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, xxx 2018

Spis treści

Rozdział 1

Wstęp

Rozdział 2

Preliminaria

2.1 Notacja

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

- $\mathbb{R}$ - zbiór liczb rzeczywistych
- $\mathbb{N}$ - zbiór liczb naturalnych
- $\mathrm{O} = \{\mathrm{O}_1, \mathrm{O}_2, \ldots, \mathrm{O}_n\}$ - zbiór obiektów przestrzennych, tj. opisywanych przez wiele atrybutów, $n \in \mathbb{N}$
- $X = [x_{ij}]$ - macierz danych, gdzie x_{ij}-oznacza wartość j-tej zmiennej dla i-tego obiektu, gdzie $i = 1, \ldots, n$, $j = 1, \ldots, m,$, $n, m \in \mathbb{N}$
- X - w rozdziale drugim przez X najczęściej oznaczać będziemy dowolny zbiór, jednak w dalszych rozdziałach najczęściej służyć on będzie do opisywania macierzy zawierającej podstawowe (surowe) dane do analizy
- x^S - oznaczenie zmiennej mającej charakter stymulanty
- x^D - oznaczenie zmiennej mającej charakter destymulanty
- x^N - oznaczenie zmiennej mającej charakter nominanty
- Y - w rozdziale drugim używana jest najczęściej do oznaczenie zmiennej losowej
- Ω - oznaczenie dowolnej przestrzeni zdarzeń elementarnych ω, z rodziną podzbiorów $\mathcal{F}$
- $\mathcal{B}$ - rodzina zbiorów borelowskich
- $\mathfrak{B}$ - rodzina wszystkich zbiorów otwartych
- $\leqslant$ - oznaczenie relacji częściowego porządku, przy dołożeniu warunku spójności oznaczać będzie relację liniowego porządku
- G - oznaczenie ogólnego 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, których wytłumaczenie znajduje się poniżej.

- Statystyka matematyczna [5, w oparciu o rozdział 1] Statystyka matematyczna zajmuje się metodami wnioskowania o całej zbiorowości statystycznej na podstawie zbadania pewnej jej części zwanej próbką lub próbą.
- Model statystyczny [2, w oparciu o rozdział 2] Modelem statystycznym nazywamy przestrzeń próby doświadczenia tj. wartości zmiennych losowych o jednakowym rozkładzie, rodzinę podzbiorów zbioru zmiennych oraz prawdopodobieństwo występowania danej zmiennej. Można tutaj wskazać analogię do rachunku prawdopodobieństwa, tj. uporządkowanej trójki $(\Omega, \mathcal{F}, P)$.
- Cecha statystyczna [9, Rozdział 1] Cecha statystyczna jest to liczbowy opis przedmiotu dociekań tj. konkretnej dziedziny życia społeczno-gospodarczego.Służy ona do scharakteryzowania podmiotu badania.

Definicja 2.2.1. *Macierz [1, w oparciu o rozdział 1.1] Niech X będzie skończonym podzbiorem liczb $\mathbb{R}$. Macierzą wymiaru $m \times n$ (tzn. o m wierszach i n kolumnach) nazywamy prostokątną tablicę utworzoną z elementów zbioru X, postaci:*

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

gdzie: $a_{ij} \in X$, *dla* $1 \leqslant i \leqslant m, \quad 1 \leqslant j \leqslant n$

W analizie danych macierze wykorzystywane są nader często. Jest tak gdyż, pozwalają w jasny i przystępny sposób prezentować duże ilości danych. Wyszczególnia się więc wiele różnych rodzajów macierzy, pod względem rodzajów i cech danych które zawierają i prezentują. I tak zdefiniujmy w szczególności:

Definicja 2.2.2 (Macierz obserwacji [9, Rozdział 2])**.** *Niech $m > 1$ oraz $n > 1$ będą liczbami naturalnymi. Macierzą obserwacji nazywamy macierz rozmiaru $n \times m$ postaci*

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

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

Definicja 2.2.3. *Macierz odległości zmiennych [10, Rozdział 1.6] Macierzą odległości cech zmiennych nazywamy macierz, której elementami są odległości między parami badanych obiektów:*

$$D = [d_{ii'}].$$

gdzie:
dii'-odległość między i-tym a i'-tym obiektem, dla $i, i' = 1, 2, \dots, n$

W statystyce posługujemy się pojęciami skal do opisu różnych typów danych, które mogą podlegać naszej analizie. I tak zdefiniujmy następujące rodzaje skal:

- Skala [10, w oparciu o rozdział 1.2] Skalą nazywamy pewien skończony zbiór, umożliwiający porównywanie obiektów na podstawie wartości lub własności wybranych zmiennych.

- Skala porządkowa [10, Rozdział 1.2] Skala ta pozwala na stwierdzeniu o identyczności lub różnicy porównywanych obiektów, a także na porównywanie wariantów cech zaobserwowanych w obiektach. Nie pozwala ona określić odległości między obiektami. Umożliwia zliczanie obiektów uporządkowanych (liczby relacji równości, nierówności, większości i mniejszości). Przykład zmiennych przedstawianych na skali porządkowej: wykształcenie, kolejność zawodników na podium, oceny w systemie szkolnym.

- Skala przedziałowa [10, Rozdział 1.2] Jest to skala, która w stosunku do skali porządkowej, pozwala obliczyć odległość między obiektami, dokonując pomiaru cech za pomocą liczb rzeczywistych. Dla skali tej możliwe jest korzystanie z operacji dodawania oraz odejmowania. Dla skali tej istnieje charakterystyczna wartość - punkt zerowy. Jest on wyznaczany w sposób umowny, umożliwia on do zachowania różnic między wartościami cechy, przy zmianie jednostek miary. Przykład zmiennych przedstawianych na skali przedziałowej: temperatura, rok urodzenia.

- Skala ilorazowa [10, Rozdział 1.2] Skala ta, podobna jest do skali przedziałowej, z tym że występuje w niej zero bezwględne - punkt, który mówi o tym, że dana zmienna nie występuje, oraz ogranicza lewostronnie zakres skali ilorazowej. Powoduje to, że można na tej skali obok operacji dodawania i odejmowania, dokonywać także dzielenia i mnożenia, a tym samym przedstawiać dowolną wartość cechy danego obiektu jako wielokrotność wartości cechy dla innego obiektu. Przykład zmiennych przedstawianych na skali ilorazowej: napięcie elektryczne, bezrobocie, inflacja.

Definicja 2.2.4. *Stymulanta [10, Rozdział 1.5] Stymulantami nazywane są zmienne, których wysokie wartości badany w badanych obiektach są pożądane z punktu widzenia rozpatrywanego zjawiska.*

Definicja 2.2.5. *Destymulanta [10, Rozdział 1.5] Destymulantami nazywane są zmienne, których wysokie wartości badany w badanych obiektach są niepożadane z punktu widzenia rozpatrywanego zjawiska.*

Definicja 2.2.6. *Nominanta [10, Rozdział 1.5] Nominantami nazywane są zmienne, których odchylenia wartości w badanym obiekcie od wartości (lub przedziału wartości) uznawanych za najkorzystniejsze są niepożądane z punktu widzenia rozpatrywanego zjawiska.*

2.2.1 Wybrane operacje statystyczne dla zmiennych

Normalizacja zmiennych

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

- standaryzacja,
- unitaryzacja,
- przekształcenie ilorazowe

- rangowanie zmiennych

1. Celem standaryzacji jest otrzymanie zmiennych o odchyleniu standardowym równym 1 i wartości oczekiwanej równej 0. W tym celu dla każdej zmiennej, będącej cechą obiektu oblicza się odchylenie standardowe na podstawie wartości tej zmiennej dla wszystkich obiektów, a także wartość oczekiwaną na tej samej zasadzie. W kolejnym kroku dla każdego obiektu liczymy jego znormalizowana wartość tj. od wartości zmiennej odejmujemy wartość oczekiwaną dla danej cechy, a otrzymaną różnicę dzielimy przez odchylenie standardowe dla tej cechy,co można to zapisać w postaci

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

   gdzie z_{ij}-wartość znormalizowanej zmiennej j-tej zmiennej, w i-tym obiekcie.

2. W pracy skorzystałam również z unitaryzacji, stosowanej w celu uzyskania zmiennych o ujednoliconym zakresie zmienności, u mnie przedział ten to $[0, 1]$. W tym celu od wartości zmiennej w obiekcie, odejmuje się minimalną wartość występującej dla tej cechy, a następnie różnica ta dzielona jest przez różnicę między maksymalną a minimalną wartością zmiennej, poddanej normalizacji. Znormalizowana zmienna jest postaci

$$z_{ij} = \frac{x_{ij} - \min_i\{x_{ij}\}}{\max_i\{x_{ij}\} - \min_i\{x_{ij}\}}, \quad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, m.$$

3. Kolejna metoda normalizacyjna, wykorzystana w pracy to przekształcenie ilorazowe. Stosuje się je w celu odniesieniu wartości zmiennej do ustalonej wartości - może to być wartość oczekiwana danej zmiennej na tle analizowanych obiektów, wartość minimalna lub maksymalna tej cechy. W pracy za tę wartość przyjęłam wartość oczekiwaną zmiennej. Każda wartość zmiennej dla danego obiektu jest dzielona przez wartość oczekiwaną tej zmiennej, a postać znormalizowanej zmiennej to

$$z_{ij} = \frac{x_{ij}}{\mathrm{E}(x_j)}, \quad \mathrm{E}(x_j) \neq 0 \quad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, m.$$

4. Przy zastosowaniu metody rang, wykorzystuje się normalizację rangową. Stosuje się ją przeważnie, gdy zmienne opisujące obiekty mierzone są na skali porządkowej. W pierwszym kroku wartości zmiennych opisujących obiekty 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 obiekty w uporządkowanym zbiorze. Postać zmiennej znormalizowanej rangowo

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

   gdzie h-ranga nadana i-temu obiektowi znajdującemu się na h-tym miejscu w uporządkowanym zbiorze, ze względu na wartość j-tej zmiennej.

Stymulacja zmiennych

W celu ujednolicenia charakteru danych 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 oraz przekształcenie różnicowe. W zależności od skali na której mierzone są zmienne, należy stosować odpowiednie przekształcenie stymulacyjne.

W przypadku przekształcenia ilorazowego można go stosować tylko dla zmiennych mierzonych na skali ilorazowej. Stymulacja destymulant wygląda nastepująco

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

dla nominant

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

Dla zmiennych mierzonych na skali ilorazowej czy też przedziałowej. Stymulacja dla destumulant prezetuje się następująco

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

dla nominant

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

2.3 Podstawowe pojęcia rachunku prawdopodobieństwa oraz statystyki

Na potrzeby pracy, zostały wykorzystane pojęcia rachunku prawdopodobieństwa oraz 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 [11, Rozdział 8.1] Rodzinę $\mathcal{F}$ podzbiorów, niepustego zbioru Y nazywamy ciałem zbiorów, jeżeli spełnia ona następujące warunku:*

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

Definicja 2.3.2. *σ-algebra/ciało zbiorów[11, Rozdział 8.1] Ciało zbiorów $\mathcal{F}$ nazywamy σ-ciałem zbiorów, jeżeli spełnia ona warunek dla dowolnych zbiorów $A_n \in \mathcal{F}, n \in \mathbb{N}$, mamy $\bigcup\limits_{i=1}^{\infty} A_n \in \mathcal{F}$.*

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

Definicja 2.3.4. *Przestrzeń zdarzeń elementarnych [7, w oparciu 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 2.3.5. *Miara zbioru [4, Rozdział 2.10] Funkcję μ określoną na ciele $\mathcal{F}$ podzbiorów zbioru Ω nazywamy miarą, jeśli spełnia następujące warunki:*

1. $\mu(A) \in [0, \infty]$ *dla każdego zbioru* $A \in \mathcal{F}$,

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

Definicja 2.3.7. *Funkcja mierzalna [11, w oparciu o rozdział 8.2] Niech* Y *będzie niepustym zbiorem,* $\mathcal{F}$ σ*-ciałem na* Y *i* $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$. *Funkcję* $f : Y \to \overline{\mathbb{R}}$ *nazywamy mierzalną, jeżeli zbiór* $\{y \in Y : f(y) > a\}$
jest mierzalny przy dowolnym $a \in \mathbb{R}$.

Definicja 2.3.8. *Zdarzenie losowe [7, w oparciu 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 [7, w oparciu o rozdział 1.1] Prawdopodobieństwem nazywamy dowolną funkcję* P *o wartościach rzeczywistych, określoną na* σ*-ciele zdarzeń* $\mathcal{F} \subset 2^{\Omega}$, *spełniającą warunki:*

1. $P(A) \geqslant 0 \quad \forall A \in \mathcal{F}$

2. $P(\Omega) = 1$

3. *Jeśli* $A_i \in \mathcal{F}$, $i = 1, 2, ...$ *oraz* $A_i \cap A_j$ *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 [7, w oparciu 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 określonym na* $\mathcal{F}$.

Definicja 2.3.11. *Zmienna losowa [7, 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*
$\{\omega : Y(\omega) < y\} \in \mathcal{F}$ *dla każdego* $y \in \mathbb{R}$

Definicja 2.3.12. *Wektor losowy [6, Rozdział 5.1] Wektorem losowym nazywamy odwzorowanie* $Y : \Omega \to \mathbb{R}^n$*, 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* Y *[6, Rozdział 5.1] Rozkładem prawdopodobieństwa zmiennej losowej* Y *o wartościach w* $\mathbb{R}$ *nazywamy funkcję* μ_Y *określoną 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 [6, Rozdział 5.1] Mówimy, że zmienna losowa* Y *ma rozkład dyskretny, jeśli istnieje przeliczany zbiór* $S \in \mathbb{R}$*, taki że* $\mu_Y(S) = 1$.

Definicja 2.3.15. *Gęstość i rozkład ciągły [6, 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 oczekiwana [7, Rozdział 2.6] Niech* X *będzie zmienną losową typu dyskretnego lub ciągłego. Wartością oczekiwaną zmiennej losowej* X *nazywamy*

$$\mathrm{E}(X) = \begin{cases} \sum\limits_{i=1}^{n} x_i p_i & \textit{, jeśli zmienna ma rozkład dyskretny} \\ \int\limits_{-\infty}^{\infty} x f(x)dx & \textit{, jeśli zmienna ma rozkład ciągły} \end{cases}$$

Definicja 2.3.17. *Wariancja [6, Rozdział5.6] Niech* Y *będzie zmienną losową. Jeśli* $\mathrm{E}(Y - \mathrm{E}Y)^2 < \infty$*, to liczbę tę nazywamy wariancją zmiennej losowej* Y *o o wartościach rzeczywistych i oznaczamy:*
$\mathrm{Var}\, Y = \mathcal{D}^2 Y = \mathrm{E}(Y - \mathrm{E}Y)^2$.

Definicja 2.3.18. *Odchylenie standardowe [6, Rozdział 5.6] Niech* Y *będzie zmienną losową. Odchyleniem standardowym zmiennej losowej* Y *nazywamy pierwiastek z wariancji.*
$\sigma_Y = \mathcal{D}(Y) = \sqrt{\mathcal{D}^2 Y}$

Definicja 2.3.19. *Rozkład normalny (Gaussa) [6, 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 Podstawowe pojęcia teorii grafów

W pracy zostaną opisane zarówno metody porządkowania liniowego jak i nieliniowego tzn. w ujęciu matematycznym - porządku częściowego, w tym celu należy wprowadzić definicje związane z teorią grafów, niezbędne przy opisywaniu metod porządkowania nieliniowego.

W celu wprowadzeniu złożonych definicji, należy wcześniej podać podstawowe pojęcia dotyczące grafów. Uprzednio zostanie jeszcze wprowadzona definicja pary uporządkowanej oraz nieuporządkowanej, gdyż pojęcia te zostały wykorzystane w definicji grafu.

Definicja 2.4.1. *Para uporządkowana [8, w oparciu o rozdział 3] Niech dane będą dwa elementy a i b. Parą uporządkowaną nazywamy parę postaci $< a, b >$, gdzie element a jest poprzednikiem, zaś element b jest następnikiem.*
$< a, b >= \{\{a\}, \{a, b\}\}$.

Definicja 2.4.2. *Para nieuporządkowana [8, w oparciu o rozdział 3] Niech dane będą dwa elementy a i b. Parą nieuporządkowaną nazywamy zbiór postaci $\{a, b\}$, 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 utworzą krawędź, a parą zbiorów definiującą graf.

Definicja 2.4.3. *Graf [12, w oparciu 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 2.4.4. *Pętle [12, Rozdział 2] Pętlami w grafie nazywamy krawędzie reprezentowane przez $\{a\}$ gdzie a jest pewnym wierzchołkiem, tj. łączące wierzchołek z samym sobą.*

Definicja 2.4.5. *Wierzchołki sąsiednie, krawędzie 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 2.4.6. *Krawędzie sąsiednie [12, Rozdział 2] Dwie krawędzie e i f grafu G są sąsiednie, jeśli mają wspólny wierzchołek, tj*

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

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

Definicja 2.4.7. *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 od wierzchołka v_0 do wierzchołka v_m. Liczbę krawędzi na trasie nazywamy długością trasy.

Definicja 2.4.8. *Ścieżka [12, Rozdział 3] Trasą, w której wszystkie krawędzie są różne, nazywamy ścieżką.*

Definicja 2.4.9. *Droga [12, Rozdział 3] Ścieżkę, w której wierzchołki $v_0, v_1, \ldots, v_m$ są różne (z wyjątkiem, być może, równości $v_0 = v_m$), nazywamy drogą.*

Definicja 2.4.10. *Droga zamknięta/ścieżka zamknięta [12, Rozdział 3] Droga lub ścieżka jest zamknięta, jeśli $v_0 = v_m$.*

Definicja 2.4.11. *Cykl [12, Rozdział 3] Cyklem nazywamy drogę zamkniętą lub ścieżkę zamkniętą, w której jedynym powtarzającym się wierzchołkiem jest jej początek, będący jednocześnie końcem.*

Definicja 2.4.12. *Graf spójny [12, Rozdział 3] Graf jest spójny wtedy i tylko wtedy, gdy każda para wierzchołków jest połączona drogą.*

Definicja 2.4.13. *Drzewo [12, Rozdział 4] Drzewem nazywamy graf spójny, nie zawierający cykli.*

Definicja 2.4.14. *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 oznaczamy zwykle w postaci pary uporządkowanej $G =< V, E >$*

Uwaga 1. *Każdy graf jednoznacznie wyznacza pewną relację dwuargumentową (binarną) w zbiorze V. Można również powiedzieć odwrotnie, że każda relacja dwuargumentowa (binarna) r w zbiorze V, wyznacza jednoznacznie graf zorientowany, którego węzłami są elementy zbioru V, z kolei krawędziami są uporządkowane pary (v, v'), należące do r.*

Uwaga 2. *Niech dany będzie digraf D składający się z niepustego zbioru skończonego wierzchołków $V(D)$ i skończonej rodziny krawędzi$E(D)$. W momencie gdy $\forall a, b \in V(D) \quad < a, b > \in E(D)$ oraz $< b, a > \in E(D)$, to taki graf skierowany, staje się grafem niezorientowanym.*

$$\begin{tikzcd} a \arrow[d, leftrightarrow] \\ b \end{tikzcd}$$

2.5 Wybrane pojęcia z teorii mnogości, topologii i algebry liniowej

2.5.1 Relacja porządkująca

W niniejszej pracy skupiam się na zagadnieniu porządkowania danych wielowymiarowych. Konieczne jest zatem przywołanie odpowiednich 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 ze względu na korzystanie z tych pojęć, przy definiowaniu właściwości porządkowania liniowego zbioru obiektów. Dodatkowo przytoczona została definicja zbioru skończonego, ze względu na zastosowanie metod porządkowania na zbiorze skończonym.

Definicja 2.5.1. *Relacja [8, 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 2.5.2. *Relacja porządkująca (częściowego porządku) [3, Rozdział 2] Niech dana relacja ρ, którą oznaczać będziemy przez $\leqslant$, będzie określona 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),*

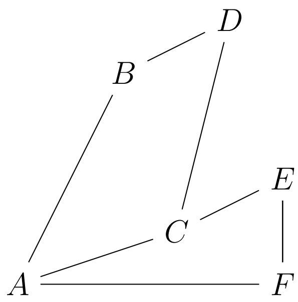
3. *jeśli $x \leqslant y$ i $y \leqslant z$, to $x \leqslant z$ (przechodniość).*

Przykład 1. *Częściowego porządku na zbiorze Wykorzystanie częściowego porządku na płaszczyźnie $\mathbb{R}^2$, obrazuje diagram Hassego, będący grafem skierowanym, którego wierzchołki zostały poddane relacji porządkowania i reprezentują elementy skończonego zbioru $X \subset \mathbb{R}$. Aby go skonstruować, należy postępować według poniższych kroków:*

- *Punkty obrazujące elementy zbioru X, umieszcza się na płaszczyźnie.*

- *Punkt $x \in X$ łączony jest odcinkiem z punktem $y \in X$, jeśli x jest następnikiem y, czyli gdy $y < x$ oraz nie istnieje taki punkt $z \in X$, że $y < z < x$.*

Zobrazowanie relacji częściowego porządku, dla punktów $A, B, C, D, E, F \in \mathbb{R}^2$



Definicja 2.5.3. *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 ona następujący warunek spójności tzn.*

$x \leqslant y$ lub $y \leqslant x$

Parę $(X, \leqslant)$ nazywamy zbiorem liniowo uporządkowanym lub łańcuchem.

Definicja 2.5.4. *Dobry porządek [3, Rozdział 2] Niech dany będzie zbiór X. Relację $\leqslant$ porządkującą zbiór X, nazywamy dobrym porządkiem na zbiorze X, gdy w każdym niepustym podzbiorze zbioru X istnieje element najmniejszy względem relacji $\leqslant$. Jeśli relacja $\leqslant$ na zbiorze X jest dobrym porządkiem, to mówimy, że para $(X, \leqslant)$ jest zbiorem dobrze uporządkowanym.*

Definicja 2.5.5. *Elementy wyróżnione [3, Rozdział 2] Niech X będzie zbiorem częściowo uporządkowanym przez relacje $\leqslant$ oraz niech $a \in X$. Mówimy, że:*

1. *a jest elementem najmniejszym w X, gdy $\forall x \in X \quad a \leqslant x$*

2. *a jest elementem minimalnym w X, gdy $\neg((\exists x \in X) \quad x < a)$*

3. *a jest elementem największym w X, gdy $\forall x \in X \quad x \leqslant a$*

4. *a jest elementem maksymalnym w X, gdy $\neg((\exists x \in X) \quad a < x)$*

Definicja 2.5.6. *Ograniczenie górne [3, Rozdział 2] Niech $A \subseteq X$, gdzie $(X, \leqslant)$ jest zbiorem uporządkowanym. Element $x \in X$ nazywamy ograniczeniem górnym zbioru A względem relacji $\leqslant$, gdy $a \leqslant x$ dla każdego $a \in A$.*

Definicja 2.5.7. *Ograniczenie dolne [3, Rozdział 2] Niech $A \subseteq X$, gdzie $(X, \leqslant)$ jest zbiorem uporządkowanym. Element $y \in X$ nazywamy ograniczeniem dolnym zbioru A względem relacji $\leqslant$, gdy $y \leqslant a$ dla każdego $a \in A$.*

Definicja 2.5.8. *Zbiór ograniczony z góry, zbiór ograniczony z dołu, zbiór ograniczony [3, Rozdział 2] Niech $A \subseteq X$, gdzie $(X, \leqslant)$ jest zbiorem uporządkowanym. Zbiór nazywamy ograniczonym z góry (ograniczonym z dołu), jeśli ma on ograniczenie górne (dolne).*
Zbiór ograniczony z dołu i z góry nazywamy ograniczonym.

Definicja 2.5.9. *Kres górny [3, Rozdział 2] Niech $A \subseteq X$, gdzie $(X, \leqslant)$ jest zbiorem uporządkowanym. Jeśli zbiór A jest ograniczony z góry i wśród ograniczeń 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. *$a \leqslant x_0$ dla każdego $a \in A$,*
2. *jeśli $a \leqslant x$ dla każdego $a \in A$, to $x_0 \leqslant x$.*

Definicja 2.5.10. *Kres dolnym [3, Rozdział 2] Niech $A \subseteq X$, gdzie $(X, \leqslant)$ jest zbiorem uporządkowanym. Jeśli zbiór A jest ograniczony z dołu i wśród ograniczeń dolnych zbioru A istnienie element największy x_0, to element ten nazywamy kresem dolnym zbioru A i oznaczamy symbolem $infA$. Tak więc $x_0 = infA$, gdy spełnione są następujące warunki:*

1. *$y_0 \leqslant a$ dla każdego $a \in A$,*
2. *jeśli $y \leqslant a$ dla każdego $a \in A$, to $y \leqslant y_0$.*

Definicja 2.5.11. *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 2.5.12. *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 on n-elementowy, tj. mocy równej n.*

Definicja 2.5.13. *Zbiór przeliczalny [3, Rozdział 5] Mówimy, że zbiór X jest przeliczalny, gdy jest skończony lub jest równoliczny z $\mathbb{N}$.*

2.5.2 Przestrzenie metryczne, miary odległości

Niezbędnym jest również wprowadzenie podstawowych pojęć z topologii, ze względu na stosowanie funkcji odległości w celu uporządkowania obiektów.

Definicja 2.5.14. *Metryka [8, Rozdzial 9] Niech X będzie niepustym zbiorem, wtedy funkcję $\mathrm{d} : X \times X \to [0, \infty)$, nazywamy metryką jeśli spełnione są warunki:*

1. $\forall x, y \in X \quad \big(\mathrm{d}(x, y) = 0 \iff x = y\big)$,
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 2.5.15. *Przestrzeń metryczna [8, Rozdział 9] Niech X będzie niepustym zbiorem, d metryką, wówczas parę (X, d) nazywamy przestrzenią metryczną.*

Przykład 2. *Metryka euklidesowa w* $\mathbb{R}^2$ *Niech* $\mathrm{d}_e : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ *będzie 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 3. *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 4. *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*

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

2.5.3 Rachunek macierzowy

Definicja 2.5.16. *Grupa [1, Rozdział 0] Grupą nazywamy zbiór* P *z działaniem* $\cdot : P \times P \to P$ *dla którego są spełnione następujące warunki:*

1. *Dla dowolnych* $a, b, c \in P \quad (a \cdot b) \cdot c = a \cdot (b \cdot c)$ *(łączność)*
2. *Istnieje element* $e \in P$ *(nazywany elementem neutralnym grupy), taki że dla każdego* $a \in P$ *mamy,* $e \cdot a = a \cdot e = a$
3. *Dla każdego* $a \in P$ *istnieje element* $b \in P$ *(nazywany elementem odwrotnym do* a*), taki że* $a \cdot b = b \cdot a = e$*.*

Dodatkowo, mówimy że grupa P *jest przemienna, gdy dla dowolnych* $a, b \in P \quad a \cdot b = b \cdot a$*.*

Definicja 2.5.17. *Pierścień [1, Rozdział 0] Pierścieniem nazywamy zbiór* R *z dwoma działaniami: z dodawaniem* $+ : R \times R \to R$ *i z mnożeniem* $\cdot : R \times R \to R$*, dla których są spełnione następujące warunki:*

1. *Zbiór* R *z działaniem* $+$ *jest grupą przemienną. Element neutralny działania* $+$*oznaczamy przez 0.*
2. *Dla dowolnych* $a, b, c \in R \quad (a \cdot b) \cdot c = a \cdot (b \cdot c)$ *(łączność mnożenia).*
3. *Dla dowolnych* $a, b, c \in R$

$$a \cdot (b + c) = a \cdot b + a \cdot c$$
$$(a + b) \cdot c = a \cdot c + b \cdot c$$

(rozdzielność mnożenia względem dodawania).

Definicja 2.5.18. *Ciało [1, Rozdział 0] Ciałem nazywamy zbiór* K *z działaniami:* $+ : K \times K \to K$ *oraz* $\cdot : K \times K \to K$*, takimi że:*

1. *Zbiór* K *z działaniami* $+$ *i* $\cdot$ *jest pierścieniem przemiennym z jedynką.*
2. *Zbiór* $K^* = K \backslash \{0\}$ *z działaniem* $\cdot$ *jest grupa.*

Definicja 2.5.19. *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 5. *Macierz A o m-wierszach i n-kolumnach najczęściej zapisuje się postaci* $A = [a_{ij}]_{i \leqslant m, j \leqslant n}$, *tj.*

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

Rozdział 3

Metody porządkowania

Rozdział ten został opracowany w oparciu o [10, Rozdział 2], omówiono w nim ogólnie metody porządkowania zarówno liniowego jak i nieliniowego, po to by w kolejnym rozdziale szczegółowo przyjrzeć się wybranym metodą wraz z przedstawieniem ich algorytmów oraz dokładnych opisów matematycznych, z uwzględnieniem ich wad oraz zalet.

Metody porządkowania liniowego pozwalają na ustalenie hierarchii obiektów ze względu na określone kryterium. Z kolei metody porządkowania nieliniowego nie pozwalają na ustaleniu hierarchii, natomiast w wyniku uporządkowania możliwe jest wskazanie dla każdego z obiektów poddanych porządkowaniu, wskazaniu obiektów podobnych ze względu na opisujące je zmienne.

3.1 Metody porządkowania liniowego

Porządkowanie liniowe obiektów polega, w ujęciu geometrycznym, na rzutowaniu na prostą punktów reprezentujących obiekty, umieszczonych w wielowymiarowej przestrzeni zmiennych. Takie postępowanie pozwala na ustalenie hierarchii obiektów, czyli uporządkowanie ich od obiektu stojącego najwyżej w tej hierarchii do obiektu znajdującego się najniżej. Poniżej zostaną przedstawione własności uporządkowania liniowego obiektów, wraz z podaniem ich matematycznej interpretacji.

- każdy obiekt ma przynajmniej jednego sąsiada i nie więcej niż dwóch sąsiadów,
- jeżeli sąsiadem i-tego obiektu jest i'-ty obiekt, to jednocześnie sąsiadem i'-tego obiektu jest i-ty obiekt,
- dokładnie dwa obiekty mają tylko jednego sąsiada.

Powyżej wymienione własności są wynikiem posiadania jedynie skończonej ilości obiektów, które podane są uporządkowaniu. W następnej części chciałabym

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

Na początku zacznę od 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. Mówi się, 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 $X \times X$, gdzie X jest zbiorem ze skończoną liczbą obiektów, 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 oznacza 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 obiekt 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 2.5.3 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 od obu $x \neq z \neq y$ 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 ogólności 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 oznaczeń możemy przyjąć, że będą nimi a oraz b. Ustalmy zatem, że $a \leqslant x, b \leqslant x$. Wobec definicji 2.5.3 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. Zupełnie analogicznie postępujemy dla przypadku $x \leqslant a, x \leqslant b$. Zatem uzyskujemy sprzeczność, będącą efektem przypuszczenia, że mogą istnieć takie 3 elementy a, b, c. Zatem ostatecznie $\overline{\overline{\{y \in X, ySx\}}} \in \{1, 2\}$.

2. Niech $x, y \in X$ oraz 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 oznacza, 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. Załóżmy, 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, oczywiście 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 $x_m \leqslant y_m$ oraz $y_m \leqslant x_m$. Wobec antysymetryczności z definicji 2.5.2 mamy, że $x_m = y_m$ wbrew naszemu przypuszczeniu, że są od siebie różne. Pozostaje pokazać, że element minimalny ma pojedynczego sąsiada. Przypuśćmy, że $y, z \in X$ są dwoma różnymi sąsiadami dla x_m. Wtedy $x_m \leqslant y \vee y \leqslant x_m$ oraz $x_m \leqslant z \vee z \leqslant x_m$. Skoro x_m jest minimalny to musi to oznaczać, że

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

  Wobec spójności z definicji 2.5.3 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 $x_M \leqslant y_M$ oraz $y_M \leqslant x_M$. Wobec antysymetryczności z definicji 2.5.2, mamy że $x_M = y_M$, co wbrew naszemu przypuszczeniu daje, że elementy te są od siebie różne. Pozostaje pokazać, że element maksymalny ma pojedynczego sąsiada. Przypuśćmy, że $y, z \in X$ są dwoma różnymi sąsiadami dla x_M. Wtedy $x_M \leqslant y \vee y \leqslant x_M$ oraz $x_M \leqslant z \vee z \leqslant x_M$. Skoro x_M jest minimalny to musi to zatem oznaczać

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

  Wobec spójności z definicji 2.5.3 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 oraz 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 $x \leqslant x_j$, który będzie sąsiadem z x, zatem xSx_j. Zatem ostatecznie 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 ostatecznie y_iSx i xSx_M, co przeczy założeniu że x ma pojedynczego sąsiada.

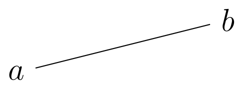
  Zajmijmy się teraz trzecim przypadkiem, tj. gdy ani x_m oraz x_M nie są sąsiadami elementu x. Z faktu, iż zbiór X posiada element minimalny x_m, który nie jest sąsiadem elementu x, wynika że istnieje taki element $x_1 \in X$, że $x_m \leqslant x_1 \leqslant x$. Co więcej istnieje taki $x_2 \in X$, że $x_m \leqslant x_1 \leqslant x_2 \leqslant x$. 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.

□

Powyżej wykazane własności są często podawane, niemal na równi z definicją takiego uporządkowania, w pozycjach książkowych omawiających praktyczny aspekt porządkowania danych. Poniżej podane zostaną przykłady takich relacji, które też posiadają powyższy zestaw własności, ale nie opisują relacji będących porządkami liniowymi.

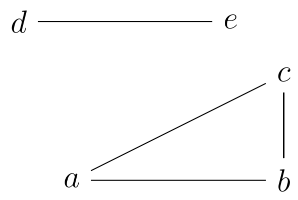
Przykład 6. *Rozważmy zbiór dwuelementowy $X = \{a, b\}$ gdzie $a \leqslant b$ jest jedynym punktem tej relacji. Tak zdefiniowana relacja spełnia wszystkie własności, ale nie spełnia założenia o zwrotności - zatem relacja ta nie jest liniowym porządkiem.*

Diagram Hassego prezentujący tę relację, jest postaci:



Przykład 7. *Rozważmy zbiór $X = \{a, b, c, d, e\}$ oraz 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:



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

W podsumowaniu tej sekcji należy podkreślić, że by uporządkować liniowo obiekty, charakteryzujące je zmienne muszą być mierzone przynajmniej na skali porządkowej. Gdy zmienne te mierzone są na skali przedziałowej lub ilorazowej, należy dokonać ich normalizacji, dla zapewnienia ich porównywalności.

Metody porządkowania liniowego można podzielić na metody diagramowe, procedury oparte na zmiennej syntetycznej oraz procedury iteracyjne bazujące funkcji kryterium dobroci uporządkowania tzn. funkcji, którą się przyjmuje, lub też tworzy się ją aby w kolejnych iteracjach szukać takiego uporządkowania, które optymalizuje zbiór wartości tej funkcji. W kolejnej sekcji zostaną pokrótce przedstawione różne metody, z wyszczególnieniem najważniejszych założeniach każdej z nich.

3.1.1 Metody diagramowe

W metodach diagramowych stosuje się graficzną reprezentację macierzy odległości zwanej diagramem. Macierz konstruowana jest w oparciu o odległości między obiektami, wyznaczone za pomocą dowolnej metryki. Porządkowanie obiektów polega na porządkowaniu diagramu, tj. przestawieniu wierszy i odpowiadających im kolumn diagramu, tak aby symbole graficzne reprezentujące najmniejsze odległości skupiały się wzdłuż głównej przekątnej, zaś w miarę oddalania się od głównej przekątnej znajdowały się symbole graficzne odpowiadające coraz to większym odległością.

Jako narzędzie pomocnicze w porządkowaniu danych, może stanowić kryterium postaci:

$$F^1 = \sum_{i=1}^{n} \sum_{i'>1}^{n} d_{ii'} w_{ii'}$$

gdzie:

$d_{ii'}$ - odległość euklidesowa między i-tym i i'-tym obiektem .

$w_{ii'}$ - wagi elementów macierzy odległości, zdefiniowane w oparciu o jeden z następujących wzorów:

$$w_{ii'} = \frac{|i - i'|}{n - 1},$$

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

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

Dodatkowo wagi elementów macierzy odległości tworzą macierz wag postaci:
$W = [w_{ii'}], \quad i, i' = 1, 2, \ldots, n.$

3.1.2 Metody oparte na zmiennych syntetycznych

W tym podrozdziale zostaną opisane metody porządkowania oparte na zmiennych syntetycznych, tj. funkcji których wartości będą służyć do porządkowania danych. Metody oparte na zmiennych syntetycznych dzielimy na wzorcowe i bezwzorcowe. Poniżej zostaną one opisane szczegółowo, jednak wcześniej zostaną przedstawione wzory wyznaczające zmienną syntetyczną.

Sposoby wyznaczania zmiennej syntetycznej

1. dla średniej arytmetycznej:

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

2. dla średniej geometrycznej:

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

3. dla średniej harmonicznej

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

gdzie: s_i - wartość zmiennej syntetycznej w i-tym obiekcie, w_j - waga j-tej zmiennej.

Metody bewzorcowe

W metodach tych, unormowane wartości podanych zmiennych wejściowych są uśrednianie, przez przypisywanie im odpowiednich wag. Poniżej zostaną omówione wybrane metody porządkowania bezwzorcowego.

Metoda rang

Metoda ta bazuje na normalizacji rangowej, w związku z tym zmienne mierzone są na skali porządkowej. Dla każdego obiektu wyznacza się sumę przyporządkowanych mu rang ze względu na wszystkie zmienne. Na końcu obliczana jest wartość zmiennej syntetycznej, jako średniej wartości rang. W oparciu o tę wartość następuje porządkowanie obiektów. Wzór na obliczenie wartości zmiennej syntetycznej:

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

gdzie:
z_{ij}-zmienna znormalizowana rangowo, tj. $z_{ij} = h$ dla $x_{hj} = x_{ij}, \quad h, i = 1, 2, \ldots, n$. gdzie:
h-ranga nadana i-temu obiektowi znajdującemu się na h-tym miejscu w uporządkowanym szeregu obiektów ze względu na j-tą zmienną.

Metoda sum

Metoda ta bazuje na konstrukcji zmiennej syntetycznej przy pomiarze zmiennych na skali ilorazowej lub przedziałowej. Dla każdego obiektu obliczana jest wartość zmiennej syntetycznej, jako średnia arytmetyczna wartości zmiennych przy przyjęciu jednakowych wag dla każdej zmiennej. Eliminowane są ujemne wartości zmiennej syntetycznej przy wykorzystaniu przekształcenia:

$$s_i' = s_i - \min\{s_i\}, \quad i = 1, 2, \dots, n.$$

Końcowa postać zmiennej syntetycznej otrzymywana jest po przeprowadzeniu normalizacji według wzoru:

$$s_i'' = \frac{s_i'}{\max\{s_i'\}}, \quad i = 1, 2, \dots, n.$$

Powyższe przekształcenia powodują unormowanie miary syntetycznej w przedziale [0,1]. Powyższa wielkość wykorzystywana jest do uporządkowania obiektów.

Metoda wzorcowe

W metodach tych zakłada się istnienie obiektu wzorcowego $O_0 = [z_{0j}], \quad j = 1, 2, \dots, m$, w którym zmienne wejściowe $z_{0j}, j = 1, 2, \dots, m$, będące współrzędnymi obiektu wzorcowego, przyjmują optymalne wartości, które to mogą być ustalane na podstawie ogólnie przyjętych norm, subiektywnej opinii dotyczącej obserwowanego obiektu, lub też opinii ekspertów. Poszczególne metody różnią się sposobem wyznaczania obiektu wzorcowego, poniżej zostaną one przedstawione.

Metoda Hellwiga

W metodzie tej, obiekt wzorcowy wyznaczony jest na podstawie wystandaryzowanych zmiennych wejściowych. Współrzędnym obiektu wzorcowego przyporządkowuje się maksimum, gdy zmienne wejściowe są stymulantami lub minimum gdy zmienne są destymulantami. Obiekty są uporządkowywane na podstawie odległości od obiektu wzorcowego, przy wykorzystaniu odległości euklidesowej. Miara syntetyczna jest postaci:

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

gdzie:
współrzędne obiektu wzorcowego są obliczane na podstawie wzoru:

$$z_{0j} = \begin{cases} \max\limits_i\{z_{ij}\} & \text{dla } z_j^S, \quad j = 1, 2, \dots, m, \quad i = 1, 2, \dots, n, \\ \min\limits_i\{z_{ij}\} & \text{dla } z_j^D, \quad j = 1, 2, \dots, m, \quad i = 1, 2, \dots, n, \end{cases}$$

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

$$d_0 = \overline{d_0} + 2S(d_0),$$

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

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

Wartości miary s_i zazwyczaj są z przedziału $[0, 1]$. Należy tu zaznaczyć, że wartości miary są tym wyższe, im mniej jest oddalony obiekt od obiektu wzorcowego.

Metoda Walesiaka

Metoda ta bazuje na konstrukcji zmiennej syntetycznej w oparciu o badanie odległości obiektów od obiektu wzorcowego , przy wykorzystaniu uogólnionej miary odległości. Umożliwia ona porządkowanie obiektów, jeżeli opisujące je charakterystyki są mierzone przynajmniej na skali porządkowej. W takim przypadku, zmienne wejściowe o postaci nominant muszą zostać podaje stymulacji. Z kolei gdy zmienne są mierzone na skali przedziałowej lub ilorazowej, należy je znormalizować. Miara syntetyczna oparta na uogólnionej mierze odległości przyjmuje postać:

$$s_i = \frac{1}{2} - \frac{\sum_{j=1}^{m} w_j a_{i0j} b_{0ij} + \sum_{j=1}^{m} \sum_{i''=1}^{n} w_j a_{ii''j} b_{0i''j}}{2\left[\left(\sum_{j=1}^{m} \sum_{i''=1}^{n} w_j a_{ii''j}^2\right) \cdot \left(\sum_{j=1}^{m} \sum_{i''=1}^{n} w_j b_{0i''j}^2\right)\right]^{\frac{1}{2}}} \tag{3.1}$$

gdzie:

- a_{ii^*j}-miernik odległości i-tego obiektu o j-tej zmiennej, od i^*-tego obiektu, gdzie $i^* = 0, i^{''}$, przy czym $i^{''} \neq 0$
- b_{0i^*j}-miernik odległości obiektu wzorcowego o j-tej zmiennej, od i^*-tego obiektu, gdzie $i^* = i, i^{''}$, przy czym $i^{''} \neq 0$
- w_j-waga j tej zmiennej, dla której spełnione są warunki:

$$w_j \in [0, m] \quad \wedge \quad \sum_{j=1}^{m} w_j = m$$

Ostateczna postać zmiennej syntetycznej zależy od skali pomiaru zmiennych. Jeśli zmienne charakteryzujące obiekty mierzone są na skali ilorazowej lub przedziałowej, stosowane jest następujące podstawienie:

$a_{ii^*j} = z_{ij} - z_{i^*j}$ dla $i^* = 0, i^{''}$,

$b_{0i^*j} = z_{0j} - z_{i^*j}$ dla $i^* = i, i^{''}$.

gdzie:

z_{0j}-wystandaryzowana wartość j-tej zmiennej dla obiektu wzorcowego

Z kolei, gdy zmienne charakteryzujące obiekty mierzone są na skali porządkowej to stosowne jest podstawienie:

$$a_{ii^*j} = \begin{cases} 1 & \text{dla } z_{ij} > z_{i^*j}, \\ 0 & \text{dla } z_{ij} = z_{i^*j}, \quad i^* = 0, i^{'}, \\ -1 & \text{dla } z_{ij} < z_{i^*j}, \end{cases} \tag{3.2}$$

$$b_{0i^*j} = \begin{cases} 1 & \text{dla } z_{0j} > z_{i^*j}, \quad i^* = i, i^{'}, \\ 0 & \text{dla } z_{0j} = z_{i^*j}, \quad i^* = i, i^{'}, \\ -1 & \text{dla } z_{0j} < z_{i^*j}. \end{cases} \tag{3.3}$$

Zmienna syntetyczna przyjmuje wartości z przedziału $[0, 1]$. Czym niższa wartość zmiennej syntetycznej, tym bliżej wzorca leży dany obiekt.

Metoda dystansowa

W metodzie tej zmienna syntetyczna wyznaczana jest na podstawie odległości każdego obiektu od obiektu wzorca, przy wykorzystaniu np. metryki euklidesowej. Miara syntetyczna, przy wykorzystaniu przekształcenia unitaryzacyjnego, jest postaci:

$$s_i = \left(\frac{d_{i0} - \min_i\{d_{i0}\}}{\max_i\{d_{i0}\} - \min_i\{d_{i0}\}}\right)^p, \quad i = 1, 2, \ldots, m \tag{3.4}$$

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

3.1.3 Metody iteracyjne

W metodach tych przyjmowania jest funkcja kryterium dobroci porządkowania i w kolejnych iteracjach poszukiwane jest takie uporządkowanie liniowe obiektów, które optymalizuje wartość funkcji kryterium, aż do osiągnięcia przez nią wartości optymalnej tj. maksymalnej lub minimalnej.

Metoda Szczotki

W metodzie tej poszukiwane jest takie liniowe uporządkowanie obiektów, dla którego funkcja kryterium dobroci uporządkowania osiąga maksimum:

$$F^2 = \sum_{i^{'}=1}^{n-1} i^{'} \sum_{i=1}^{n-i^{'}} d_{ii'} \to \max \tag{3.5}$$

gdzie:

$d_{ii'}$ - odległość euklidesowa miedzy i-tym i $i^{'}$-tym obiektem.

Sposób postępowania:

W pierwszym kroku dokonywane jest dowolne liniowe uporządkowanie obiektów, dla którego obliczana jest wartość funkcji kryterium (3.5). W kolejnym etapie obliczana jest wartość funkcji kryterium dla każdej możliwej transpozycji pary obiektów. Uporządkowanie to stanowi punkt wyjścia do oceny, czy kolejna transpozycja dowolnej pary obiektów pozwoli na wzrost wartości funkcji kryterium. Powyższe postępowanie kontynuowane jest tak długo, aż transpozycja dowolnej pary obiektów nie prowadzi do wzrostu wartości funkcji kryterium.

3.1.4 Metody gradientowe

W metodach gradientowych dąży się do takiego liniowego uporządkowania obiektów, które jak najmniej zniekształca relacje strukturalne porządkowanego zbioru obiektów. Od strony geometrycznej oznacza to, że odległości pomiędzy punktami reprezentującymi obiekty w przestrzeni jednowymiarowej, określonej przez zmienną syntetyczną, w jak najmniejszym stopniu zniekształcają odległości pomiędzy tymi punktami w przestrzeni wielowymiarowej, określonej przez zmienne wejściowe. Metody gradientowe poszukują takich współrzędnych punktów reprezentujących obiekty w przestrzeni jednowymiarowej, dla których funkcja dobroci uporządkowania osiąga minimum, co można przedstawić za pomocą wariantów:

$$F^3 = \frac{\sum_{i,i^{'}=1, i\neq i^{'}}^{n} (d^s_{ii'} - d_{ii'})^2}{\sum_{i,i^{'}=1 i<i^{'}}^{n} d_{ii'}} \to \min \tag{3.6}$$

$$F^4 = \sum_{i,i^{'}=1, i<i^{'}}^{n} \left(\frac{d^s_{ii'} - d_{ii'}}{d_{ii'}} \right)^2 \to \min \tag{3.7}$$

$$F^5 = \frac{1}{\sum_{i,i^{'}=1, i\neq i^{'}}^{n} d_{ii'}} \sum_{i,i^{'}=1 i<i^{'}}^{n} \frac{\left(d^s_{ii'} - d_{ii'} \right)^2}{d_{ii'}} \to \min \tag{3.8}$$

gdzie:

$d_{ii'}$ - odległość euklidesowa miedzy i-tym i $i^{'}$-tym obiektem.

3.2 Metody porządkowania nieliniowego

Metody porządkowania nieliniowego nie pozwalają na ustaleniu hierarchii obiektów, lecz na określeniu dla każdego z nich, stopnia podobieństwa do innych obiektów, ze względu na ich charakterystyki.

Aby uporządkować nieliniowo obiekty, charakteryzujące je zmienne powinny być mierzone przynajmniej na skali przedziałowej lub ilorazowej. Gdy zmienne te mierzone są na skali przedziałowej lub ilorazowej, należy dokonać ich normalizacji, dla zapewnienia ich porównywalności.

Metody porządkowania nieliniowego można podzielić na metody dendrytowe i metody aglomeracyjne. Metody dendrytowe prowadzą do powstania dendrytu, będącego ilustracją graficzną położenia względem siebie obiektów ze względu na ich podobieństwo. Z kolei metody aglomeracyjne prowadzą do utworzenia drzewka połączeń, będącego graficzną ilustracją hierarchii łączenia obiektów, ze względu na ich podobieństwo.

3.2.1 Metody dendrytowe

Metody dendrytowe opierają się na regułach i pojęciach teorii grafów. Porządkowanie dendrytowe polega na przyporządkowaniu obiektom poszczególnych wierzchołków dendrytu, w tym celu budowany jest dendryt. Poniżej zostaną opisane przykłady metod dendrytowych, tj. taksonomia wrocławska oraz metoda Prima.

Taksonomia wrocławska

W metodzie tej obiekty dzielone są grupy obiektów najbardziej do siebie podobnych, tj. takich, których odległość między sobą jest najmniejsza. W tym celu w każdym wierszu(kolumnie) macierzy odległości D, wyznaczamy jest element najmniejszy:

$$d_i i^{'} = \max_{i^{'}} d_{ii'}, \quad i, i^{'} = 1, 2, \ldots, n, i \neq i^{'}. \tag{3.9}$$

Otrzymane pary najbardziej podobnych do siebie obiektów, przedstawiane są w postaci grafu niezorientowanego, Długość krawędzi łączących wierzchołki grafu, są proporcjonalne do odległości między obiektami. Może się zdarzyć, że wśród wyznaczonych par połączeń, pojawią się połączenia występujące dwukrotnie, jedno z nich zostanie wyeliminowane, ponieważ kolejność połączeń w dendrycie nie jest istotne. Warto również zwrócić uwagę, na fakt iż w dendrycie danych obiekt może występować tylko jeden raz, w związku z tym jeżeli w łączeniu występują wielokrotnie te same obiekty, to zostaną one połączone w zespoły zwane skupieniami. Metoda kończy działanie, w momencie uzyskania grafu spójnego.

Metoda Prima

W odróżnieniu od taksonomii wrocławskiej, metoda Prima nie wymaga operowania cały czas pełną, wyjściową macierzą odległości. W trakcie tworzenia dendrytu, na każdym etapie zbiór porządkowanych obiektów jest klasyfikowany do jednego z dwóch podzbiorów, np. A i B. Niech zbiór A będzie pierwszym z nich a zbiór B drugim. Pierwszy z nich zawiera obiekty należące na danym etapie do dendrytu, zaś drugi zawiera obiekty nie należące na tym etapie do dendrytu.

Na początku procedury, zbiór A jest zbiorem pustym, z kolei zbiór B zawiera wszystkie obiekty. W pierwszym kroku do zbioru A zostaje włączony dowolny obiekt, nie ma to wpływu na ostateczną postać dendrytu. Następnie do zbioru A zostają włączone te obiekty zbioru B, najbardziej podobne do obiektów należących już do zbioru A. W tym celu w pierwszym kroku algorytmu zostaje stworzony wektor d, zawierający odległości wybranego obiektu zbioru A, od pozostałych obiektów zbioru B.

W powstałym dendrycie wierzchołkami są obiekty przechodzące kolejno do zbioru A, z kolei wiązadłami łączącymi wierzchołki są minimalne wartości elementów wektora d, otrzymanego w kolejnych krokach przyłączania obiektów do dendrytu.

3.2.2 Metody aglomeracyjne

Istotą metod aglomeracyjnych jest utworzenie drzewka połączeń - dendrogramu. W ten sposób zobrazowana jest hierarchia łączenia obiektów, na podstawie zmniejszającego się podobieństwa między obiektami włączonymi do dendrogramu, w kolejnych etapach a obiektami należącymi już do dendrogramu. Hierarchia połączeń określa wzajemnie położenie względem siebie obiektów oraz grup obiektów powstających w kolejnych etapach tworzenia drzewka. Grupy podobnych do siebie obiektów tworzą oddzielne gałęzie. Punktem wyjściem metod aglomeracyjnych stanowi założenie, że każdy obiekt stanowi odrębną, jednoelementową grupę ($\mathrm{G_r}, r = 1, 2, \ldots, z$).

W kolejnych krokach następuje łączenie ze sobą grupy obiektów najbardziej do siebie podobnych ze względu na wartości opisujących je zmiennych. Podobieństwo weryfikowane jest na podstawie odległości między grupami obiektów.

Na początku odległości między jednoelementowymi grupami obiektów $\mathrm{G_1}, \ldots, \mathrm{G_z}$ są elementami wyjściowej macierzy odległości **D**. W macierzy **D** poszukiwane są najmniejsze odległości pomiędzy grupami obiektów:

$$d_{rr'} = \min_{ii'} d_{ii'}, \quad i = 1, 2, \ldots, n_r, \quad i^{'} = 1, 2, \ldots, n_{r'}, \quad r, r^{'} = 1, 2, \ldots, z, r \neq r^{'}. \tag{3.10}$$

gdzie:

$d_{rr'}$ - odległość r-tej od $r^{'}$-tej grupy.

Ogólny wzór wyznaczania odległości nowo powstałej grupy $\mathrm{G_{r''}}$, powstałej w wyniku połączenia grup $\mathrm{G_r}$ i $\mathrm{G_{r'}}$, od pozostałych grup G_r''', przy tworzeniu drzewka połączeń ma postać:

$$d_{r'''r''} = \alpha_r d_{r'''r} + \alpha_{r'} d_{r'''r'} + \beta d_{rr'} + \gamma |d_{r'''r} - d_{r'''r'}| \tag{3.11}$$

gdzie: $\alpha_r, \alpha_{r'}, \beta, \gamma$ - współczynniki przekształceń odmienne dla różnych metod aglomeracyjnych

Poszczególne metody aglomeracyjne, różnią się między sobą sposobami wyznaczania odległości między obiektami. Poniżej zostały wymienione najczęściej stosowane metody aglomeracyjne, które będą dokładniej omówione w kolejnym podrozdziale.

- metoda najbliższego sąsiedztwa (metoda pojedynczego wiązania)
  parametry przekształceń $\alpha_r = 0,5 \quad \alpha_{r'} = 0,5 \quad \beta = 0 \quad \gamma = 0,5.$
- metoda najdalszego sąsiedztwa (metoda pełnego wiązania)
  parametry przekształceń $\alpha_r = 0,5 \quad \alpha_{r'} = 0,5 \quad \beta = 0 \quad \gamma = -0,5.$
- metoda średniej międzygrupowej (metoda średnich połączeń)
  parametry przekształceń $\alpha_r = \frac{n_r}{n_r + n_{r'}} \quad \alpha_{r'} = \frac{n_{r'}}{n_r + n_{r'}} \quad \beta = 0 \quad \gamma = 0).$
- metoda mediany
  parametry przekształceń $\alpha_r = 0,5 \quad \alpha_{r'} = 0,5 \quad \beta = -025 \quad \gamma = 0.$
- metoda środka ciężkości
  parametry przekształceń $\alpha_r = \frac{n_r}{n_r + n_{r'}}; \alpha_{r'} = \frac{n_{r'}}{n_r + n_{r'}} \quad \beta = \frac{-n_r n_{r'}}{(n_r + n_{r'})^2} \quad \gamma = 0.$
- metoda Warda
  parametry przekształceń $\alpha_r = \frac{n_r + n_{r'''}}{n_r + n_{r'} + n_{r'''}} \quad \alpha_{r'} = \frac{n_{r'} + n_{r'''}}{n_r + n_{r'} + n_{r'''}} \quad \beta = \frac{-n_{r'''}}{n_r + n_{r'} + n_{r'''}} \quad \gamma = 0.$

Metoda najbliższego sąsiedztwa

W metodzie tej odległość między dwoma grupami obiektów jest równa odległości pomiędzy najbliższymi obiektami (sąsiadami), które należą do dwóch różnych grup obiektów. Odległość ta opisana jest wzorem:

$$d_{rr'} = \min_{ii'} d_{ii'}(\mathbf{O_i} \in \mathbf{G_r}, \mathbf{O_{i'}} \in \mathbf{G_{r'}}), \tag{3.12}$$

$$i = 1, 2, \ldots, n_r, \quad i' = 1, 2, \ldots, n_{r'}, \quad r, r' = 1, 2, \ldots, z, \quad r \neq r',$$

gdzie:

$$\mathbf{O_i} = [z_{ij}], \quad j = 1, 2, \ldots, m. \tag{3.13}$$

Metoda najdalszego sąsiedztwa

W metodzie tej odległość między dwoma grupami obiektów jest równa odległości pomiędzy najdalszymi obiektami (sąsiadami), które należą do dwóch różnych grup obiektów. Odległość ta opisana jest wzorem:

$$d_{rr'} = \max_{ii'} d_{ii'}(\mathbf{O_i} \in \mathbf{G_r}, \mathbf{O_{i'}} \in \mathbf{G_{r'}}), \tag{3.14}$$

$$i = 1, 2, \ldots, n_r, \quad i' = 1, 2, \ldots, n_{r'}, \quad r, r' = 1, 2, \ldots, z, \quad r \neq r',$$

Metoda średniej międzygrupowej

W metodzie tej odległość między dwoma grupami obiektów równa jest średniej arytmetycznej odległości między wszystkimi parami obiektów należących do dwóch różnych wzór. Odległość ta opisana jest wzorem:

$$d_{rr'} = \frac{1}{n_r n_{r'}} \sum_{i'=1}^{n_{r'}} \sum_{i=1}^{n_r} d_{ii'}(\mathbf{O_i} \in \mathbf{G_r}, \mathbf{O_{i'}} \in \mathbf{G_{r'}}) \tag{3.15}$$

$$r, r' = 1, 2, \ldots, z, \quad r \neq r'.$$

Metoda mediany

W metodzie tej odległość między grupami obiektów jest równa medianie odległości pomiędzy wszystkimi parami obiektów należących do dwóch grup. Odległość ta opisana jest wzorem:

$$d_{rr'} = \mathbf{med}_{i,i'}\{d_{ii'}(\mathbf{O_i} \in \mathbf{G_r}, \mathbf{O_{i'}} \in \mathbf{G_{r'}})\}, \tag{3.16}$$

$$i = 1, 2, \ldots, n_r, \quad i' = 1, 2, \ldots, n_{r'}, \quad r, r' = 1, 2, \ldots, z, \quad r \neq r'.$$

Metoda środków ciężkości

W metodzie tej odległość między dwoma grupami jest równa odległości między środkami ciężkości tych grup. Odległość ta opisana jest wzorem:

$$d_{rr'} = d_{i^c i'^c}(\mathbf{O_i^c} = \overline{\mathbf{O}}_\mathbf{r} \in \mathbf{G_r}, \mathbf{O_{i'^c}} = \overline{\mathbf{O}}_\mathbf{r} \in \mathbf{G_{r'}}), \tag{3.17}$$

$$i = 1, 2, \ldots, n_r, \quad i' = 1, 2, \ldots, n_{r'}, \quad r, r' = 1, 2, \ldots, z, \quad r \neq r'.$$

gdzie:

$d_{i^c i'^c}$ - odległość środka ciężkości r-tej grupy od środka ciężkości $r\prime$-tej grupy,

$\overline{\mathbf{O}}_{\mathbf{i^c}}, \overline{\mathbf{O}}_{\mathbf{i'^c}}$ - środki ciężkości odpowiednio r-tej i $r^{'}$-tej grupy obiektów. przy czym:

$$\mathbf{O}_{i^c} = \overline{\mathbf{O}}_r = \frac{1}{n_r}\sum_{\mathbf{i=1}}^{\mathbf{n_r}}\mathbf{O}_i, \tag{3.18}$$

$$\mathbf{O}_{i'^c} = \overline{\mathbf{O}}_{r'} = \frac{1}{n_{r'}}\sum_{\mathbf{i'=1}}^{\mathbf{n_{r'}}}\mathbf{O}_{i'}. \tag{3.19}$$

Metoda Warda

W metodzie tej odległości między dwoma grupami obiektów nie można przedstawić wprost za pomocą odległości między obiektami należącymi do tych grup. Dwie grupy obiektów podczas tworzenia drzewka połączeń, na dowolnym etapie są łączone w jedną grupę,w celu zminimalizowania sumy kwadratów odchyleń wszystkich obiektów z tych dwóch grup od środka ciężkości nowej grupy, powstałej w wyniku połączenia tych dwóch grup. Proces ten oznacza, że na każdym etapie łączenia grup obiektów, w jedną grupę łączy się te grupy, które charakteryzują się najmniejszym zróżnicowaniem ze względu na opisujące je zmienne. Zróżnicowanie badania się przy pomocy kryterium $ESS(ErrosSumofSquares)$ sformułowanego przez J.H. Warda, które jest postaci:

$$ESS = \sum_{\mathbf{i''=1}}^{\mathbf{n_r''}} d^2_{i'' i''^c}(\mathbf{O}_{i''} \in \mathbf{G}_{r''}, \mathbf{O}_{i''^c} = \overline{\mathbf{O}}_{r''} \in \mathbf{G}_{r''}), \tag{3.20}$$

gdzie: $d_{i'' i''^c}$ - odległość $i^{''}$-tego obiektu, należącego do nowo powstałej $r^{''}$-tej grupy od środka ciężkości tej grupy,

$$\mathbf{O}_{i''^c} = \overline{\mathbf{O}}_{r''} = \frac{1}{n_{r''}}\sum_{\mathbf{i''=1}}^{\mathbf{n_r''}}\mathbf{O}_{i''}. \tag{3.21}$$

Rozdział 4

Zastosowanie wybranych metod porządkowania danych wielowymiarowych

4.1 Opis zbioru

Zbiór danych jest opracowaniem własnym, na podstawie ofert sprzedaży samochodów osobowych, zamieszczonych na portalu *www.otomoto.pl* w okresie listopad - grudzień 2017 roku. Zebrane dane dotyczą szczegółowych informacji odnośnie 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. Oprócz 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 oraz bezwypadkowe,czy jest sprowadzane, jaki jest kraj aktualnej rejestracji, czy było serwisowane, czy sprzedający jest pierwszym właścicielem. Dodatkowo oprócz powyższych, został dodany atrybut najbardziej interesujący kupującego - czyli cena oraz województwo tj. miejsce skąd wystawiana została oferta.

	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, opisujących 61 różnych rekordów, tj. obiektów reprezentowanych przez wiersze, którym przypisano pewne wartości atrybutów. Wśród zebranych danych można wyróżnić zarówno zmienne jakościowe, jak i ilościowe.

Zmiennymi jakościowymi są atrybuty:

- 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

Wśród zmiennych jakościowych można wyróżnić zmienne porządkowe, nominalne oraz binarne. W stworzonym zbiorze danych zmiennymi binarnymi są atrybuty:

- pierwszy.wlasciciel,
- ABS,
- serwisowany,
- komputer.pokladowy,
- ESP,
- bezwypadkowy,
- uszkodzony.

Pozostałe atrybuty są zmiennymi nominalnymi.

Zmiennymi ilościowymi są atrybuty:

- cena.netto[pln],
- cena.brutto[pln],
- moc,
- pojemnosc.skokowa[cm^3],
- rok.produkcji,
- przebieg[km],
- l.drzwi.

Wśród zmiennych ilościowych można wyróżnić zmienne skokowe oraz dyskretne. W stworzonym zbiorze danych, zmiennymi skokowymi są:

- moc,
- pojemnosc.skokowa[cm^3],
- rok.produkcji,
- przebieg,
- l.drzwi.

Z kolei atrybuty: cena.netto[pln], cena.brutto[pln] są zmiennymi ciągłymi.

Bibliografia

[1] Grzegorz Banaszak and 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 and Sławomir Turek. *Teoria mnogości*. Państwowe Wydawnictwo Naukowe, Warszawa, 2007.

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

[5] Jerzy Greń. *Statystyka matematyczyna: modele i zadania*. Państwowe Wydawnictwo Naukowe, Warszawa, 1984.

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

[7] W. Krysicki, J. Bartos, W. Dyczka, K. Królikowska, and M. Wasilewski. *Rachunek prawdopodobieństwa i statystyka matematyczna w zadaniach: część I. rachunek pradopodobieństwa*. Państwowe Wydawnictwo Naukowe, Warszawa, 1999.

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

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

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

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

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