru - zawiera 8 najbardziej różniących się o biektów. Różnice dla każdej z metod będą zaczynały się o d sekcji związanej z transformacją danych.

Import danych

Wobec powyższego, podobnie jak przy metodach porządkowania nieliniowego, użytkownik musi zaimportować dane, które chce poddać porządkowaniu.

```r
library(readxl)
zbior_danych <- read_excel("datasets/8_Rozniacych_sie_obiektow.xlsx",
                           sheet = "Arkusz1", col_types = c("numeric","text",
                           "text", "text", "text", "text",
                              "numeric", "numeric", "text",  "numeric",
                              "text", "text", "text", "text","text",
                                  "text", "numeric", "text", "text",
                                   "text", "numeric", "numeric",
                                   "numeric", "numeric", "numeric"))
```

Podgląd danych:

```r
head(zbior_danych)
```

```
## # A tibble: 6 x 29
##       Nr MARKA  MODEL  WERSJA TYP   WOJEWODZTWO `CENA.BRUTTO_[p~ `MOC_[km]`
##    <dbl> <chr>  <chr>  <chr>  <chr> <chr>                  <dbl>      <dbl>
## 1   1.00 Mazda  3      II     komp~ zachodniop~            25200        105
## 2   2.00 Jaguar XF     X260   kombi dolnoslask~           323600        240
## 3   3.00 Subaru B9 Tr~ <NA>   suv   malopolskie            38900        245
## 4   4.00 Volks~ Golf   VII    kombi lodzkie               113900        150
## 5   5.00 Peuge~ 508    <NA>   kombi slaskie                42500        115
## 6   6.00 o~pel   Antara <NA>   suv   lodzkie                24000        150
## # ... with 21 more variables: `POJEMNOSC.SKOKOWA_[cm3]` <dbl>,
## #   ROK.PRODUKCJI <dbl>, `PRZEBIEG_[km]` <dbl>, KOLOR <chr>, L.DZRZWI
## #   <dbl>, RODZAJ.PALIWA <chr>, SKRZYNIA.BIEGOW <chr>, NAPED <chr>,
## #   KRAJ.AKTUALNEJ.REJESTRACJI <chr>, KRAJ.POCHODZENIA <chr>,
## #   STATUS.POJAZDU.SPROWADZONEGO <chr>, PIERWSZY.WLASCICIEL <dbl>,
## #   KTO.SPRZEDAJE <chr>, STAN <chr>, SERWISOWANY <chr>, ABS <dbl>,
## #   KOMPUTER.POKLADOWY <dbl>, ESP <dbl>, KLIMATYZAJCA <dbl>, BEZWYPADKOWY
## #   <dbl>, USZKODZONY <dbl>
```

Podzbiór danych

W kolejnym, kroku po przyjrzeniu się zbiorowi danych, użytkownik musi zadecydować na których danych ilościowych chce pracować - ważna jest znajomość danych. Dodatkowo pierwszą kolumną musi być kolumna zawierająca numery indeksów o biektów, ze względu na to, że w wyniku zastosowania funkcji o dpowiedzialnej za porządkowanie, zostaną zwrócone w kolejności malejącej numery indeksów, mówiące o kolejności uporządkowania. W związku z tym, za pomocą poniższej procedury użytkownik tworzy podzbiór zaimportowanego zbioru, gdzie w cudzysłowie wpisuje nazwy kolumn zawierających zmienne ilościowe, wybrane do porządkowania (przyjmijmy założenie, że podzbiór będzie nazywał się dane_porzadkowanie - będzie to pomocne w dalszej części programu). U nas wybranymi kolumnami są: cena, moc, pojemność, rok produkcji, przebieg.

```r
dane_porzadkowanie<-zbior_danych[c("Nr","CENA.BRUTTO_[pln]","MOC_[km]",
                                   "POJEMNOSC.SKOKOWA_[cm3]",
                                   "ROK.PRODUKCJI","PRZEBIEG_[km]")]
```

Transformacje danych

Przed zastosowaiem metod, należy dokonać ich transformacji. W pierwszym kroku należy dokonać stymulacji destymulant - dla metody rang i metody Hellwiga zostanie użyte przekształcenie ilorazowe, dla metody sum przekształcenie ilorazowe. Po stymulacji można przejść do transformacji normalizacyjnej - dla metod: sum i rang użyjemy normalizacji, natomiast dla metody Hellwiga standaryzacji.

4.3.5 Metoda sum

Stymulacja

```
dane_porzadkowanie<-stymulacja_przeksztalcenie_roznicowe(dane_porzadkowanie,
"PRZEBIEG_[km]")
```

Transformacja normalizacyjna

```
dane_porzadkowanie<-unitaryzacja(dane_porzadkowanie)
```

Funkcja porządkująca

Chcąc przeprowadzić porządkowanie znormalizowanych danych za pomocą metody sum, należy użyć poniższej funkcji:

```
funkcja_porzadkowanie_metoda_sum<-function(x)
  {
  x[,"zmienna_syntetyczna"] <-0 #ostatnia kolumna to zmienna_syntetyczna
  for(i in 1:nrow(x))
  {
    for(j in 2:(ncol(x)-1))
    {
      x[i,ncol(x)]=x[i,ncol(x)]+x[i,j]
    }#liczba kolumn - (kolumna_nr_indeksu+kolumna_zmienna_synt)
  x[i,ncol(x)]=x[i,ncol(x)]/(ncol(x)-2)
  }
  #wyeliminowanie ujemnych wartosci zmiennej syntetycznej
  min_zmienna=min(x$zmienna_syntetyczna)
  for(i in 1:nrow(x))
  {
    x[i,ncol(x)]=x[i,ncol(x)]-min_zmienna
  }
  #normalizacja zm. syntetycznej
  max_zmienna=max(x$zmienna_syntetyczna)
  for(i in 1:nrow(x))
  {
    x[i,ncol(x)]=x[i,ncol(x)]/max_zmienna
  }
  x<-x[order(-x$zmienna_syntetyczna),]
  return(x[1])
}
```

Wywołanie funkcji dla zbioru dane_porzadkowanie

```
funkcja_porzadkowanie_metoda_sum(dane_porzadkowanie)
```

```
## # A tibble: 8 x 1
##       Nr
##    <dbl>
## 1  2.00
## 2  4.00
## 3  3.00
## 4  5.00
## 5  6.00
## 6  1.00
## 7  8.00
## 8  7.00
```

Funkcja zwraca nam indeksy uporządkowanych o biektów, tj. 1 miejsce zajął o biekt z numerem indeksu 2, 2 miejsce o biekt z numerem indeksu równym 4, z kolei miejsce o statnie zajął o biekt o numerze indeksu równym 7.

4.3.6 Metoda rang

Stymulacja

```
dane_porzadkowanie<-stymulacja_przeksztalcenie_ilorazowe(dane_porzadkowanie,
"PRZEBIEG_[km]")
```

Transformacja normalizacyjna

```
dane_porzadkowanie<-unitaryzacja(dane_porzadkowanie)
```

Funkcja porządkująca

Chcąc przeprowadzić porządkowanie znormalizowanych danych za pomocą metody rang, należy użyć poniższej funkcji:

```
funkcja_porzadkowanie_metoda_rang<-function(x)
  {
  y<-x
  for (i in 2:ncol(x))
  {
    x[ncol(x)+1]=rank(-x[i])
  }
  x[,"zmienna_syntetyczna"] <-0 #ostatnia kolumna to zmienna_syntetyczna
  for(i in 1:nrow(x))
  {
    for(j in (ncol(y)+1):(ncol(x)-1))
    {
      x[i,ncol(x)]=x[i,ncol(x)]+x[i,j]
```

```
      j=j+1
    }
    x[i,ncol(x)]=x[i,ncol(x)]/(ncol(x)-7)
  }
  x<-x[order(x$zmienna_syntetyczna),]
  print("Numery indeksów o~biektów po uporządkowaniu: ")
  return(x[1])
}
```

Wywołanie funkcji dla zbioru dane_porzadkowanie

```
funkcja_porzadkowanie_metoda_rang(dane_porzadkowanie)
```

```
## [1] "Numery indeksów o~biektów po uporządkowaniu: "

## # A tibble: 8 x 1
##      Nr
##   <dbl>
## 1  2.00
## 2  4.00
## 3  3.00
## 4  5.00
## 5  6.00
## 6  1.00
## 7  8.00
## 8  7.00
```

Na podstawie powyższego wyniku, widać, że 1 miejsce zajęła o ferta sprzedaży z numerem indeksu 2, 2 miejsce o ferta o numerze indeksu równym 4, a o statnie o ferta z numerem indeksu numer 7. Porównując wyniki porządkowania tej metody, z metodą sum zauważamy, że wyniki są takie same.

4.3.7 Metoda Hellwiga

Stymulacja

```
dane_porzadkowanie<-stymulacja_przeksztalcenie_ilorazowe(dane_porzadkowanie,
"PRZEBIEG_[km]")
```

Transformacja normalizacyjna

```
dane_porzadkowanie<-standaryzacja(dane_porzadkowanie)
```

Funkcja porządkująca

Chcąc przeprowadzić porządkowanie znormalizowanych danych za pomocą metody Hellwiga, należy użyć poniższej funkcji:

```
funkcja_porzadkowanie_metoda_Hellwiga<-function(x)
  {
  obiekt_wz=0
  for (j in 2:ncol(x))
  {
  obiekt_wz[j]=max(x[j])
  }
  odleg<- x[c("Nr" )]
  for (i in 1:nrow(x))
  {
  SUMKA=0
  for (j in 2:ncol(x))
  {
    SUMKA=SUMKA+(x[i,j]-obiekt_wz[j])^2
  }
  odleg[i,2]=sqrt(SUMKA) #kolumna zawierajaca o~dleglosci
  }
  d_0=0
  suma=0
  srednia=0
  odchylenie=0
  for (j in 2:ncol(odleg))
  {
    suma[j]=sum(odleg[j])
    srednia[j]=suma[j]/nrow(odleg)
    suma_kwadratow=0
    kwadrat=0
    for(i in 1:nrow(odleg))
    {
      kwadrat=(odleg[i,j]-srednia[j])^2
      suma_kwadratow=suma_kwadratow+kwadrat
    }

    odchylenie[j]=sqrt(suma_kwadratow/nrow(odleg))
    d_0=srednia[j]+2*odchylenie[j]  #d_0 to po prostu wartosc
    }
#ostatnia kolumna to jak zawsze zmienna syntetyczna
    x[,"zmienna_syntetyczna"] <-0
    for (i in 1:nrow(x))
    {
      x[i,ncol(x)]=1-(odleg[i,2]/d_0)
    }
    x<-x[order(-x$zmienna_syntetyczna),]
    return(x[1])
    }
```

Wywołanie funkcji dla zbioru dane_porzadkowanie

```
metoda_Hellwiga(dane_porzadkowanie)
```

```
## # A tibble: 8 x 1
##      Nr
##   <dbl>
## 1  2.00
## 2  4.00
## 3  3.00
## 4  6.00
## 5  5.00
## 6  1.00
## 7  8.00
## 8  7.00
```

Na podstawie powyższego wyniku, widać, że 1 miejsce zajęła o ferta sprzedaży z numerem indeksu 2, a o statnie o ferta z numerem indeksu o numerze 7. Wobec powyższego wyniki porządkowania tą metodą są inne, niż wyniki metody sum.

4.3.8 Porównanie wyników metod porządkowania liniowego dla całego zbioru

Analizując wyniki porządkowania zaprezentowane w poprzedniej sekcji, zauważamy że dla tak małego zbioru uzyskaliśmy zgodność rezultatów dla metody sum o raz rang. Sprawdzimy teraz jak będzie w przypadku zastosowania tych metod dla całego zbioru. W związku z tym należy zaimportować cały zbiór danych. W dalszej kolejności, należy stworzyć trzy kopie zbioru, dla każdej z metod porządkowania. Następnie przeprowadzamy o dpowiednie transformacje o raz stosujemy algorytmy porządkowania.

Import danych

Tak samo jak przy implementacji metod porządkowania nieliniowego, importujemy pełen zbiór o biektów.

```
library(readxl)
zbior_danych <- read_excel("datasets/zbior_danych.xlsx",
                           sheet = "DANE_INNA_WERSJA")
```

Podzbiór danych

Mając zaimportowany zbiór danych, tworzymy podzbiór na którym będziemy pracować.

```
dane_porzadkowanie<-zbior_danych[c("Nr","CENA.BRUTTO_[pln]","MOC_[km]",
                                   "POJEMNOSC.SKOKOWA_[cm3]",
                                   "ROK.PRODUKCJI","PRZEBIEG_[km]")]
```

Transformacja danych

Ze względu na to, że metody: sum, rang, Hellwiga bazują na różnych przekształceniach, tworzymy 3 podzbiory, które poddamy wcześniej już zaprezenowanym metodom stymulacji.

```
dane_rang<-stymulacja_przeksztalcenie_ilorazowe(dane_porzadkowanie,"PRZEBIEG_[km]")
dane_hellwig<-stymulacja_przeksztalcenie_ilorazowe(dane_porzadkowanie,"PRZEBIEG_[km]")
dane_sum<-stymulacja_przeksztalcenie_roznicowe(dane_porzadkowanie,"PRZEBIEG_[km]")
```

Teraz dokonamy przekształcenia normalizacyjnego wystymulowanych podzbiorów.

```
dane_sum<-unitaryzacja(dane_sum)
dane_rang<-unitaryzacja(dane_rang)
dane_hellwig<-standaryzacja(dane_hellwig)
```

4.3.9 Zastosowanie funkcji o dpowiedzialnych za porządkowania

Wykorzystamy tutaj wcześniej już wyjaśnione funkcje o dpowiedzialne za porządkowanie.

```
dane_sum<-funkcja_porzadkowanie_metoda_sum(dane_sum)
dane_rang<-funkcja_porzadkowanie_metoda_rang(dane_rang)
dane_hellwig<-funkcja_porzadkowanie_metoda_Hellwiga(dane_hellwig)
```

4.3.10 Porównanie wyników

W celu porównania wyników uporządkowania zostały stworzone trzy tabele pomocnicze. Każda tabela zawiera trzy kolumny, w dwóch pierwszych kolumnach znajdują się indeksy o biektów po uporządkowaniu. Z kolei w kolumnie trzeciej -„porownanie" znajduje się jedna z dwóch wartości: 0 lub 1. o dpowiednio wartość 1 jest przypisywana tym rekordom, dla których zgadza się kolejność uporządkowania przy zastosowaniu dwóch różnych metod porządkowania. Dodatkowo zostały również stworzone trzy tabele o nazwie „podsumowanie". W nich zliczane są wystąpienia wartości: 0 o raz 1 w kolumnie poprzedniej tabeli - „porownanie".

Para 1: metoda rang i metoda sum

```
tabela_porownawcza=data.frame(dane_rang,dane_sum)
names(tabela_porownawcza)<-c("dane_rang","dane_sum")
tabela_porownawcza$porownanie=0
#inwersja
for(i in 1:nrow(tabela_porownawcza))
{
  if(tabela_porownawcza$dane_sum[i]==tabela_porownawcza$dane_rang[i])
  {
    tabela_porownawcza$porownanie[i]=1
  }
}
head(tabela_porownawcza,15)
```

```
##   dane_rang dane_sum porownanie
## 1        49       49          1
## 2        50       50          1
## 3        53       53          1
## 4        48       16          0
## 5        52       48          0
```

```
## 6          16        52          0
## 7          57        38          0
## 8          38        57          0
## 9          37        37          1
## 10         58        58          1
## 11         44        44          1
## 12         17        46          0
## 13         56        17          0
## 14         20         1          0
## 15         46        41          0
```

```
#podsumowanie
podsumowanie=as.data.frame(table(tabela_porownawcza$porownanie))
names(podsumowanie)<-c("wartość","ilosc wystąpień")
podsumowanie
```

```
##   wartość ilosc wystąpień
## 1       0              51
## 2       1              10
```

Para 2: metoda rang i metoda Hellwiga

```
tabela_porownawcza=data.frame(dane_rang,dane_hellwig)
names(tabela_porownawcza)<-c("dane_rang","dane_hellwig")
tabela_porownawcza$porownanie=0
#inwersja
for(i in 1:nrow(tabela_porownawcza))
{
  if(tabela_porownawcza$dane_hellwig[i]==tabela_porownawcza$dane_rang[i])
  {
    tabela_porownawcza$porownanie[i]=1
  }
}
head(tabela_porownawcza,15)
```

```
##    dane_rang dane_hellwig porownanie
## 1         49           53          0
## 2         50           49          0
## 3         53           50          0
## 4         48           16          0
## 5         52           57          0
## 6         16           48          0
## 7         57           52          0
## 8         38           58          0
## 9         37           38          0
## 10        58           37          0
## 11        44           46          0
## 12        17           54          0
## 13        56           56          1
## 14        20           36          0
## 15        46            7          0
```

```r
#podsumowanie
podsumowanie=as.data.frame(table(tabela_porownawcza$porownanie))
names(podsumowanie)<-c("wartość","ilosc wystąpień")
podsumowanie
```

```
##   wartość ilosc wystąpień
## 1       0              56
## 2       1               5
```

Para 3: metoda sum i metoda Hellwiga

```r
tabela_porownawcza=data.frame(dane_sum,dane_hellwig)
names(tabela_porownawcza)<-c("dane_sum","dane_hellwig")
tabela_porownawcza$porownanie=0
#inwersja
for(i in 1:nrow(tabela_porownawcza))
{
  if(tabela_porownawcza$dane_hellwig[i]==tabela_porownawcza$dane_sum[i])
  {
    tabela_porownawcza$porownanie[i]=1
  }
}
head(tabela_porownawcza,15)
```

```
##    dane_sum dane_hellwig porownanie
## 1        49           53          0
## 2        50           49          0
## 3        53           50          0
## 4        16           16          1
## 5        48           57          0
## 6        52           48          0
## 7        38           52          0
## 8        57           58          0
## 9        37           38          0
## 10       58           37          0
## 11       44           46          0
## 12       46           54          0
## 13       17           56          0
## 14        1           36          0
## 15       41            7          0
```

```r
#podsumowanie
podsumowanie=as.data.frame(table(tabela_porownawcza$porownanie))
names(podsumowanie)<-c("wartość","ilosc wystąpień")
podsumowanie
```

```
##   wartość ilosc wystąpień
## 1       0              59
## 2       1               2
```

4.3.11 Podsumowanie

Na podstawie powyższych wyników zauważamy, że najwięcej zgodności wyniku porządkowania jest widoczne dla pary pierwszej - metody rang o raz metody sum. W przypadku kolejnych par, zgodność ta jest już niewielka.

Rozdział 5

Podsumowanie

W niniejszej pracy rozważone zostały wybrane zastosowania statystycznych metod porządkowania danych wielowymiarowych. Celem było porównanie tych metod praktycznej statystyki z matematyczną teorią porządków, jak również zaprezentowanie ich zastosowania w praktycznym przykładzie. W pracy dostrzeżono wiele analogii pomiędzy porządkami liniowymi i częściowymi, a rozważanymi metodami porządkowania liniowego o raz nieliniowego. Ponadto, w efekcie przeprowadzonych eksperymentów, zaobserwowano kilka własności wybranych metod porządkowania. Praca zawiera również wszystkie kody źródłowe przeprowadzonych eksperymentów, które w naszej o cenie mogą zostać ponownie wykorzystane w innych zastosowaniach. Podstawowe metody porządkowania danych wielowymiarowych, przypuszczają możliwość występowania w nich porządków liniowych. Przypuszczenie to jest co prawda wysoce niepewne, gdyż dla danych wielowymiarowych, bardziej spodziewane są porządki częściowe. Stosowanie zatem tych algorytmów, jeśli wyniki nie są wyraźnie jednoznaczne, może się o kazać wątpliwym. Na korzyść tych algorytmów przemawia natomiast wysoka zgodność z matematyczną teorią takich porządków. W pracy wykazaliśmy, że niesłusznym jest definiowanie porządków liniowych za pomocą jedynie 3 poszukiwanych własności. Jednak o bserwujemy, że algorytmy te pracując na danych, które posiadają naturalne uporządkowanie liniowe, istotnie zwracają zgodne i prawdziwe uporządkowania. W dalszej części rozpoznawaliśmy pozostałe algorytmy porządkowania danych, generujących inne matematyczne struktury. Aby w pełni zrozumieć zależności pomiędzy porządkami częściowymi, a algorytmami wyznaczającymi porządki nieliniowe w danych, niezbędna o kazała się podstawowa wiedza z zakresu teorii grafów. Matematyczna teoria porządków może być bowiem z łatwością przedstawiana na grafach nazywanych diagramami Hassego. Z drugiej strony algorytmy statystyki o kazują się poszukiwać w skończonych zbiorach danych właśnie struktur będących (lub przypominających) grafy. Reprezentowanie porządków na tych strukturach o kazuje się mieć dodatkowe atuty w postaci przejrzystej wizualizacji uzyskiwanych wyników. Występują tu jednak pewne istotne różnice. o bserwujemy, iż nie każdy porządek częściowy, mógłby za pomocą tychże algorytmy zostać wykryty. Algorytmy te bowiem generują grafy nazywane dendrytami, które nie wyczerpują rodziny wszystkich porządków częściowych. W rozdziale 4 zaprezentowaliśmy implementacje wybranych metod porządkowania liniowego, tj. wybraliśmy 3 różne metody: metodę rang, sum o raz Hellwiga. Metoda Hellwiga zaliczana jest do metod wzorcowych, czyli zakłada istnienie o biektu wzorcowego. Jako, że zmienne o pisujące o biekty (zgromadzone o ferty sprzedaży aut), są w większości stymulantami, stąd też przyjęliśmy koncepcję, że wartości współrzędnych o biektu wzorcowego przyjmą wartość maksymalną dla każdej zmiennej. Podkreślamy to gdyż zaobserwowaliśmy różnice uzyskiwane przez różne grupy metod. Metody wzorcowe generowały o dmienne porządki, w porównaniu z metodą sum czy też metodą rang. Wyniki, o których mówimy, znajdują się w sekcji 4.3.10. Jednakże przyglądając się wynikom porządkowania zawartym w sekcji 4.3.4, widać iż dla tych trzech metod o trzymujemy wyniki wspólne dla o biektów skrajnych, znajdujących się najwyżej lub najniżej w tym porządkowaniu. Różnice zauważamy głównie dla o biektów ze środka. Rozważone zostały też algorytmy nieliniowe, w szczególności dla metod aglomeracyjnych zaobserwowano, że istotna jest różnica w porządkach przy zastosowaniu o dległości liczonej wzorem najbliższego o raz najdalszego sąsiada. Umożliwiły nam o ne zobrazowanie wewnętrznych różnic pomiędzy tymi wersjami algorytmu porządkowania danych. Stąd też można zauważyć, że podobnie jak przy wynikach porządkowania liniowego dla małego zbioru, zaobserwować można zgodność uporządkowania tych metod dla o biektów skrajnych, znajdujących się najwyżej lub najniżej w tym uporządkowaniu. Wybrane wbudowane funkcje, o bok samodzielnie stworzonych funkcji porządkowania przedstawionych w sekcji 4.3.4, mogą zostać użyte do innych zbiorów danych, ze względu na ich o gólność i uniwersalność.

Bibliografia

[1] Grzegorz Banaszak, Wojciech Gajda. *Elementy algebry liniowej (część 1).* Wydawnictwo Naukowo-Techniczne, Warszawa, 2002.

[2] Jarosław Bartoszewicz. *Wykłady ze statystyki matematycznej.* Państwowe Wydawnictwo Naukowe, Warszawa, 1996.

[3] Aleksander Błaszczyk, Sławomir Turek. *Teoria mnogości.* Państwowe Wydawnictwo Naukowe, Warszawa, 2007.

[4] Patric Billingsley. *Prawdopodobieństwo i miara.* Państwowe Wydawnictwo Naukowe, Warszawa, 1987.

[5] Jacek Jakubowski, Rafał Sztencel. *Wstęp do teorii prawdopodobieństwa.* SCRIPT, Warszawa, 2004.

[6] Włodzimierz Krysicki, Jerzy Bartos, Wacław Dyczka, Krystyna Królikowska, Mariusz Wasilewski. *Rachunek prawdopodobieństwa i statystyka matematyczna w zadaniach: część I. rachunek prawdopodobieństwa.* Państwowe Wydawnictwo Naukowe, Warszawa, 1999.

[7] Kazimierz Kuratowski. *Wstęp do teorii mnogości i topologii.* Państwowe Wydawnictwo Naukowe, Warszawa, 2004.

[8] Andrzej Młodak. *Analiza taksonomiczna w statystyce regionalnej.* Centrum Doradztwa i Informacji Difin, Warszawa, 2006.

[9] Tomasz Panek, Jan Karol Zwierzchowski. *Statystyczne metody wielowymiarowej analizy porównawczej: teoria i zastosowania.* Oficyna Wydawnicza, Szkoła Główna Handlowa, Warszawa, 2013.

[10] Ryszard Rudnicki. *Wykłady z analizy matematycznej.* Państwowe Wydawnictwo Naukowe, Warszawa, 2006.

[11] Arkadiusz Sołtysiak, Piotr Jaskulski. `http://www.antropologia.uw.edu.pl/MaCzek/maczek.html`. dostęp: 02.07.2018.

[12] Robin J. Wilson. *Wprowadzenie do teorii grafów.* Państwowe Wydawnictwo Naukowe, Warszawa, 2008.