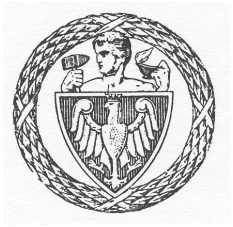


POLITECHNIKA WARSZAWSKA

WYDZIAŁ MATEMATYKI
I NAUK INFORMACYJNYCH

PRACA DYPLOMOWA MAGISTERSKA

Analiza gradacyjna – teoria i zastosowania

Grade Correspondence Analysis - theory and applications

Autor:
Elżbieta Anna Kłopotek

Promotor
dr inż. Przemysław Biecek

WARSZAWA, CZERWIEC 2012

.....
podpis promotora

.....
podpis autora

Streszczenie

Gradacyjna analiza danych (GAD), zaproponowana i rozwijana w Instytucie Podstaw Informatyki PAN, jest metodą wizualizacji pewnych podstawowych konceptów dotyczących analizy par zmiennych w próbkach niezależnych oraz zależnych.

Niniejsza praca stanowi próbę syntetycznej ekspozycji GAD, wskazania podobieństw i różnic w stosunku do pojęć znanych z użycia m.in. klasycznej statystyki oraz wskazania obszarów zastosowania GAD, czemu służy m.in. przeprowadzone w ramach pracy studium danych nt. ocen egzaminacyjnych studentów.

Wskazano na użyteczność wizualizacji zależności ocen z poszczególnych przedmiotów za pomocą krzywych koncentracji (uogólnionych krzywych Lorenza), stopniowanych krzywych koncentracji, wykresów współczynnika polowego (jako substytutu indeksu Giniego) oraz indeksu Giniego-Simpsona (uogólnienie GAD klasycznego indeksu Giniego-Simpsona). Dla celów wspomnianych badań przygotowano zestaw programów w języku R, który stał się podstawą ilustracji w niniejszym tekście.

Skupiono się na tych jej składowych, które są istotne z punktu widzenia danych z systemu USOS, będących przedmiotem prac eksperymentalnych, tzn. analizie par zmiennych o takich samych zbiorach wartości w niezależnych próbkach oraz analizie par zmiennych dla jednej próbki. Ograniczono się do zmiennych porządkowych dyskretnych.

W pracy przestawiono definicje pojęć GAD, ich interpretację oraz wybrane własności, opierając się na literaturze przedmiotu, uzupełniając je własnymi przemyśleniami.

W szczególności w niniejszej pracy określono zakres indeks Giniego-Simpsona GAD, zbadano wpływ przekodowania dziedzin zmiennych na kształt krzywej koncentracji, określono wartość minimalną i maksymalną funkcji nadreprezentacji, a dla niezależnych zmiennych jej wartość oczekiwaną. Udowodnione twierdzenia mają znaczenie dla implementacji programów wizualizacyjnych, a także celowości transformacji dziedzin zmiennych.

Ponadto drobnym wkładem jest uporządkowanie definicji krzywej koncentracji oraz powierzchni koncentracji.

Podstawowym walorem GAD jest nacisk na wizualną interpretację pojęć. Na podstawie wizualizacji krzywych koncentracji oraz map nadreprezentacji dość łatwo można dostrzec jakie przedmioty łatwiej zaliczyć, czy też który z prowadzących ten sam przedmiot jest surowszy w ocenianiu studentów, jak uczą się poszczególnych przedmiotów studentki w porównaniu ze studentami, jaka jest łatwość zaliczania w kolejnych terminach egzaminów, na ile dobre oceny z jednych przedmiotów zależą od innych, czy uprawianie sportu wpływa na uzyskiwane wyniki nauczania.

Podsumowując wydaje się, że w fazie eksploracji danych gradacyjna analiza danych może ułatwić dostrzeżenie pewnych tendencji poprzez dotarcie do obrazowej wyobraźni analityka i przez to stymulować jego prace badawcze.

Abstract

Grade Correspondence Analysis (GCA, in Polish Gradacyjna analiza danych GAD), proposed and developed at the Institute of Computer Science, is a method of visualizing certain basic concepts for the analysis of pairs of variables in independent and dependent samples.

This thesis is an attempt to expose synthetically GCA, identify similarities and differences in relation to concepts known from usage in classic statistics and to identify areas of GCA application. In particular, the methodology is applied to examination grades of students of one Polish university.

The usefulness of visualization of dependencies between grades of individual subjects using concentration curves (generalized Lorenz curves), graded concentration curves, area factor graphs (as a substitute for the Gini index) and the Gini-Simpson index (the GCA generalization of the classical Gini-Simpson index) was demonstrated. For the purposes of these studies there was a set of programs in R, which became the basis of the illustrations in the text.

We focus on the components, which are significant from a point of view of the data from the USOS system that was used for experimental work, i.e. analysis of pairs of variables with the same set of values from independent samples and analysis of pairs of variables of a single sample under restriction to discrete ordinal variables.

In the thesis basic GCA definitions, their interpretations and selected properties are presented, based on the bibliography while complementing them with own thoughts.

In particular, range of Gini-Simpson index is investigated, the influence of recoding variables domains on the shape of the curve of concentration is tested, minimum and maximum values for overrepresentation functions and expected values of overrepresentation function for independent variables are identified. Selected theorems are proven. They are important for implementation of visualization programs and provide basis for making decisions on transformation of variable domains.

In addition, a small contribution is a clarification of the definition of the concentration curve and surface concentration.

The main advantage of GCA is emphasis on the visual interpretation of various concepts. On the basis of visualization of the concentration curves and overrepresentation maps it is easy to see, what exams are easier to pass, or which teacher engaging in the same subject is stricter in assessing students, how are female and male students learning, how easy is it to pass in the subsequent dates of examinations, how good is the evaluation of some subjects depending on others, whether sport affects the learning achievements etc..

In conclusion, it appears that GCA in the phase of data mining , can make it easier to see certain trends through pictorial visualization for the analyst and help him to do his job.

Podziękowania

Przede wszystkim pragnę gorąco podziękować memu Promotorowi, Panu Doktorowi Przemysławowi Bieckowi, którego niezwykła życzliwość, doświadczenie oraz pomoc pomogły mi przygotować przedłożoną pracę.

Spis treści

Rozdział 1

Wprowadzenie

1.1 Motywacja

Eksploracyjna analiza danych to dział statystyki, do rozwoju którego przyczyniły się w dużym stopniu prace Tukeya [Tuk77]. Podczas gdy klasyczna statystyka koncentrowała się na testowaniu postawionych hipotez, eksploracyjna analiza danych zmierza do sformułowania hipotez nt. przyczyn obserwowanych zjawisk, bada poprawność (spełnianie) założeń dla wnioskowania statystycznego, wspiera dobór właściwych technik i narzędzi statystycznych do badań hipotez oraz formułuje zadania zbierania dalszych danych poprzez obserwacje i/lub doświadczenia. Tukey zachęcał statystyków m.in. do korzystania z metod wizualizacji danych, nim zaczną formułować hipotezy, badane na nowych kolekcjach danych [FM00].

Do metod eksploracyjnej analizy danych zalicza się między innymi zapoczątkowaną w Instytucie Podstaw Informatyki Polskiej Akademii Nauk *gradacyjną analizę danych* (GAD) [KPR04]. Charakteryzuje się ona tym, że:

- Bazuje na tzw. przekształceniach gradacyjnych i korzysta z pomiaru koncentracji[1] do analizy odpowiedniości[2]

  Metoda ta jest stosowana szczególnie często w naukach biologicznych oraz socjologicznych, gdzie często występują macierze kontyngencji., analizy skupień, analizy regresji, oceny regularności danych i rozwiązywania innych zagadnień analizy danych.

- Przekształcenie gradacyjne zamienia parę zmiennych tak, aby ich rozkłady brzegowe były (prawie) równomierne.

- Podstawowymi narzędziami GAD są tzw. miary nierówności, krzywe koncentracji oraz koncept regularności rozkładów dwuwymiarowych.

- Można ją określić jako metodę analizy bezmodelowej (nieparametrycznej), gdyż pomija ona założenia o rozkładzie analizowanych zmiennych.

- Dopuszcza zmienne mierzone w różnych skalach: ilorazowej, porządkowej jak i nominalnej.

[1] Do podstawowych narzędzi GAD należą tzw. krzywe koncentracji, opisane w definicji 3.1.1 na stronie 19. Reprezentują one zagęszczenie (stosunek prawdopodobieństw wartości) jednej zmiennej losowej w stosunku do drugiej przy założeniu, że obie mają takie same dziedziny.

[2] Analiza odpowiedniości (korespondencji) w eksploracyjnej analizie macierzy kontengencji usiłuje zidentyfikować ciągłe zmienne ukryte, których wyrazem są kategoryczne zmienne wiersza i kolumny. Przypisywane wartości ciągłe winny maksymalizować współczynnik korelacji Pearsona pomiędzy zmiennymi ukrytymi.

- Zakłada reprezentację danych z próbki w postaci dwuwymiarowej tablicy kontyngencji kolumn i wierszy, przy czym zakłada się, że istnieją jakieś ukryte czynniki (zmienne ukryte, nie obserwowane) porządkujące wiersze / kolumny w taki sposób, że wyraźna staje się zależność w danych, w szczególności zależność wierszowej i kolumnowej porządkującej zmiennej ukrytej.
- Podobnie jak klasyczna analiza korespondencji oblicza znormalizowane czynniki ukryte wierszowy i kolumnowy w taki sposób, by uzyskać maksymalną pozytywną zależność.
- Podczas gdy klasyczna analiza korespondencji optymalizowała wartość współczynnika korelacji Pearsona jako miary zależności, GAD optymalizuje współczynnik ρ Spearmana. Dlatego zadowala się skalami porządkowymi. Normalizacja w GAD jest także odmienna od klasycznej analizy korespondencji.

GAD znalazła liczne zastosowania w praktycznej analizie danych, m.in. w np. danych medycznych [KMPW05] czy sytuacji ekonomicznej firm [Cio05].

1.2 Cel pracy

Celem pracy było zbadanie własności algorytmów stosowanych w analizie gradacyjnej pod kątem możliwości zastosowania w eksploracyjnej analizie danych. W części teoretycznej należało przedstawić przegląd metod gradacyjnej analizy danych oraz postawić i udowodnić wybrane twierdzenia dotyczące zgodności tych metod. W części praktycznej należało zaimplementować algorytmy realizujące przedstawione metody gradacyjnej analizy danych w języku R, jak również przedstawić wyniki zastosowania wymienionych algorytmów do analizy zbioru danych z systemu USOS.

1.3 Układ pracy

W rozdziale 2 przedstawiono podstawowe koncepcje GAD w zakresie analizy jednej zmiennej.

Rozdział 3 prezentuje metody wizualizacji dla par zmiennych pochodzących z niezależnych próbek, w szczególności omawiana jest krzywa koncentracji oraz jej odniesienie do klasycznej krzywej Lorenza.

W rozdziale 4 omawiana jest tzw. mapa nadreprezentacji, będąca proponowanym przez GAD sposobem wizualizacji zależności pary zmiennych.

W rozdziałach 2-4 przedstawiono wybrane twierdzenia, dotyczące własności prezentowanych koncepcji, oraz ich dowody.

Rozdział 5 poświęcony jest zastosowaniu GAD do analizy danych nt. ocen egzaminacyjnych wybranej grupy studentów z systemu USOS.

Pracę podsumowuje rozdział 6.

Rozdział 2

Analiza gradacyjna jednej zmiennej

2.1 Bazowe pojęcia analizy gradacyjnej - ujęcie nieformalne

Metody gradacyjne mierzą niepodobieństwo między rozkładami. Dzięki nim można wykrywać trendy w danych, podpopulacje związane ze strukturami funkcyjnymi, obserwacje odstające, tworzyć naturalne grupy obserwacji (klastry). Opierają się na pojęciu koncentracji. Definiują jeden rozkład w odniesieniu do drugiego za pomocą tzw. krzywej koncentracji.

Krzywa koncentracji to krzywa opisująca stopień koncentracji rozkładu zmiennej losowej względem rozkładu innej zmiennej losowej (tzw. rozkład bazowy).

Wartość nadreprezentacji jest ilorazem względnej proporcji dla rozkładu i względnej proporcji rozkładu bazowego dla każdej kategorii rozkładu.

2.2 Podstawowe definicje

Niech $(\mathbb{R}, \tau)$ będzie przestrzenią topologiczną. Zbiór borelowski definiuje się jako element należący do najmniejszego σ-ciała przestrzeni $\mathbb{R}$, które zawiera wszystkie podzbiory otwarte $\mathbb{R}$.

Definicja 2.2.1 Zmienną losową *na przestrzeni probabilistycznej* $(\Omega, \mathcal{F}, \mathbb{P})$ *nazywamy dowolną rzeczywistą funkcję mierzalną* $\nu : \Omega \to \mathbb{R}$, *tzn. funkcję* ν *spełniającą warunek*

$$\nu^{-1}(B) \in \mathcal{F}$$

dla każdego zbioru borelowskiego $B \subseteq \mathbb{R}$.

W poniższej ekspozycji gradacyjnej analizy danych bazujemy na książce [KPR04]. Wprowadzane pojęcia w dużej mierze odpowiadają konceptom klasycznej statystyki, ale dla utrzymania spójności z pracami twórców GAD będziemy się w dalszym ciągu posługiwać ich terminologią.

Definicja 2.2.2 *Zmienną losową nazwiemy* zmienną lilipucią *(ozn.* ξ*), jeśli przyjmuje wartości z przedziału [0,1], jest dyskretna, ciągła bądź dyskretna-ciągła. Dystrybuanta oznaczana przez* F_ξ *jest niemalejącą krzywą, leżącą w kwadracie jednostkowym, łączącą punkty (0,0) oraz (1,1).*

W klasycznej statystyce zmiennymi lilipucimi są np. zmienne losowe o rozkładzie *beta*.

Zbiór wszystkich lilipucich zmiennych losowych nazywamy *Jednowymiarowym Modelem Lilipucim* i oznaczamy przez $\mathbb{ULM}$ (z ang. the Univariate Lilliputian Model)

Definicja 2.2.3 *Zmienna lilipucia jest* ciągła, *jeśli jej dystrybuanta jest ciągła i jej pochodna wyrażona wzorem:*

$$f_\xi(u) = \frac{dF_\xi(u)}{du}$$

zwana gęstością zmiennej ξ, istnieje prawie wszędzie w sensie miary Lebesgue'a. Dziedziną u jest przedział [0,1].

Funkcję odwrotną do dystrybuanty nazywać będziemy funkcję kwantylową.

Definicja 2.2.4 Zmienna lilipucia jest dyskretna, *jeśli istnieje ciąg punktów u_i, $u_i \in [0,1]$, $i = 1, 2, ...$ i ciąg prawdopodobieństw $p_\xi(u_i)$ takich, że $\sum_i p_\xi(u_i) = 1$. Dystrybuanta F_ξ spełnia:*

$$p_\xi(u_i) = dF_\xi(u_i) = dF_\xi(u_i+) - dF_\xi(u_i-) > 0$$

Podstawowym wkładem GAD jest wprowadzenie nowych metod wizualizacji podstawowych pojęć statystycznych.

Zacznijmy od wizualizacji dystrybuanty zmiennej lilipuciej. Przykład takiej dystrybuanty widzimy na rysunku 3.1 na stronie 20. Chwilowo zignorujmy fakt, iż osie są opisane zmiennymi - w następnym rozdziale będziemy mówić o zamianie pary zmiennych w zmienną lilipucią. Wtedy będziemy sobie wyobrażać rozkład zmiennej z osi Y w przestrzeni "skrzywionej" przez zmienną osi X, czyli w tym wypadku zmiennej Bazy Danych względem Algebry. Na tej ilustracji jak i na wszystkich innych w niniejszej pracy prezentowane są dane pochodzące z systemu USOS, opisanego szerzej w rozdziale 5. Obecnie spójrzmy na ten wykres jako wykres dystrybuanty zmiennej lilipuciej Bazy-danych-względem-Algebry. Zarówno dziedzina jak i zbiór wartości tej zmiennej to przedziały [0,1]. Sama krzywa dystrybuanty łączy punkt [0,0] z punktem [1,1] na wykresie i jest niemalejąca (jak na dystrybuantę przystało).

Odpowiadającą jej funkcję kwantylową widzimy na rysunku 2.1

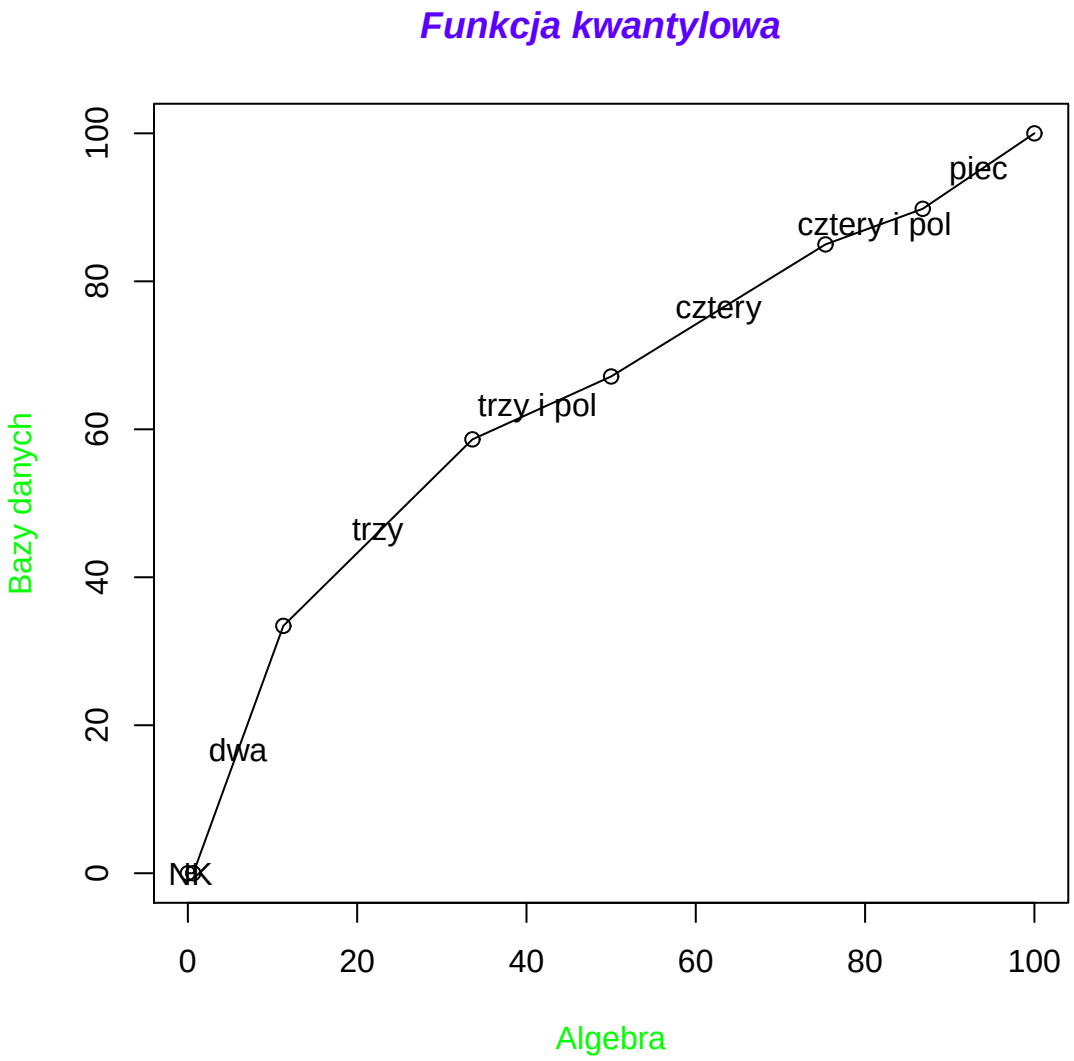
Wartość oczekiwana zmiennej lilipuciej może być przedstawiona jako pewien obszar na kwadracie jednostkowym. Jest ona równa polu nad krzywą lilipucią, odpowiadającą zmiennej ξ, lub polu kwadratu pomniejszonemu o pole pod krzywą lilipucią. Matematycznie zdefiniowana jest jako:

$$E(\xi) = \int_0^1 u dF_\xi(u) = 1 - \int_0^1 F_\xi(u)du = \int_0^1 (1 - F_\xi(1-u))du$$

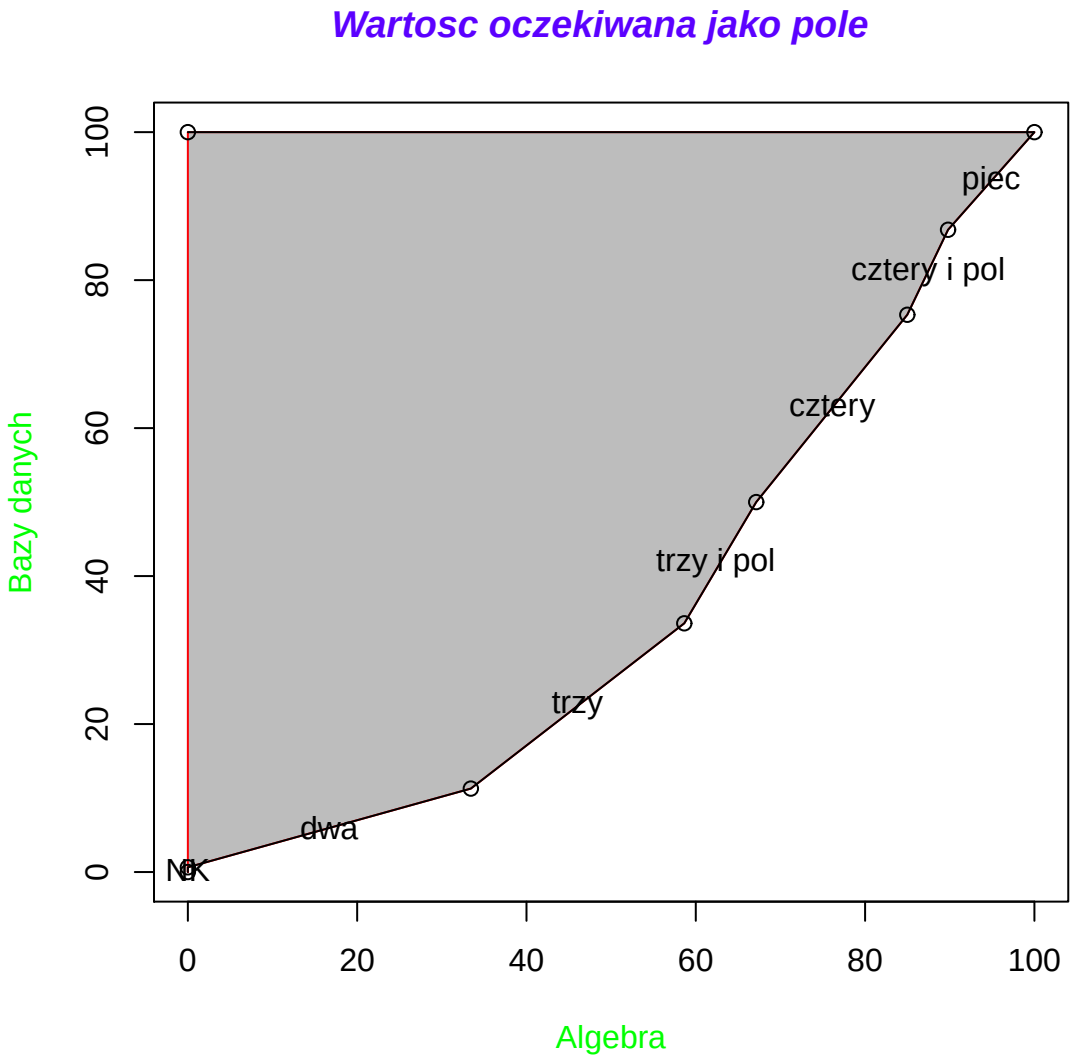
Rysunek 2.2 ilustruje powierzchnię odpowiadającą wartości oczekiwanej wspomnianej zmiennej lilipuciej Bazy-danych-względem-Algebry o dystrybuancie z rys. 3.1.

Definicja 2.2.5 Współczynnik ar (współczynnik polowy, z ang. "area") zmiennej lilipuciej *definiujemy jako:*

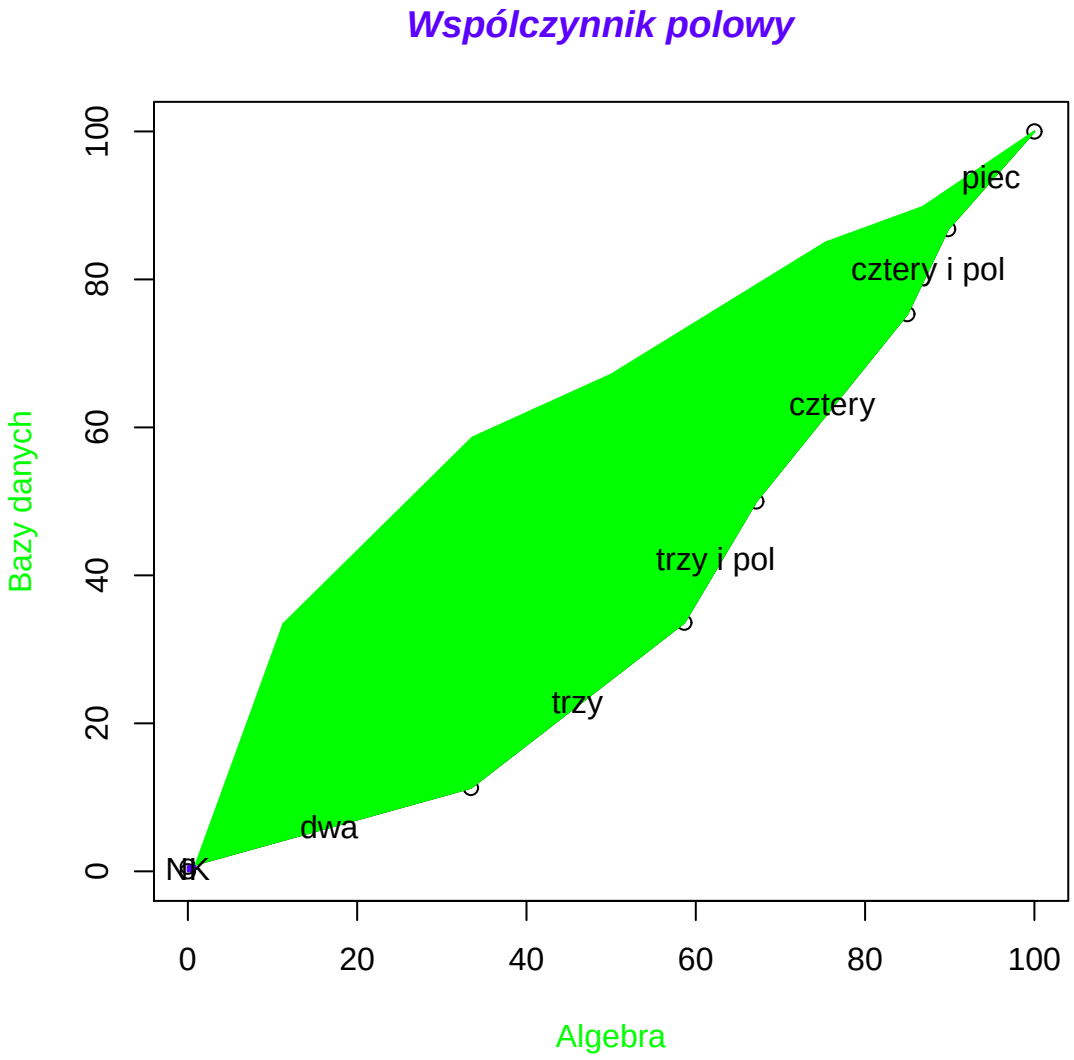
$$ar(\xi) = \frac{E(\xi) - E(U)}{1 - E(U)} = 2E(\xi) - 1 = 2\int_0^1 (u - F_\xi(u))du$$



Rysunek 2.1: Funkcja kwantylowa względem zmiennej Algebra dla zmiennej Bazy danych. Jej interpretacja jest następująca: Krzywa koncentracji mówiła nam, że tak źle (nk lub ndst), jak 35% studentów algebry uczy się tylko ok. 10% studentów baz danych. Funkcja kwantylowa mówi, że tak źle, jak 10% najgorszych studentów z baz danych uczy się ok. 35% studentów algebry.



Rysunek 2.2: Ilustracja wartości oczekiwanej względem zmiennej Algebra dla zmiennej Bazy danych



Rysunek 2.3: Współczynnik polowy względem zmiennej Algebra dla zmiennej Bazy danych

Może być on również wyrażony jako podwojona różnica dwóch obszarów: pola pomiędzy diagonalą i krzywą lilipucią (pod diagonalą) oraz pola pomiędzy diagonalą i krzywą lilipucią (nad diagonalą):

$$ar(\xi) = 2\int_{u>F_\xi(u)} (u - F_\xi(u))du - 2\int_{u<F_\xi(u)} (F_\xi(u) - u)du$$

Przykład wizualizacji tego współczynnika mamy na rysunku 2.3. Jest on zaznaczony dla wspomnianej zmiennej lilipuciej Bazy-danych-względem-Algebry o dystrybuancie z rys. 3.1. Ponieważ dystrybuanta ta leżała poniżej przekątnej, współczynnik polowy jest dodatni, a na rysunku odpowiadające mu pole oznaczono kolorem zielonym. (Gdyby dystrybuanta leżała powyżej przekątnej, współczynnik byłby ujemny, a oznaczające go pole oznaczylibyśmy kolorem niebieskim. Ogólnie wartość współczynnika to różnica między polem zielonym a polem niebieskim).

Taka interpretacja współczynnika polowego zmiennej lilipuciej odpowiada geometrycznej interpretacji indeksu Giniego znanego z klasycznej statystyki.

Indeks $ar(F_\xi(\xi))$ to numeryczna miara nieciągłości dla ξ.

Jedną z miar zmienności jest *indeks Giniego-Simpsona* ozn. $GS(\xi)$ Uśrednia on bezwzględne różnice dla wszystkich par wartości. Formalna definicja jest następująca:

Definicja 2.2.6 *Niech ξ' będzie zmienną niezależną od ξ oraz niech ma taką samą dystrybuantę jak ξ. Wtedy:*

$$GS(\xi) = \frac{E|\xi - \xi'|}{2E(\xi)} \tag{2.1}$$

Geometryczną interpretacją indeksu Giniego-Simpsona jest pole pomiędzy dystrybuantami F_ξ i F_ξ^2 podzielone przez pole nad dystrybuantą F_ξ

Twierdzenie 2.2.1 *$GS(\xi)$ leży w zakresie [0,1].*

Dowód Ponieważ funkcja F_ξ jest dystrybuantą, więc przyjmuje wartości z przedziału [0,1]. Zatem $0 \leq 1 - F_\xi(\xi) \leq 1$. Obustronnie mnożąc przez F_ξ dostajemy $0 \leq F_\xi(\xi) - F_\xi^2(\xi) \leq F_\xi(\xi)$, co po obustronnym całkowaniu od 0 do 1 daje $0 \leq \int_0^1 F_\xi(\xi)d\xi - \int_0^1 F_\xi^2(\xi)d\xi \leq \int_0^1 F_\xi(\xi)d\xi$. Wykluczając ekstremalny przypadek, gdy F_ξ ma całą masę skupioną w punkcie $\xi = 1$, ostatnia całka będzie dodatnia. Dzieląc przez nią obustronnie otrzymujemy

$$0 \leq \frac{\int_0^1 F_\xi(\xi)d\xi - \int_0^1 F_\xi^2(\xi)d\xi}{\int_0^1 F_\xi(\xi)d\xi} \leq 1$$

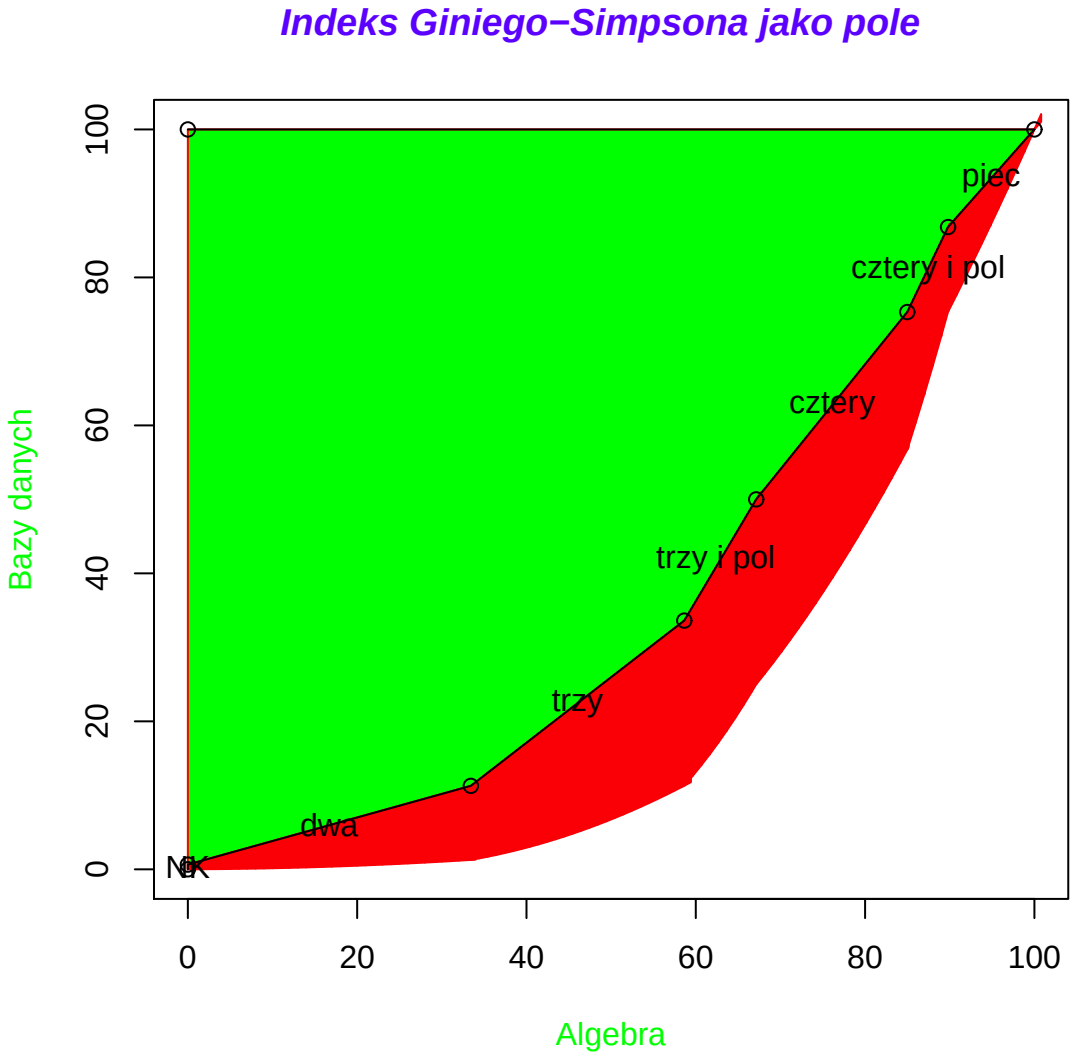
co zgodnie z przytoczoną interpretacją oznacza zamknięcie $GS(\xi)$ w przedziale [0,1], czego należało dowieść.

Rysunek 2.4 ilustruje wspomnianą geometryczną interpretację GS dla wspomnianej zmiennej lilipuciej Bazy-danych-względem-Algebry o dystrybuancie z rys. 3.1. Obszar oznaczony na czerwono to licznik, a obszar zielony to mianownik powyższego ułamka.

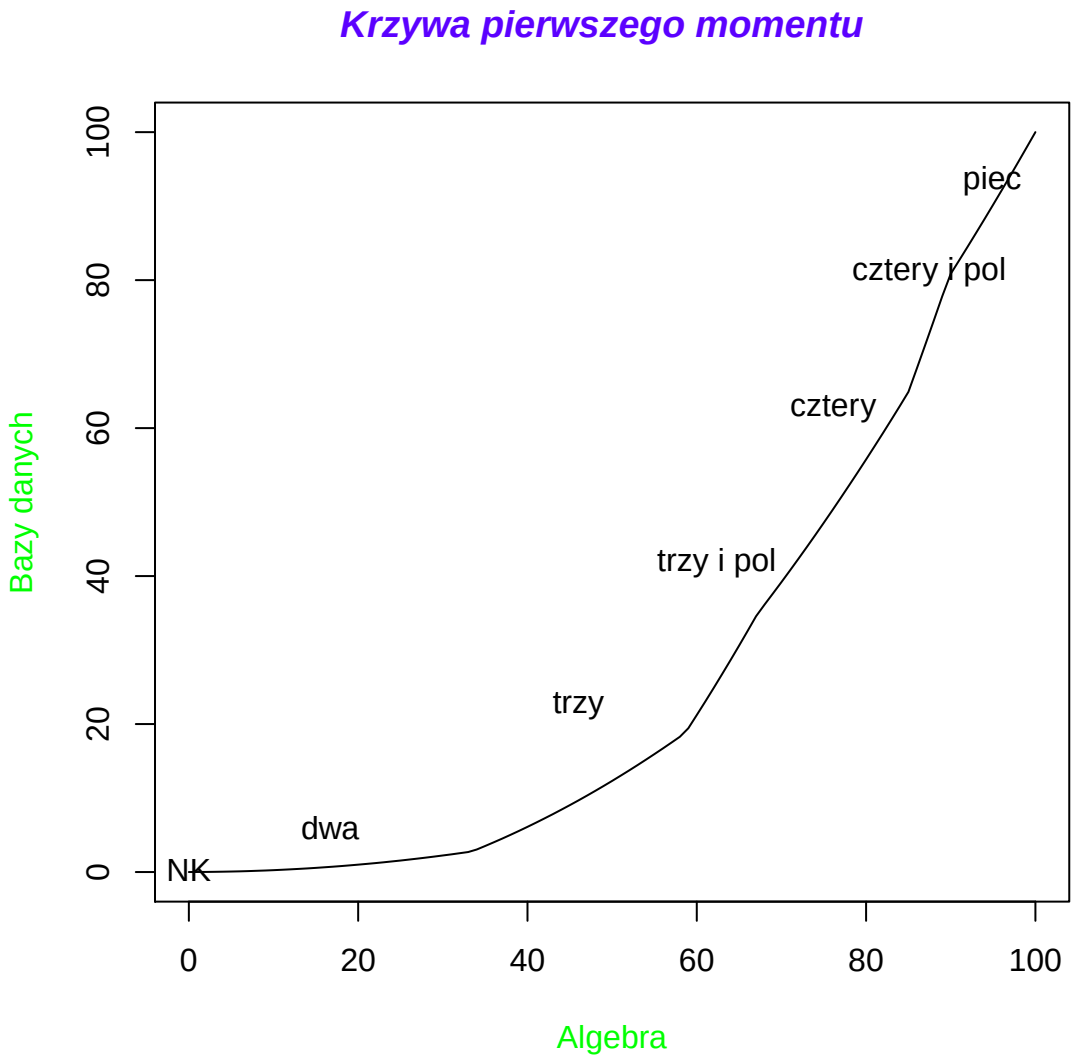
Definicja 2.2.7 *Dla każdej zmiennej $\xi \in \mathbb{ULM}$ możemy zdefiniować tzw. zmienną pierwszego momentu $\xi^{(1)}$ o dystrybuancie danej wzorem:*

$$F_{\xi^{(1)}}(u) = \frac{E(\xi; \xi \leq u)}{E(\xi)},$$

gdzie $E(\xi; \xi \leq u) = \int_0^u tdF\xi(t)$.



Rysunek 2.4: Powierzchnie indeksu Giniego-Simpsona dla zmiennej Bazy danych



Rysunek 2.5: Krzywa pierwszego momentu względem zmiennej Algebra dla zmiennej Bazy danych

Rysunek 2.5 ilustruje krzywą dystrybuanty zmiennej pierwszego momentu dla wspomnianej zmiennej lilipuciej Bazy-danych-względem-Algebry o dystrybuancie z rys. 3.1.

Definicja 2.2.8 Wariancję zmiennej lilipuciej *ozn.* $var(\xi)$ *definiujemy jako:*

$$var(\xi) = E(\xi - E(\xi))^2 = E(\xi^2) - (E(\xi))^2$$

Mamy poniższy związek:

$$\frac{var(\xi)}{E(\xi)} = E(\xi^{(1)}) - E(\xi) = \frac{ar(\xi^{(1)}) - ar(\xi)}{2}.$$

Z powyższego wynika, że "zmienność" (ang. index of dispersion, dispersion index, coefficient of dispersion, lub variance-to-mean ratio (VMR) lub Fano index) $V(\xi) = \frac{var(\xi)}{E(\xi)}$ może być przedstawiona w następujący sposób:

$$\begin{aligned} V(\xi) &= \tfrac{ar(\xi^{(1)}) - ar(\xi)}{2} \\ &= \tfrac{2\int_0^1 (u - F_{\xi^{(1)}}(u))du - 2\int_0^1 (u - F_\xi(u))du}{2} \\ &= \int_0^1 (u - F_{\xi^{(1)}}(u) - u + F_\xi(u))du \\ &= \int_0^1 (F_\xi(u) - F_{\xi^{(1)}}(u))du \\ &= \left(\int_0^1 F_\xi(u)du\right) - \left(\int_0^1 F_{\xi^{(1)}}(u)du\right). \end{aligned} \tag{2.2}$$

Wobec tego zmienność jest obszarem między krzywymi $F_\xi(u)$ i $F_{\xi^{(1)}}(u)$.

Rysunek 2.6 ilustruje powierzchnię reprezentującą zmienność wspomnianej zmiennej lilipuciej Bazy-danych-względem-Algebry o dystrybuancie z rys. 3.1.

Wariancja z kolei to zmienność razy wartość oczekiwana, która jest mniejsza od jeden, więc można ją przedstawić jako stosownie proporcjonalny wycinek obszaru między tymi krzywymi.

Definicja 2.2.9 Wygładzoną funkcję ξ *definiujemy jako:*

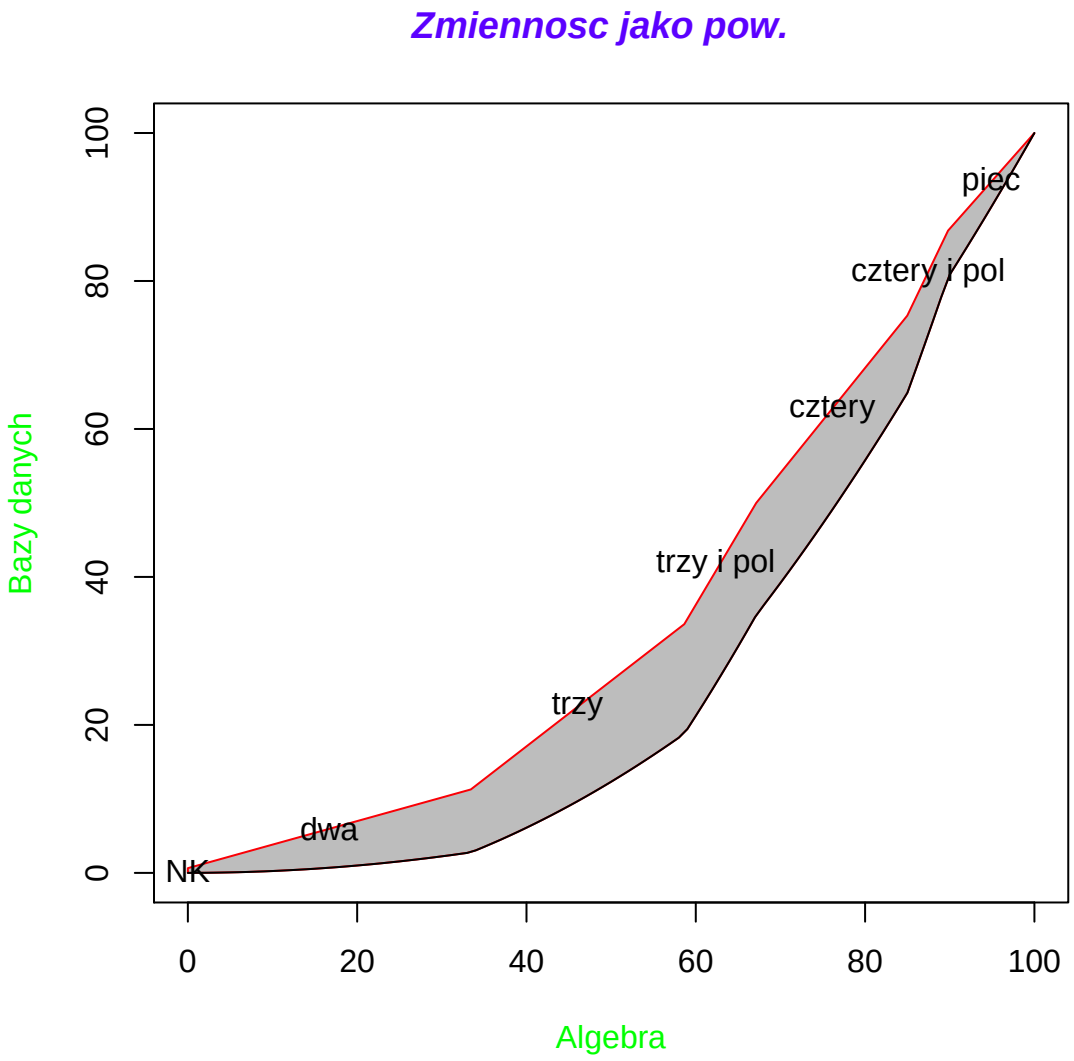
$$\tilde{F}_\xi(\xi) = \frac{F_\xi(\xi+) - F_\xi(\xi-)}{2},$$

gdzie $F_\xi(\xi+)$ *to zmienna losowa przyjmująca wartości prawostronnej granicy funkcji* $F_\xi(u)$ *w punkcie* $\xi = u$*, a* $F_\xi(\xi-)$ *to zmienna losowa przyjmująca wartości lewostronnej granicy funkcji* $F_\xi(u)$ *w punkcie* $\xi = u$*.*

Niech F_Z będzie dystrybuantą zmiennej losowej Z.

Lemat 2.2.2 *Jeśli Z jest zmienną ciągłą, to zmienna* $F_Z(Z)$ *ma rozkład jednostajny na przedziale* $[0, 1]$*.*

Dowód Zachodzi to, ponieważ taka transformacja przekształca wartości zmiennej Z z przedziału $[-\infty, z]$ do $[F_Z(-\infty), F_Z(z)] = [0, u]$. $F_Z(z)$ przyjmuje wartości z przedziału $[0, 1]$ z prawdopodobieństwem $F_Z(z) = u$, czyli ma rozkład jednostajny.



Rysunek 2.6: Powierzchnia zmienności względem zmiennej Algebra dla zmiennej Bazy danych. Powierzchnia zmienności to pole między dystrybuantą (krzywa koloru czerwonego) a krzywą pierwszego momentu (krzywa koloru czarnego).

Taką transformację opisuje tzw. *funkcja prawdopodobieństwa przejścia* $G_Z([0,u];z)$ zdefiniowana następująco:

$$G_Z([0,u];z) = \begin{cases} 1 & \text{dla } F_Z(z) \leq u \\ 0 & \text{dla } F_Z(z) > u \end{cases}$$

Zmienna losowa otrzymana po zastosowaniu funkcji G_Z ma rozkład jednostajny, gdyż:

$$\int_0^1 G_Z([0,u];z)dF_Z(z) = \int_{F_Z(z)\leq u} dF_Z(z) = u$$

Jeśli F_Z jest zmienną nieciągłą, to dla każdego $u \in (0,1)$ istnieje punkt z_u taki, że $F_Z(z_u-) \leq F_Z(z_u)$. Dlatego musimy użyć innej funkcji prawdopodobieństwa przejścia tzw. *monotonicznej funkcji prawdopodobieństwa przejścia* F_Z^* zdefiniowanej następująco:

$$F_Z^*([0,u];z) = \begin{cases} 1 & \text{dla } F_Z(z) \leq u \\ \frac{u-F_Z(z-)}{F_Z(z)-F_Z(z-)} & \text{dla } F_Z(z-) < u \leq F_Z(z) \\ 0 & \text{dla } F_Z(z) > u \end{cases}$$

Przejście ze zmiennej losowej X o wartościach rzeczywistych do zmiennej lilipuciej U o rozkładzie jednostajnym nazwiemy *jednostajnym przekształceniem gradacyjnym zmiennej losowej X.*

Dzięki przekształceniom gradacyjnym otrzymujemy uciągloną dystrybuantę dla zmiennej lilipuciej, zwaną krzywą koncentracji, bądź dla pary zmiennych lilipucich zwaną powierzchnią koncentracji.

Rozdział 3

Metody gradacyjne dla pary rozkładów niezależnych

3.1 Krzywe koncentracji

Niech zmienne losowe X i Y mają dystrybuanty istniejące wszędzie na zbiorze liczb rzeczywistych.

Definicja 3.1.1 *Rozpatrzmy krzywą w* $[0,1] \times [0,1]$ *zadaną wzorem:*

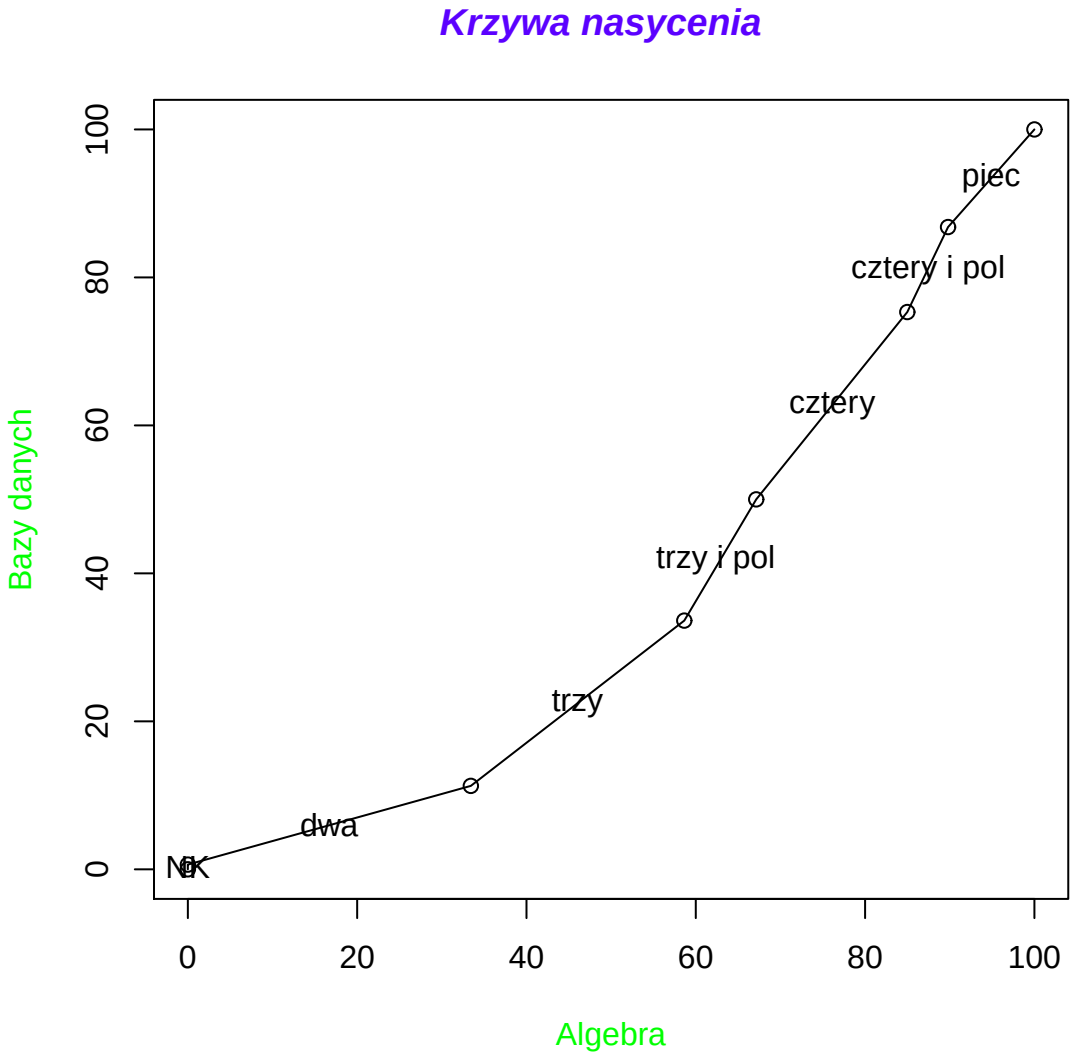
$$Z = \{(F_X(x), F_Y(x)); x \in \mathbb{R}\}$$

Niech krzywa C *składa się z krzywej* Z *oraz następujących punktów:*

- $(0,0) = (F_X(-\infty), F_Y(-\infty))$
- $(1,1) = (F_X(+\infty), F_Y(+\infty))$
- *Dla punktu* x_0 *niech* $(F_X(x_0), F_Y(x_0))^+ ::= (\lim_{x \to x_0^+} F_X(x), \lim_{x \to x_0^+} F_Y(x))$. *Jeśli* $(F_X(x_0), F_Y(x_0))^+ \neq (F_X(x_0), F_Y(x_0))$, *wtedy* C *zawiera także* $\lambda(F_X(x_0), F_Y(x_0))^+ + (1-\lambda)(F_X(x_0), F_Y(x_0))$ *dla wszystkich* $\lambda \in [0,1]$.
- *Dla punktu* x_0 *niech* $(F_X(x_0), F_Y(x_0))^- ::= (\lim_{x \to x_0^-} F_X(x), \lim_{x \to x_0^-} F_Y(x))$. *Jeśli* $(F_X(x_0), F_Y(x_0))^- \neq (F_X(x_0), F_Y(x_0))$, *wtedy* C *zawiera także* $\lambda(F_X(x_0), F_Y(x_0))^- + (1-\lambda)(F_X(x_0), F_Y(x_0))$ *dla wszystkich* $\lambda \in [0,1]$.

Ostatnie dwie kropki dotyczą też wartości niewłaściwych x *czyli* $-\infty$ *oraz* $+\infty$. *Krzywa ta łączy punkty* $(0,0)$ *i* $(1,1)$ *i nazywamy ją* krzywą koncentracji Y do X *(lub krzywą nasycenia) i oznaczamy przez* $C(Y:X)$.

W sposób nieformalny mówimy, że $C(Y:X)$ zawiera Z i jest uzupełniana przez interpolację liniową. Na rysunku 3.1 narysowano przykładowy wykres krzywej koncentracji dla zmiennej Bazy danych względem zmiennej Algebra, pochodzących z bazy USOS (patrz rozdział 5). Dziedzinami obu zmiennych są oceny: NK (nie klasyfikowani), oraz oceny 2, 3, 3.5, 4, 4.5 i 5. Nachylenie poszczególnych odcinków powstałej łamanej mówi nam o proporcjach między częstościami poszczególnych ocen dla obu przedmiotów. I tak udział dwójek z Baz danych jest zdecydowanie wyższy niż z Algebry (ponad 30% kontra ok. 10%), podczas gdy kąt nachylenia dla 4.5 jest powyżej 45^o, co świadczy o wyższym udziale tych ocen na Bazach



Rysunek 3.1: Krzywa koncentracji względem zmiennej Algebra dla zmiennej Bazy danych

danych w porównaniu z Algebrą. Jeśli krzywa koncentracji przebiega poniżej przekątnej, jak ma to miejsce na rysunku, to znaczy że generalnie udział danej oceny i ocen niższych jest mniejszy dla Baz danych niż dla Algebry (albo mówiąc odwrotnie z baz danych uzyskuje się lepsze oceny niż z Algebry).

O parze zmiennych $Y : X$ mówimy, że jest ona C-równoważna parze zmiennych $Q : S$, jeśli ich krzywe $C(Y : X)$ i $C(Q : S)$ są identyczne.

Twierdzenie 3.1.1 *Dla pary zmiennych porządkowych (nad tą samą dziedziną), krzywa koncentracji nie zależy od kodowania wartości liczbami rzeczywistymi przy zachowaniu porządku dziedziny.*

Dowód Aby przekonać się o prawdziwości tego twierdzenia zauważmy, że wartości zmiennych nie odgrywają przy konstrukcji krzywej żadnej roli, a jedynie fakt, że dla tej samej wartości zmiennych porównujemy ich dystrybuanty i wartości funkcji dystrybuant odgrywają rolę przy kreśleniu krzywej.

W szczególności np. krzywe koncentracji dla zmiennych X, Y będą takie same jak dla X^3, Y^3.

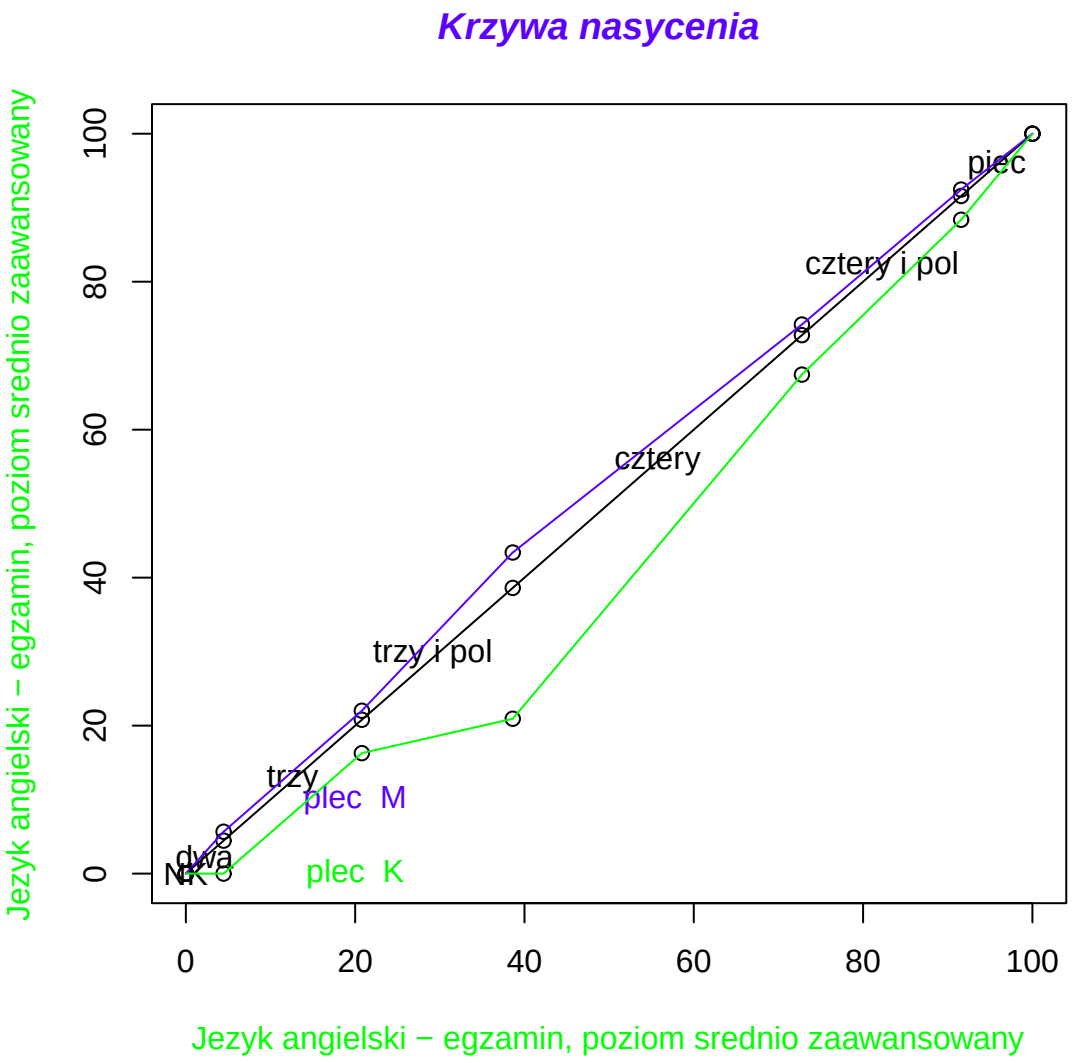
W GAD rozpatruje się także krzywe koncentracji, w których jedna ze zmiennych przebiega podzbiór obiektów, które przebiega druga z nich. Wtedy takich podzmiennych nanosi się zwykle kilka. I tak na rysunku 3.2 zmienna Język angielski - egzamin, poziom średnio zaawansowany naniesiona na oś X jest przyrównywana do tej samej zmiennej zawężonej raz do zbioru mężczyzn, a raz do zbioru kobiet. Na rysunku widać, że studentki mają lepsze oceny od studentów.

Odpowiednikiem krzywej koncentracji w statystyce klasycznej jest krzywa Lorenza (patrz [Kle05] oraz rozdz.3.2), która podobnie jest łamaną na kwadracie jednostkowym.

3.2 Krzywa Lorenza a krzywa nasycenia GAD

Krzywa Lorenza jest stosowana w ekonomii do wizualizacji dystrybucji dochodów bądź bogactwa w aspekcie "nierówności" lub "koncentracji" [Kle05]. Zakłada się pewną populację obiektów o, z których każdy jest opisany pewną ciągłą addytywną zmienną (cechą) $c(o)$, np. o to gospodarstwa domowe, a $c(o)$ to ich roczne dochody. Niech ciąg $\{o_i\}$ będzie ciągiem wszystkich tych obiektów bez powtórzeń taki, że dla $i > 0$ zachodzi $c(o_{i-1}) \leq c(o_i)$. Niech obiektów tych będzie n+1; Krzywa Lorenza składa się z punktów $\left(\frac{i}{n}, \frac{\sum_{j=0}^{i} c(o_j)}{\sum_{j=0}^{n} c(o_j)}\right)$ dla $i = 0, ..., n$ oraz odcinków łączących kolejne punkty na płaszczyźnie.

Krzywa nasycenia (lub koncentracji) zdefiniowana w ramach GAD jest w pewnym sensie uogólnieniem krzywej Lorenza i graficznie może się wydawać łudząco podobna, ale istnieje zasadnicza różnica koncepcyjna. Mianowicie po pierwsze dominującym obszarem zastosowania Krzywej Lorenza są zagadnienia ekonomiczne, podczas gdy twórcy GAD eksplorowali takie dziedziny jak medycyna, wyniki wyborów i zakładają, że możliwe jest prezentowanie danych z dowolnych dziedzin, np. ocen studentów na uczelni. Po drugie cecha obiektu, który może być przedmiotem porównania, nie musi być ciągła, ale dopuszcza się porządkowe a nawet nominalne wartości. Po trzecie, to nie cecha c jest przedmiotem zainteresowania lecz jej kumulatywne sumy. Wreszcie po czwarte - i to chyba jest fundamentalna różnica - krzywe nasycenia GAD nadają się do (i są chyba przeznaczone) analizy niezależnych próbek. W tym



Rysunek 3.2: Płciowa krzywa koncentracji dla zmiennej Język angielski - egzamin, poziom średnio zaawansowany. Zmienna dla całego zbioru (czarna krzywa) jest przyrównywana do tej samej zmiennej zawężonej raz do zbioru mężczyzn (krzywa niebieska), a raz do zbioru kobiet (krzywa zielona). Na rysunku widać, że studentki mają lepsze oceny od studentów.

kontekście jedna z realizacji krzywej może odzwierciedlać statystykę testu Manna-Witneya-Wilcoxona.

Warto ponadto podkreślić, że krzywa Lorenza ze swej natury jest wypukła [Kle05], a krzywa nasycenia GAD - niekoniecznie.

3.3 Współczynnik polowy

Definicję indeksu ar można rozszerzyć dla par zmiennych losowych $(X;Y)$ o wartościach rzeczywistych: Niech para zmiennych $Y : X$ będzie C-równoważna parze zmiennych $\xi : U$. Wtedy:

$$ar(Y : X) := ar(\xi)$$

Zachodzi wtedy:

$$ar(\xi) = ar(\xi : U)$$

Jest uzupełnieniem indeksu Giniego-Simpsona (który będziemy oznaczać jako $GS(\xi)$), o którym piszemy dalej.

3.4 Indeks Giniego a współczynnik polowy

W ekonomii indeks Giniego jest rozważany jako podwojone pole między krzywą Lorenza a odcinkiem ((0,0),(1,1)). Indeks Giniego przyjmuje wartości od 0 (dla społeczeństwa egalitarnego) po 1 (dla społeczeństwa totalitarnego, ze skupieniem bogactwa w jednym ręku).

W GAD pojęcie to w oczywisty sposób zostało uogólnione do współczynnika polowego. Jednakże uogólnienie krzywej Lorenza do krzywej nasycenia niesie ze sobą podstawową konsekwencję polegającą na możliwości górowania tejże nad przekątną - dla pola powyżej przekątnej zarezerwowano więc wartości ujemne, a dla poniżej - dodatnie. W GAD współczynnik polowy z ekonomii może zatem przyjmować wartości od -1 do 1.

3.5 Indeks Giniego-Simpsona

Niech para zmiennych $Y : X$ będzie C-równoważna parze zmiennych $\xi : U$. $GS(\xi)$ będzie wtedy miarą odmienności zmiennych X, Y.

3.6 Klasyczny indeks Giniego-Simpsona a indeks GS w GAD

Indeks GS w GAD, wspomniany wyżej, to adaptacja klasycznego indeksu Giniego-Simpsona, stosowanego pierwotnie w ekologii, sociologii, psychologii oraz w nauce o zarządzaniu. Klasyczny indeks Giniego-Simpsona wyraża prawdopodobieństwo, że dwa obiekty wylosowane ze zwracaniem będą różnego rodzaju (np. gatunku). W ekologii nazywa się go stąd prawdopodobieństwem spotkania międzygatunkowego (PIE) [JJvEC02] i wyrażany jest formułą

$$1 - \sum_{i=1}^{N} p_i^2$$

gdzie N to np. liczba gatunków. Inne nazwy tego indeksu to indeks Gibbsa-Martina lub indeks Blaua.

Indeks GS w GAD różni się nieco koncepcyjnie po pierwsze dlatego, że bierze pod uwagę poniekąd "odległość" między losowanymi obiektami (stąd ma mianownik, który jakby normalizuje przez średnią wartość prawdopodobieństwa). Po drugie obiekty przynależące do jednej grupy nie są traktowane identycznie. Milcząco zakłada się w analogii np. do ocen studentów, że podział na kategorie (oceny studentów) jest nieco sztuczny, powstały przez progowanie, wobec czego w tle za stopniującymi kategoriami jest jakieś domyślne kontinuum.

3.7 Funkcjonał gradacyjny

Niech P_X oraz Q_Y będą miarami indukowanymi przez zmienne losowe odpowiednio X i Y. *Iloraz wiarogodności* $h_{Y:X}$ jest zdefiniowany na zbiorze $\frac{dP_X}{d(P_X+Q_Y)}(z) > 0$ jako:

$$h_{Y:X} = \frac{dQ_Y}{d(P_X+Q_Y)} / \frac{dP_X}{d(P_X+Q_Y)}(z).$$

Jeśli funkcjonał ten jest rosnący, to wtedy rozkład jest reprezentowany przez wypukłą dystrybuantę, a ciągła zmienna lilipucia ma niemalejącą gęstość.

Rozdział 4

Analiza rozkładów zależnych

4.1 Rozkłady dwuwymiarowe

Na wzór Jednowymiarowego Modelu Lilipuciego wprowadźmy Dwuwymiarowy Model Lilipuci.

Definicja 4.1.1 *Parę zmiennych losowych nazwiemy* lilipucią parą zmiennych *(ozn. ξ, ν), jeśli przyjmuje wartości z produktu przedziałów $[0,1] \times [0,1]$, jest dyskretna, ciągła bądź dyskretna-ciągła. Dystrybuanta oznaczana przez $F_{\xi,\nu}$ jest niemalejącą powierzchnią, leżącą w kostce jednostkowej, łączącej punkty (0,0,0) oraz (1,1,1).*

Zbiór wszystkich lilipucich par zmiennych losowych nazywamy *Dwuwymiarowym Modelem Lilipucim* i oznaczamy przez $\mathbb{BLM}$ (z ang. the Bivariate Lilliputian Model)

4.2 Powierzchnia koncentracji

Niech zmienne losowe X i Y mają dystrybuanty F_X, F_Y oraz łączną dystrybuantę F_{XY} istniejące wszędzie na zbiorze liczb rzeczywistych.

Definicja 4.2.1 *Rozpatrzmy powierzchnię w $[0,1] \times [0,1] \times [0,1]$ zadaną wzorem:*

$$Z = \{(F_X(x), F_Y(y), F_{XY}(x,y)); x, y \in \mathbb{R}\}$$

Niech powierzchnia S składa się z powierzchni Z oraz następujących punktów:

- $(0,0,0) = (F_X(-\infty), F_Y(-\infty), F_{XY}(-\infty, -\infty))$
- $(1,1,1) = (F_X(+\infty), F_Y(+\infty), F_{XY}(-\infty, -\infty))$
- *Dla punktu (x_0, y_0) niech*
  - $(F_X(x_0), F_Y(y_0), F_{XY}(x_0,y_0))^{++} ::= (\lim_{x\to x_0^+} F_X(x), \lim_{y\to y_0^+} F_Y(y), \lim_{x\to x_0^+, y\to y_0^+} F_{XY}(x,y))$.
  - $(F_X(x_0), F_Y(y_0), F_{XY}(x_0,y_0))^{+\cdot} ::= (\lim_{x\to x_0^+} F_X(x), F_Y(y_0), \lim_{y\to x_0^+} F_{XY}(x,y_0))$.
  - $(F_X(x_0), F_Y(y_0), F_{XY}(x_0,y_0))^{\cdot+} ::= (F_X(x_0), \lim_{y\to y_0^+} F_Y(y), \lim_{y\to y_0^+} F_{XY}(x_0,y))$.
  - $(F_X(x_0), F_Y(y_0), F_{XY}(x_0,y_0))^{--} ::= (\lim_{x\to x_0^+} F_X(x), \lim_{y\to y_0^-} F_Y(y), \lim_{y\to x_0^+, y\to y_0^-} F_{XY}(x,y))$.
  - $(F_X(x_0), F_Y(y_0), F_{XY}(x_0,y_0))^{-\cdot} ::= (\lim_{x\to x_0^-} F_X(x), F_Y(y_0), \lim_{x\to x_0^-} F_{XY}(x,y_0))$.

- $(F_X(x_0), F_Y(y_0), F_{XY}(x_0, y_0))^{\cdot -} ::= (F_X(x_0), \lim_{y \to y_0^-} F_Y(y), \lim_{y \to y_0^-} F_{XY}(x_0, y))$.
- $(F_X(x_0), F_Y(y_0), F_{XY}(x_0, y_0))^{+-} ::= (\lim_{x \to x_0^+} F_X(x), \lim_{y \to y_0^-} F_Y(y), \lim_{x \to x_0^+, y \to y_0^-} F_{XY}(x, y))$.
- $(F_X(x_0), F_Y(y_0), F_{XY}(x_0, y_0))^{-+} ::= (\lim_{x \to x_0^-} F_X(x), \lim_{y \to y_0^+} F_Y(y), \lim_{x \to x_0^-, y \to y_0^+} F_{XY}(x, y))$.

- *Jeśli* $(F_X(x_0), F_Y(y_0), F_{XY}(x_0, y_0))^{++} \neq (F_X(x_0), F_Y(y_0), F_{XY}(x_0, y_0))$, *wtedy* S *zawiera także górną powierzchnię* $\lambda_X \lambda_Y (F_X(x_0), F_Y(y_0), F_{XY}(x_0, y_0))^{++} + (1 - \lambda_X)\lambda_Y (F_X(x_0), F_Y(y_0), F_{XY}(x_0, y_0))^{\cdot +} + \lambda_X (1 - \lambda_Y)(F_X(x_0), F_Y(y_0), F_{XY}(x_0, y_0))^{+\cdot} + (1 - \lambda_X)(1 - \lambda_Y)(F_X(x_0), F_Y(x_0))$ *dla wszystkich* $\lambda_X, \lambda_Y \in [0, 1]$.

- *Jeśli* $(F_X(x_0), F_Y(y_0), F_{XY}(x_0, y_0))^{--} \neq (F_X(x_0), F_Y(y_0), F_{XY}(x_0, y_0))$, *wtedy* S *zawiera także powierzchnię* $\lambda_X (F_X(x_0), F_Y(y_0), F_{XY}(x_0, y_0))^{--} + (1 - \lambda_X)(1 - \lambda_Y)(F_X(x_0), F_Y(y_0), F_{XY}(x_0, y_0))^{\cdot -} + \lambda_Y (F_X(x_0), F_Y(x_0))$ *oraz powierzchnię* $\lambda_X (F_X(x_0), F_Y(y_0), F_{XY}(x_0, y_0))^{--} + (1 - \lambda_X)(1 - \lambda_Y)(F_X(x_0), F_Y(y_0), F_{XY}(x_0, y_0))^{-\cdot} + \lambda_Y (F_X(x_0), F_Y(x_0))$ *a pochodna na tych powierzchniach przyjmowana jest umownie* $\frac{f_{xy} - f_{xy}^{--}}{(f_x - f_x^-)(f_y - f_y^-)}$ *dla wszystkich* $\lambda_X, \lambda_Y \in [0, 1]$. *i tak dalej ....*

Dotyczy to też wartości niewłaściwych x oraz y czyli $-\infty$ *oraz* $+\infty$. *Powierzchnia ta łączy punkty* $(0, 0, 0)$ *i* $(1, 1, 1)$ *i nazywamy ją* powierzchnią koncentracji Y do X *i oznaczamy przez* $S(Y : X)$.

O parze zmiennych $Y : X$ mówimy, że jest ona S-równoważna parze zmiennych $Q : S$, jeśli ich powierzchnie $S(Y : X)$ i $S(Q : S)$ są identyczne.

4.3 Mapy nadreprezentacji

W przestrzeni BLM dla dystrybuanty reprezentowanej przez powierzchnię $S(\nu : \xi)$ mapą nadreprezentacji nazywamy funkcję pochodną $\frac{\delta S}{\delta \xi \delta \nu}$, zdefiniowaną wszędzie tam, gdzie istnieje.

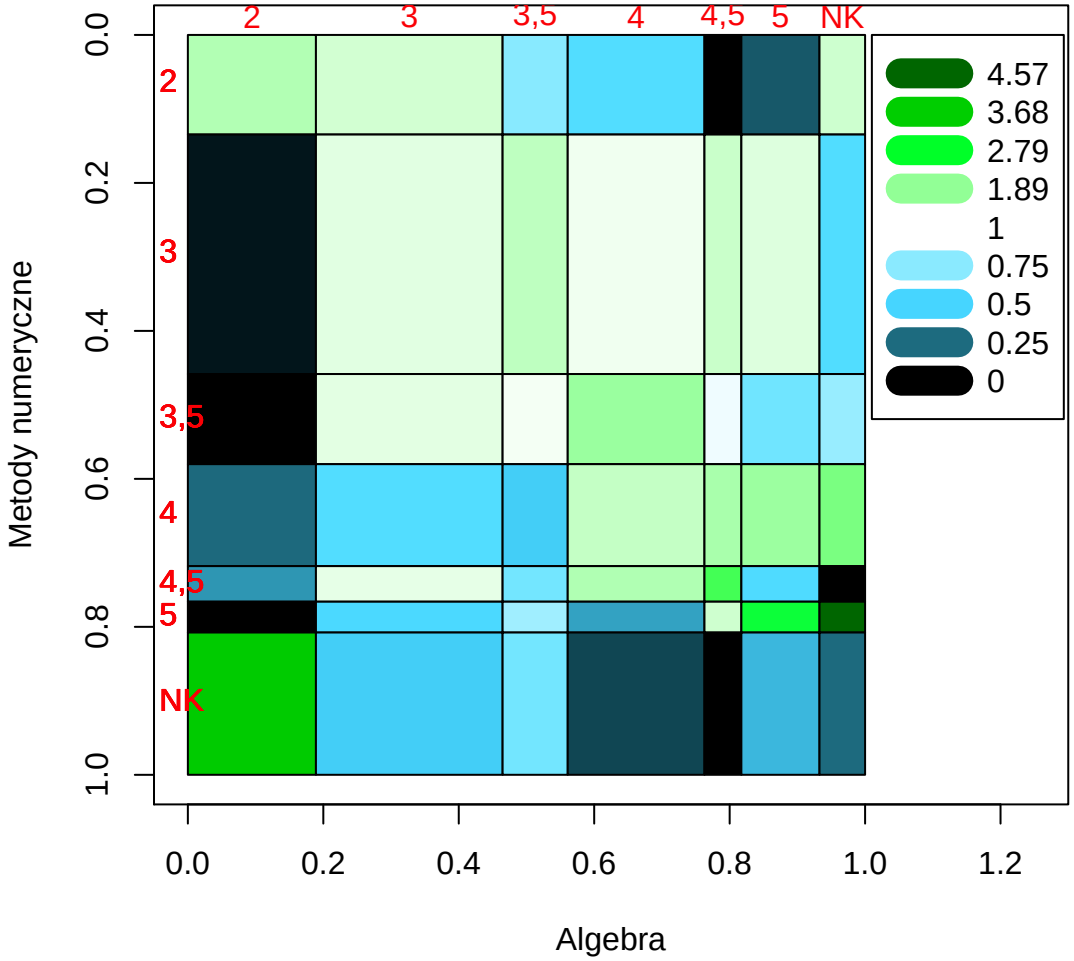
Mapy nadreprezentacji są użytecznym narzędziem przy wizualizacji trendów zależności.

Na rysunku 4.1 zaprezentowana została mapa nadreprezentacji dla dwóch przedmiotów akademickich: Algebra i Metody numeryczne. Pokazuje zależności między ocenami studentów, którzy uczęszczali na oba te przedmioty. Im większe jest nasycenie koloru zielonego, tym większa wartość nadreprezentacji odpowiadająca komórce mapy. Natomiast zwiększenie nasycenia koloru niebieskiego oznacza zmniejszanie się wartości nadreprezentacji. Niekiedy wartości reprezentowane przez odcienie koloru niebieskiego nazywane są niedoreprezentacją.

Gdy spojrzymy na komórkę z liczbą 4.5, zauważymy, że jest ona prawie symetryczna. Oznacza to, że mniej więcej tyle samo ludzi ma 4.5 z Algebry i Metod numerycznych. Wielkość komórki wskazuje na ich niewielką liczbę w stosunku do wszystkich osób zapisanych na te przedmioty. Komórka jest koloru jasno-zielonego, czyli zagęszczenie studentów w tej komórce jest około 2 razy większe, niż jakie byłoby, gdyby byli oni rozłożeni losowo. Podobnie mamy dla prostokąta z liczbą 5, z tym, że zagęszczenie studentów jest tu prawie 3 razy większe. Patrząc na ten rysunek można dojść do wniosku, że oceny skrajne (2, 4.5 i 5) są dla obu przedmiotów skojarzone.

Dwuwymiarowy indeks koncentracji nazwiemy *współczynnikiem* ρ^* *Spearmana* i zdefiniujemy jako:

$$\rho^*(X, Y) = \frac{\int_0^1 \int_0^1 Cop_{(X,Y)}(u, v) du dv - \int_0^1 \int_0^1 uv du dv}{\int_0^1 \int_0^1 Fr^+_{(X^*, Y^*)}(u, v) du dv - \int_0^1 \int_0^1 uv du dv}$$



Rysunek 4.1: Mapa nadreprezentacji względem zmiennej Algebra dla zmiennej Metody numeryczne. Pokazuje zależności między ocenami studentów, którzy uczęszczali na oba te przedmioty. Im większe jest nasycenie koloru zielonego, tym większa wartość nadreprezentacji odpowiadająca komórce mapy. Zwiększenie nasycenia koloru niebieskiego, oznacza zmniejszanie się wartości nadreprezentacji. Wielkość komórki wskazuje na liczbę studentów o danych ocenach w stosunku do wszystkich osób zapisanych na te przedmioty. Z rysunku widać, że oceny skrajne (2, 4.5 i 5) są dla obu przedmiotów skojarzone.

$$= 12 \int_0^1 \int_0^1 (Cop_{(X,Y)}(u,v) - uv) du dv$$

4.3.1 Dyskretne mapy nadreprezentacji

Jeśli obie zmienne X, Y są dyskretne, z tą samą dziedziną ($dom(X)$), mapę nadreprezentacji otrzymujemy w następujący sposób: Konstruujemy tabelę kontyngencji $T_{X,Y}$ taką, że $T_{X,Y}[i,j] = k$ oznacza, że dla k obiektów atrybuty przyjmują wartości: $X = dom(X)[i], Y = dom(X)[j]$. Następnie konstruujemy tablicę mapy

$$M_{X,Y}[i,j] = \frac{T_{X,Y}[i,j] \cdot (\sum_{i \in dom(X)} \sum_{j \in dom(X)} T_{X,Y}[i,j])}{(\sum_{i \in dom(X)} T_{X,Y}[i,j]) \cdot (\sum_{j \in dom(X)} T_{X,Y}[i,j])}$$

dla tych par i, j, dla których mianownik jest niezerowy, a dla pozostałych $M_{X,Y}[i,j] = 1$.

Wizualizacja mapy to prostokąt podzielony na prostokąty za pomocą kraty tak, aby pole prostokąta nr (i,j) było proporcjonalne do $(\sum_{i \in dom(X)} T_{X,Y}[i,j]) \cdot (\sum_{j \in dom(X)} T_{X,Y}[i,j])$.

Kolor tego prostokąta odzwierciedla wartość $M_{X,Y}[i,j]$.

Twierdzenie 4.3.1 *Jeżeli zmienne X,Y są niezależne, to wartość oczekiwana* $M_{X,Y}[i,j] = 1$.

Dowód Niech c oznacza liczbę elementów w komórce (i,j) tablicy T, w' liczbę elementów w wierszu i prócz komórki (i,j), k' liczbę elementów w kolumnie j prócz komórki (i,j), a N' liczbę elementów poza itym wierszem i jtą kolumną. $w' + k' + c + N' = N$ czyli łącznej liczbie elementów. $w = w' + c, k = k' + c$. Wartość oczekiwana dla ustalonego N wyraża się wzorem

$$E(M_{X,Y}[i,j]) = \sum_{c,w,k;0 \le c \le w, c \le k, w \le N, k \le N} \frac{cN}{wk} \binom{N}{c} (p_i \cdot p_j)^c \binom{N-c}{k'} ((1-p_i) \cdot p_j)^{k'}$$

$$\binom{N-c-k'}{w'} (p_i \cdot (1-p_j))^{w'} ((1-p_i) \cdot (1-p_j))^{N-c-w'-k'}$$

$$= N \sum_{c=1}^{N} \binom{N}{c} c \cdot (p_i \cdot p_j)^c \sum_{k'=0}^{N-c} \binom{N-c}{k'} \frac{1}{k'+c} ((1-p_i) \cdot p_j)^{k'}$$

$$\begin{aligned} &\sum_{w'=0}^{N-c-k'} \binom{N-c-k'}{w'} \frac{1}{w'+c} (p_i \cdot (1-p_j))^{w'} ((1-p_i) \cdot (1-p_j))^{N-w'-k'-c} \\ &= N \sum_{c=1}^{N} \binom{N}{c} c \cdot (p_i \cdot p_j)^c \sum_{k'=0}^{N-c} \binom{N-c}{k'} \frac{1}{k'+c} ((1-p_i) \cdot p_j)^{k'} (1-p_j)^{N-k'-c} \end{aligned} \tag{4.1}$$

$$\begin{aligned} &\sum_{w'=0}^{N-c-k'} \binom{N-c-k'}{w'} \frac{1}{w'+c} (p_i)^{w'} ((1-p_i))^{N-w'-k'-c} \\ &= N \sum_{c=1}^{N} \binom{N}{c} c \cdot (p_i \cdot p_j)^c (1-p_i p_j)^{N-c} \sum_{k'=0}^{N-c} \binom{N-c}{k'} \frac{1}{k'+c} \left(\frac{(1-p_i) \cdot p_j}{1-p_i p_j}\right)^{k'} \left(\frac{1-p_j}{1-p_i p_j}\right)^{N-k'-c} \end{aligned}$$

$$\sum_{w'=0}^{N-c-k'} \binom{N-c-k'}{w'} \frac{1}{w'+c} (p_i)^{w'} ((1-p_i))^{N-w'-k'-c}$$

W powyższym przyjęto sumowanie od c=1, ponieważ dla c=0 składnik znika.

Zauważmy, że

$$E(\frac{1}{1+X}) = \frac{1-(1-p)^{n+1}}{p(n+1)}$$

gdzie $X \sim Bin(n,p)$. Powód jest następujący:

$$
\begin{aligned}
E(\tfrac{1}{1+X}) &= \textstyle\sum_{x=0}^{n} \frac{1}{1+x} \cdot \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x} \\
&= \textstyle\sum_{x=0}^{n} \frac{1}{1+x} \cdot \frac{n!}{x!\cdot(n-x)!} \cdot p^x \cdot (1-p)^{n-x} \\
&= \textstyle\sum_{x=0}^{n} \frac{n!}{(x+1)!\cdot(n-x)!} \cdot p^x \cdot (1-p)^{n-x} \qquad (4.2) \\
&= \textstyle\sum_{x=0}^{n} \frac{n!}{(x+1)!\cdot(n-x)!} \frac{n+1}{n+1} \cdot p^x \frac{p}{p} \cdot (1-p)^{n-x} \\
&= \tfrac{1}{p(n+1)} \textstyle\sum_{x=0}^{n} \frac{(n+1)!}{(x+1)!\cdot(n+1-(x-1))!} \cdot p^{x+1} \cdot (1-p)^{(n+1)-(x+1)}
\end{aligned}
$$

Zastąpmy $x+1$ przez x' oraz $n+1$ przez n'

$$
\begin{aligned}
&= \tfrac{1}{p(n+1)} \textstyle\sum_{x'=1}^{n'} \frac{n'!}{x'!*(n'-x')!} \cdot p^{x'} \cdot (1-p)^{n'-x'} \\
&= \tfrac{1}{p(n+1)} \left(-(1-p)^{n'} \textstyle\sum_{x'=0}^{n'} \frac{n'!}{x'!\cdot(n'-x')!} \cdot p^{x'} \cdot (1-p)^{n'-x'} \right) \qquad (4.3) \\
&= \tfrac{1}{p(n+1)} \left(-(1-p)^{n'} + 1 \right) = \tfrac{1-(1-p)^{n+1}}{p(n+1)} \qquad .
\end{aligned}
$$

Jeśli wartości n będą "duże", a p "niezbyt małe", to jest to w przybliżeniu $\frac{1}{E(X)}$.

$E(\frac{1}{c+X})$ dla c większych od 1 będą niewątpliwie mniejsze, ale jeśli dla małych x prawdopodobieństwa będą niskie, to dla dużych x składniki sumy będą dominujące, a ponieważ $1/(1+x)$ nie będzie się "nadmiernie" różnić od $1/(c+x)$, więc $E(\frac{1}{c+X})$ nie będzie zbytnio różne od $E(\frac{1}{1+X})$ Możemy zatem przyjąć, że dla stałej N i zmiennych losowych c, w', c' zachodzi

$$
E(M_{X,Y}[i,j]) \approx \frac{NE(c)}{E(w^*)E(k)} = \frac{N \cdot N \cdot p_i \cdot p_j}{NN\left(\frac{(1-p_i)\cdot p_j}{1-p_i p_j}\right) N p_i} \approx 1
$$

(gdzie w^* to rozkład zmodyfikowany jak w powyższych wzorach). Badania symulacyjne[1] dla p_i, p_j z zakresu $0.1 - 0.9$ oraz N rzędu 100 pokazały, że oszacowanie to jest dość dobre $(+/-0.1\%)$.

Twierdzenie 4.3.2 *Minimalna możliwa wartość $M_{X,Y}[i,j]$ wynosi 0.*

Dowód jest prosty - zauważmy, że w tablicy kontyngencji minimalna zawartość komórki to 0. Stąd z definicji tablicy M, o ile mianownik jest niezerowy $M[i,j]$ wyniesie 0. Ujemne wartości nie mogą wystąpić w T, a więc i w mianowniku M.

Twierdzenie 4.3.3 *Maksymalna możliwa wartość $M_{X,Y}[i,j]$ wynosi $\sum_{i \in dom(X)} \sum_{j \in dom(X)} T_{X,Y}[i,j]$.*

Dowód Najpierw pokażemy, że ta wartość jest w ogóle możliwa, a potem, że inne wartości są zawsze nie większe.

[1] Konstruowano tabelę kontyngencji 2x2 dla dwóch niezależnych zmiennych losowych przez losowanie N obiektów ze stosownym prawdopodobieństwem przynależności do poszczególnych komórek

Zatem rozpatrzmy tabelę kontyngencji $T_{X,Y}$ taką, że wiersz i oraz kolumna j zawierają same zera z wyjątkiem komórki (i,j), która zawiera 1. Wtedy

$$\begin{aligned} M_{X,Y}[i,j] &= \frac{T_{X,Y}[i,j]\cdot(\sum_{i\in dom(X)}\sum_{j\in dom(X)}T_{X,Y}[i,j])}{(\sum_{i\in dom(X)}T_{X,Y}[i,j])\cdot(\sum_{j\in dom(X)}T_{X,Y}[i,j])} \\ &= \frac{1\cdot(\sum_{i\in dom(X)}\sum_{j\in dom(X)}T_{X,Y}[i,j])}{1\cdot 1} = \sum_{i\in dom(X)}\sum_{j\in dom(X)}T_{X,Y}[i,j] \end{aligned} \tag{4.4}$$

Ta wartość jest osiągana.

Teraz rozpatrzmy tabelę kontyngencji $T_{X,Y}$ taką, że wiersz i oraz kolumna j zawierają same zera z wyjątkiem komórki (i,j), która zawiera $k > 1$. Wtedy

$$\begin{aligned} M_{X,Y}[i,j] &= \frac{T_{X,Y}[i,j]\cdot(\sum_{i\in dom(X)}\sum_{j\in dom(X)}T_{X,Y}[i,j])}{(\sum_{i\in dom(X)}T_{X,Y}[i,j])\cdot(\sum_{j\in dom(X)}T_{X,Y}[i,j])} \\ &= \frac{k\cdot(\sum_{i\in dom(X)}\sum_{j\in dom(X)}T_{X,Y}[i,j])}{k\cdot k} = \frac{1}{k}\sum_{i\in dom(X)}\sum_{j\in dom(X)}T_{X,Y}[i,j] \end{aligned} \tag{4.5}$$

Wartość mniejsza od (4.4).

Teraz rozpatrzmy ogólną tabelę kontyngencji $T_{X,Y}$ oraz drugą $T'_{X,Y}$ takie, że dla pewnej kolumny l zachodzi: $T'_{X,Y}[i,l] = T_{X,Y}[i,l]-1$ i $T'_{X,Y}[i,j] = T_{X,Y}[i,j]+1$ a pozostałe odpowiednie komórki są identyczne.

Wtedy

$$\begin{aligned} M_{X,Y}[i,j] &= \frac{T_{X,Y}[i,j]\cdot(\sum_{i\in dom(X)}\sum_{j\in dom(X)}T_{X,Y}[i,j])}{(\sum_{i\in dom(X)}T_{X,Y}[i,j])\cdot(\sum_{j\in dom(X)}T_{X,Y}[i,j])} \\ &= \frac{(T'_{X,Y}[i,j]-1)\cdot(\sum_{i\in dom(X)}\sum_{j\in dom(X)}T'_{X,Y}[i,j])}{(\sum_{i\in dom(X)}T'_{X,Y}[i,j])\cdot(-1+\sum_{j\in dom(X)}T'_{X,Y}[i,j])} \\ &\leq \frac{(T'_{X,Y}[i,j])\cdot(\sum_{i\in dom(X)}\sum_{j\in dom(X)}T'_{X,Y}[i,j])}{(\sum_{i\in dom(X)}T'_{X,Y}[i,j])\cdot(\sum_{j\in dom(X)}T'_{X,Y}[i,j])} = M'_{X,Y}[i,j] \end{aligned} \tag{4.6}$$

co oznacza, że "czyszcząc" wiersz i wstawiając wszystko do jednej komórki (i,j) otrzymujemy coraz większe wartości $M[i,j]$. Analogicznie dzieje się dla kolumny j. A na temat wyczyszczonego wiersza i kolumny już się wypowiedzieliśmy, kiedy M jest największe: gdy $T[i,j] = 1$. To kończy dowód.

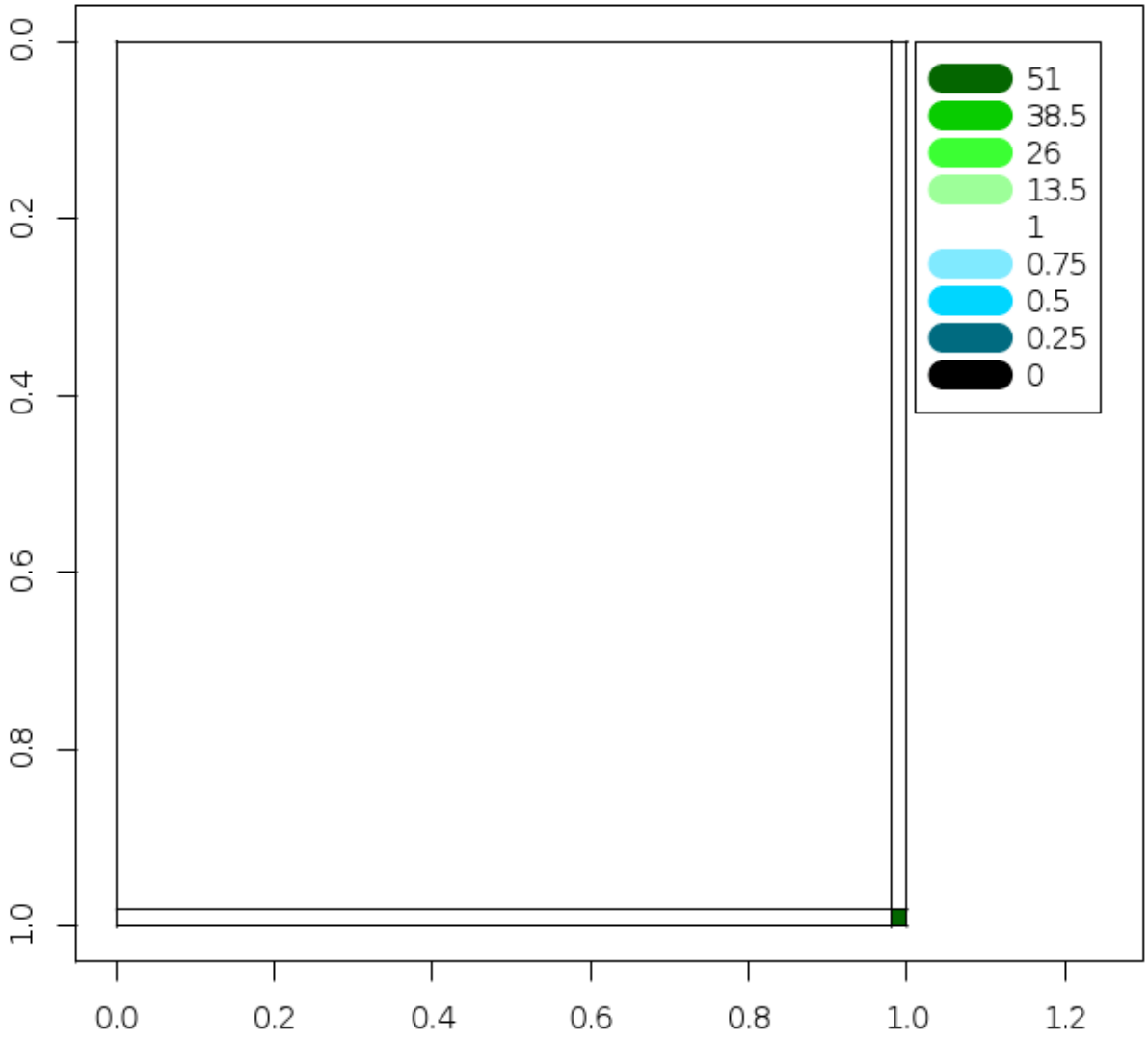
Przykład:

Załóżmy, że wśród pięćdziesięciu jeden studentów, który chodzą na Algebrę i Analizę, 50 ma ocenę 3.5 z Algebry i ocenę 4 z Analizy oraz jeden ma z obu tych przedmiotów ocenę 5. Wówczas wartość komórki tablicy mapy odpowiadająca temu studentowi będzie miała wartość 51. A cała mapa będzie wyglądać tak jak na rysunku (4.2).

Rozkład prawdopodobieństwa komórki mapy:

Niech

- w - oznacza sumę elementów w wierszu tablicy kontyngencji
- k - oznacza sumę elementów w kolumnie tablicy kontyngencji
- c - oznacza sumę elementów w komórce tablicy kontyngencji
- N - oznacza sumę wszystkich elementów w tablicy kontyngencji



Rysunek 4.2: Mapa nadreprezentacji realizująca maksimum dla przykładu.

- oraz $0 < c, k, w \leq N$, $w \geq c$, $k \geq c$, $N \geq c$,

Wtedy rozkład prawdopodobieństwa komórki mapy wyraża się wzorem:

$$p_m = P(c, k, w; N, p_i, p_j) = \binom{N}{c}\binom{N-c}{w-c}\binom{N-w-2c}{k-c} p_c^c p_w^{(w-c)} p_k^{(k-c)} p_r^{(N-w-k-2c)}$$

$$= \binom{N}{c}\binom{N-c}{w-c}\binom{N-w-2c}{k-c} (p_i p_j)^c p_i^{(w-c)} p_j^{(k-c)} (1 - p_i - p_j + p_i p_j)^{(N-w-k-2c)}$$

$$= \binom{N}{c}\binom{N-c}{w-c}\binom{N-w-2c}{k-c} p_i^w p_j^c [(1 - p_i)(1 - p_j)]^{(N-w-k-2c)}$$

A dystrybuanta wyraża się wzorem:

$$F(z) = P(\tfrac{cN}{wk} \leq z) = \sum_{\frac{cN}{wk} \leq z} p_m$$

$$= \sum_{\frac{cN}{wk} \leq z} \binom{N}{c}\binom{N-c}{w-c}\binom{N-w-2c}{k-c} p_i^w p_j^c [(1 - p_i)(1 - p_j)]^{(N-w-k-2c)}$$

Rozdział 5

Analiza danych rzeczywistych

5.1 Cel badań

Celem badań jest wizualizacja procesu studiowania na podstawie systemu USOS Uniwersytetu Warszawskiego.

5.2 Opis zbioru danych

Dane zawierają 6262 pliki z informacjami o ocenach studentów, z poszczególnych przedmiotów, wyciągniętych z systemu USOS Uniwersytetu Warszawskiego. Dane reprezentują wyniki nauczania z okresu od 1999 do 2009 z kilku wydziałów.

Każdy plik dotyczy jednego przedmiotu (egzaminu/zaliczenia za jeden semestr) i zawiera 12 zmiennych:

- os_id - identyfikator studenta (liczba naturalna jednoznacznie identyfikująca studenta)
- data_ur - miesiąc i rok urodzenia (w formacie rrrr-mm)
- plec - płeć studenta (M-mężczyzna, K-kobieta)
- cdyd_kod - rocznik, którym student uczęszczał na przedmiot (jest 11 różnych roczników, lata 1999-2009)
- prz_kod - kod przedmiotu
- nazwa - nazwa przedmiotu
- tpro_kod - sposób zaliczenia przedmiotu (egzamin lub zaliczenie)
- prot_id - identyfikator protokołu
- term_prot_id - termin, w którym został zdany/zaliczony przedmiot (przyjmuje wartości 1,2 lub 3)
- toc_kod - zmienna związana z tpro_kod (tpro-kod = EGZ => toc_kod = STD, tpro_kod = ZAL => toc_kod = ZAL, tpro_kod = ZAL-OCENA => toc_kod = STD, tpro_kod = ZAL-LUB-OCENA => toc_kod = ZAL-STD)

- wartosc - ocena z separtorem .
- opis - ocena z separtorem ,

8 plików jest pustych, pozostałe 6254 zawierają co najmniej jeden rekord (wiersz z informacjami o ocenach studenta).

Oceny studentów mają 7 poziomów (NK,2,3,3.5,4,4.5,5). Należy ocenić, jak duże powinny być próbki, aby wyniki analiz, w tym przypadku porównywalne z tworzeniem histogramów, miały sens. Spośród wielu oszacowań [But] zwróćmy uwagę na tzw. oszacowanie pierwiastkowe. Tzn. liczba przedziałów winna być mniejsza od pierwiastka z wielkości próbki, czyli w naszym przypadku wielkość próbki winna wynosić 49. Z kolei na wykładach ze statystyki [Grz10] uczono, że absolutnym minimum do liczenia takich wielkości jak średnia to 3 obserwacje. Ponieważ w pewnym sensie jeden słupek histogramu jest porównywalny z liczeniem średniej, można przyjąć, że minimalna wielkość próbki winna wynieść co najmniej 21. Dlatego też w pierwszej kolejności odrzuciłam te pliki, w których było mniej niż 21 rekordów. W rezultacie pozostały 872 pliki z przedmiotami. Jako kompromis między 49 a 21 można uznać ich średnią (czyli 35). Po odrzuceniu plików, które miały mniej niż 35 rekordów, pozostało 665 plików z przedmiotami i nimi będę się dalej zajmować w analizie. W tych plikach jest 3074 różnych studentów z 11 roczników (1999-2009).

Ostatecznie, odrzuciwszy przedmioty, kończące się zaliczeniem bez oceny, poprzez inspekcję map nadreprezentacji względem przedmiotu ALGEBRA wybrałam 25 przedmiotów, dla których mapy te wyglądały interesująco.

Kod kiru	Nazwa	Liczba studentów
1000-411ALG.txt	Algebra	353
1000-411ANA.txt	Analiza	320
0000-ANGB1E.txt	Język angielski-egzamin, poziom średnio zaawansowany	202
1000-114bRP1a.txt	Rachunek prawdopodobieństwa I (potok 1)	138
1000-134BAD.txt	Bazy danych	470
1000-135MID.txt	Matematyka w instrumentach dłużnych	346
1000-2P00PF.txt	Zaawansowane programowanie funkcyjne	129
1000-411ASD.txt	Algorytmy i struktury danych	436
1000-411EMD.txt	Elementy matematyki dyskretnej	377
1000-411MNU.txt	Metody numeryczne	323
1000-411TPR.txt	Techniki programowania	372
1000-411WDL.txt	Wstęp do logiki	384
1000-411WDM.txt	Wstęp do matematyki	408
1000-411WDP.txt	Wstęp do programowania	395
1000-412AAL.txt	Analiza algorytmów	362
1000-412BAZ.txt	Bazy danych	343
1000-412JAO.txt	Języki i automaty	292
1000-412MRJ.txt	Metody realizacji języków programowania	422
1000-412POB.txt	Programowanie obiektowe	390
1000-412RPR.txt	Rachunek prawdopodobieństwa i statystyka	348
1000-413PIS.txt	Prawne i społeczne aspekty informatyki	239
1000-413SIN.txt	Sztuczna inteligencja i systemy doradcze	260
1000-413TPR.txt	Technologia produkcji systemów informatycznych	272
1000-421PAS.txt	Laboratorium programowania	341
1000-4M00PL.txt	Programowanie w logice	348

5.3 Zastosowane metody analizy

W celu przeprowadzenia analizy stworzono grupę programów wizualizacyjnych GAD zorientowanych na w/w bazę danych.

W szczególności zaimplementowano procedury wizualizacyjne:

- `pow_oczek(x,y,wart,xtyt,ytyt)` - do rysowania wizualizacji wartości oczekiwanej dystrybuanty będącej krzywą koncentracji zmiennych x,y
- `pow_zmiennosvci(x,y,wart,xtyt,ytyt)` - do rysowania wizualizacji zmienności dystrybuanty będącej krzywą koncentracji zmiennych x,y
- `wsp_polowy(x,y,wart,xtyt,ytyt)` - do rysowania wizualizacji współczynnika polowego dystrybuanty będącej krzywą koncentracji zmiennych x,y
- `fun_kwantyl(x,y,wart,xtyt,ytyt)` - do rysowania wizualizacji funkcji kwantylowej dystrybuanty będącej krzywą koncentracji zmiennych x,y
- `krz1momentu(x,y,wart,xtyt,ytyt)` - do rysowania wizualizacji krzywej pierwszego momentu dystrybuanty będącej krzywą koncentracji zmiennych x,y
- `krzkonc(x,y,wart,xtyt,ytyt)` - do rysowania wizualizacji krzywej koncentracji zmiennych x,y

- `przyrkrzkonc(x,y,kolor,podpkolor,pozkolor)` - do rysowania wizualizacji wielu krzywych koncentracji na jednym rysunku
- `plotkwLeg(m,colors,xtyt,ytyt)` - do rysowania wizualizacji mapy nadreprezentacji
- `raport_start(nazwa_raportu,nazwa_listy)` - do generowania raportów, opisanych w następnym podrozdziale.

5.4 Wyniki i ich interpretacja

Wykresy przedstawiono w załącznikach do pracy na płycie CD w postaci raportów:

1. `raport_pow_oczek`
2. `raport_pow_zmiennosvci`
3. `raport_wsp_polowy`
4. `raport_krzyweKonc`
5. `raport_terminowa_krzkonc`
6. `raport_krz1momentu`
7. `raport_mapanadLeg`

5.4.1 Krzywa koncentracji `raport_krzyweKonc`

Krzywa koncentracji jest wyrazem wzajemnych relacji między grupami studentów dla różnych przedmiotów. W szczególności po zmieniających się kątach nachylenia krzywej widać, którą ocenę łatwiej było zdobyć.

Przykładem zastosowania analizy GAD może być analiza grup studentów realizujących ten sam przedmiot w tzw. "potokach" (np. gad.krzkonc_przed("1000-114bRP1a","1000-114bRP1b");). Jeśli założymy, że podział odbył się (z punktu widzenia zdolności studentów) w sposób losowy, to należałoby oczekiwać, że krzywa koncentracji dla dwóch potoków winna być przekątną kwadratu jednostkowego. Rzucenie okiem na stosowny wykres dla *rachunku prawdopodobieństwa* wskazuje, że nie do końca tak jest dla przedmiotu *rachunek prawdopodobieństwa I*. Jedna z grup różni się od drugiej nadmiarową liczbą studentów nieklasyfikowanych. Poza tym rzeczywiście krzywa koncentracji jest prawie prosta z wyjątkiem różnicy w liczbie ocen ponad dobrych i bardzo dobrych.

Dla *równań różniczkowych* krzywa koncentracji jest bardziej "falista", a generalnie dla ocen poniżej 5 w potoku b był większy udział tych ocen niż w potoku a, gdzie z kolei wyższa była liczba osób nieklasyfikowanych. W potoku 1 *analizy matematycznej* widzimy lekką tendencję do stawiania niższych ocen. *Geometria z algebrą liniową I* nie wykazuje specjalnych różnic między potokami.

Dla przedmiotu *Język angielski - egzamin, poziom średnio zaawansowany* ocena ndst była zdecydowanie rzadsza niż dla *algebry*, podobne zachowanie widać dla trójek. Za to pozostałe oceny stawiano częściej, w szczególności ponad dobry.

5.4.2 "Terminowa" krzywa koncentracji - `raport_terminowa_krzkonc`

Ten raport pokazuje korzyści z oglądu krzywych koncentracji dla danego przedmiotu podzbiorów danych względem całego zbioru, w tym wypadku podzbiory to zdawanie w terminie 1,2 i 3.

Dla *algebry* widzimy, że w drugim terminie rzadziej otrzymywano oceny niedostateczne niż w zbiorze ogólnym, natomiast w trzecim terminie studenci dostali same dwóje.

Podobne tendencje widzimy dla przedmiotu *analiza*, przy czym w drugim terminie widać wyraźniejszą tendencję do lepszych ocen.

Zaawansowane programowanie funkcyjne charakteryzuje najlepsza zdawalność w pierwszym terminie. W drugim podejściu obserwujemy trochę więcej dwój, ale ponad połowa osób otrzymała stopień bardzo dobry. W trzecim terminie wszyscy, którzy zdali, zaliczyli przedmiot na ocenę dostateczną.

Dla *algorytmów i struktur danych* możemy zauważyć podobne zachowanie w rozkładzie ocen pierwszego terminu w stosunku do ogólnego zbioru. Natomiast w trzecim terminie otrzymywano praktycznie same dwóje. Podobne zachowanie przedstawiają wykresy dla *elementów matematyki dyskretnej*, z tą różnicą w trzecim terminie, że poza przytłaczającą większością dwój, pojawiły się nieliczne czwórki.

Jednak nie sposób nie zauważyć, że osoby, które nie zaliczyły danego przedmiotu w pierwszym ani w drugim terminie, nie zdają również przy trzecim podejściu.

5.4.3 Wartość oczekiwana `raport_pow_oczek`

Rysunki w tym raporcie odzwierciedlają wartość oczekiwaną dystrybuanty powstałej jako krzywa koncentracji *Algebry* do pewnego przedmiotu X z listy.

Wartość oczekiwana reprezentowana jest przez pokolorowaną powierzchnię (powyżej krzywej koncentracji). Wysoka wartość oznacza, iż przedmiot X jest "łatwiejszy" od *Algebry* (gorszych studentów z X jest mniej niż z Algebry). Dotyczy to np. przedmiotu *Język angielski - egzamin, poziom średnio zaawansowany*, *Prawne i społeczne aspekty informatyki, języki i automaty, czy rachunek prawdopodobieństwa i statystyka.*

Z kolei mniejsza powierzchnia (jak np. dla *Algorytmy i struktury danych*) sugeruje, że przedmiot X jest trudniejszy od *Algebry.*

5.4.4 `raport_pow_zmiennosvci` - Zmienność

Zmienność jest wyrazem dynamiki odchyleń od średniej w stosunku do wartości oczekiwanej.

W analizowanych danych szczególnie wyróżnia się przedmiot *Zaawansowane programowanie funkcyjne* względem *Algebry*, gdyż w ich krzywej koncentracji następuje ostry przełom między grupą osób ocenianych na 2 i na 3.

Dla przedmiotu *Bazy danych* obserwujemy niską zmienność. Krzywa koncentracji jest tu silnie wypukła. Dosyć mała zmienność widoczna jest również dla *prawnych i społecznych aspektów informatyki*, przy czym dla ocen między trzy a cztery "lekko" wzrasta.

5.4.5 Współczynnik polowy `raport_wsp_polowy`

Współczynnik polowy to odpowiednik indeksu Gini. Duża jego wartość świadczy o odbieganiu rozkładu ocen jednego przedmiotu od drugiego - na niekorzyść przedmiotu na osi X.

Do ilustracji grup studentów realizujących ten sam przedmiot w tzw. "potokach" (np. gad.wsp_polowy("1000-1M05TRa","1000-1M05TRb");) można wykorzystać współczynnik polowy. Przy założeniu, że podział odbył się (z punktu widzenia zdolności studentów) w sposób losowy, oczekuje się, że współczynniki polowy dla dwóch potoków winien być jak najcieńszą przekątną kwadratu jednostkowego. Po spojrzeniu na odpowiedni wykres dla *Teorii ryzyka w ubezpieczeniach*, można zauważyć, że niezaliczenie przedmiotu (nieklasyfikowanie i dwóje) oraz ocena bardzo dobra nie zależą od tego, na który potok uczęszczał student (pole im odpowiadające jest dosyć wąskie), natomiast w potoku b, więcej osób dostało zwłaszcza oceny 3. Z kolei dla przedmiotu *Rachunek prawdopodobieństwa II*, rozkład ocen jest mniej więcej porównywalny, z tym że w potoku I jest tendencja do stawiania niższych ocen, a w potoku II mamy więcej nieklasyfikowanych osób. Wykres dla *matematyki obliczeniowej* wskazuje na to, że ciężej zdobyć zaliczenie w potoku II. Natomiast pozostałe oceny są rozłożone prawie równomiernie z "lekką" tendencją do stawiania oceny ponad dostatecznej w potoku I.

Ciekawostką może być przedmiot *topologia I*, gdzie w jednym potoku widoczna jest tendencja do stawiania niższych ocen: dwój i trój, a w drugim przeważają oceny od 3.5 w górę. Oznacza to, że potok I ciężej jest zaliczyć. Podobne zachowanie widać dla *Algebry I*, z tym, że dla potoku I przeważają dwóje i niezaliczenia, a w potoku II najwięcej jest ocen między dostateczną a dobrą. Dla przedmiotu *Wstęp do informatyki II* większość osób otrzymała oceny między dobrą a ponad dobrą dla potoku I, czyli łatwiej było zaliczyć ten potok.

5.4.6 Krzywa pierwszego momentu `raport_krz1momentu`

Krzywa pierwszego momentu odbiegająca silnie od dystrybuanty świadczy o tym, jak duża jest zmienność.

Dla przedmiotu *zaawansowane programowanie funkcyjne* krzywa pierwszego momentu jest silnie wypukła. Sugeruje to, że prawdopodobnie zmienność jest duża, podobnie jak w przypadku *elementów matematyki dyskretnej*.

Podobną tendencję wykazuje *język angielski-egzamin, poziom średnio zaawansowany*. Jednak gdy naniesiemy na jeden wykres z dystrybuantą, okaże się, że zmienność jest niewielka. Dlatego też sama krzywa pierwszego momentu niesie jedynie przesłanki o tym, jak duża jest zmienność.

5.4.7 Mapy nadreprezentacji `raport_mapanadLeg`

Mapy nadreprezentacji dotyczą zmiennych mierzonych na tej samej grupie obiektów, zatem pokazują "korelację" między nimi.

Mapa nadreprezentacji *Algebry* względem *Algebry* nie jest oczywiście informatywna, natomiast pokazuje wspomniany w pracy efekt, iż im większa jest grupa, tym z definicji ma niższy maksymalny wskaźnik nadreprezentacji.

Mapa *algebry* i *analizy* wskazuje, że na skrajnych końcach skali ocen - piątki i dwójki - są zdecydowanie nieprzypadkowe i brak umiejętności w jednej dziedzinie odzwierciedla się w drugiej. Podobnie jest z dużą wiedzą.

Ciekawostką jest być może asymetria wykresu dla przedmiotów *Algebra* kontra *Algorytmy i struktury danych*. Tu dwójka z *Algebry* jest silnie nadreprezentowana w kontekście nieklasyfikowania z *algorytmów*. Ponadto otrzymywanie ocen najwyższych (ponad dobry i bardzo dobry) jest ze sobą silnie związane - mamy dużą nadreprezntację. W analogiczny sposób wypada zestawienie *algebry* z *elementami matematyki dyskretnej, technikami programowania*,

wstępem do programowania, *analizą algorytmów*, czy też *prawnymi i społecznymi aspektami informatyki*.

Na uwagę zasługuje też mapa dla *Algebry* i *Matematyki w instrumentach dłużnych*. Widać nadreprezentację odpowiednio 4.5 z dwójkami. Analogiczną sytuację da się zauważyć dla *algebry* z *bazami danych*. Pozostałe oceny dla odpowiednich par przedmiotów zdają się nie mieć ze sobą powiązania.

Nasuwa się pytanie, czy istnieje jakieś powiązanie między tężyzną fizyczną a umysłową. A mianowicie, czy i jak wpływa zaliczenie wychowania fizycznego na zaliczenie przedmiotu ścisłego, w naszym przypadku *algebry*. Wyniki okazują się zaskakujące, bowiem w ogólności osoby nieklasyfikowane z *algebry* zaliczają wychowanie fizyczne. Tak jest dla np. *aerobiku*, *tenisa ziemnego*, *karate*, *jogi*, *badmintona*, czy *pływania*. Z kolei osoby, które chodzą na *algebrę*, nie przejmują się zaliczeniem zajęć sportowych.

Co ciekawe, w przypadku *algebra* kontra *taniec towarzyski*, osoby otrzymujące z *algebry* ocenę dostateczną i dobrą, przejmują się zaliczeniem sportu. Natomiast zaliczenie *tańca* i zaliczenie *algebry* na 4.5 nie są ze sobą związane. Kolejny wyjątek stanowi *krav-maga (samoobrona)*. Tutaj studenci, którzy otrzymali stopnie ponad dostateczne i bardzo dobre, przejmują się zaliczeniem wf-u. A niezaliczenie *krav-magi* nie zależy od zaliczenia *algebry* na 3.5.

Warto byłoby zastanowić się nad wpływem uczęszczania na przedmioty humanistyczne, wyniki osiągane w przedmiotach ścisłych. Czy pomagają w zdawalności, czy też ją utrudniają. Widać podobną zależność, co w przypadku zaliczenia sportu; osoby zaliczające przedmioty humanistyczne są nieklasyfikowane z *algebry*. Takie zachowanie widać dla np. *narzędzi i metod przetwarzania tekstu*, *protokołów komunikacyjnych*, *algorytmiki*, *matematyki przy klawiaturze*, *akademii filmowej*, *varsavianistyki* czy *praktyk pedagogicznych*.

Interesujący może być silny związek między niezaliczeniem *systemów rozproszonych* a wynikiem ponad dobrym z *algebry*. Co więcej, ani studenci czwórkowi,ani piątkowi, nie przejmują się zaliczeniem z *systemów*. Na uwagę zasługuje również fakt, że piątka z *algebry* jest niezależna z niezaliczeniem *zagadnień programowania obiektowego*, a już 4.5 z *algebry* i niezaliczenie *zagadnień* są silnie powiązane.

Rozdział 6

Podsumowanie

W niniejszej pracy przedstawiono podstawowe koncepcje gradacyjnej analizy danych (GAD). Opisano ich relacje z wybranymi pojęciami klasycznej statystyki.

Skupiono się na tych jej składowych, które są istotne z punktu widzenia analizy danych z systemu USOS, będących przedmiotem prac eksperymentalnych, tzn. analizie par zmiennych o takich samych zbiorach wartości w niezależnych próbkach oraz analizie par zmiennych dla jednej próbki. Ograniczono się do zmiennych porządkowych dyskretnych. Dlatego pominięto np. analizę skupień GAD [Cio02, Cio00, KNB02] która w swej istocie zmierza do wprowadzenia najkorzystniejszego porządku dla zmiennych nominalnych.

W pracy przestawiono definicje pojęć GAD, ich interpretację oraz wybrane własności, opierając się na literaturze przedmiotu, uzupełniając je własnymi przemyśleniami.

W szczególności w niniejszej pracy pokazano, że indeks Giniego-Simpsona mieści się w zakresie [0,1] (tw. 2.2.1). Sformułowano i naszkicowano dowód twierdzenia o przekodowaniu dziedzin zmiennych dla krzywej koncentracji (tw. 3.1.1). Sformułowano i przedstawiono dowód twierdzenia o wartości minimalnej (tw.4.3.2) oraz maksymalnej (tw.4.3.3) funkcji nadreprezentacji, a dla zmiennych niezależnych o jej wartości oczekiwanej (tw. 4.3.1).

Te trzy ostatnie twierdzenia mają znaczenie przy pisaniu programu wizualizującego mapę nadreprezentacji. Tw. 3.1.1 rozstrzyga o celowości transformacji dziedzin zmiennych. Natomiast tw. 2.2.1 uwypukla związek GS z z GAD i tradycyjnego indeksu GS.

Ponadto drobnym wkładem jest uporządkowanie definicji krzywej koncentracji (def.3.1.1) oraz powierzchni koncentracji (def.4.2.1).

Analiza gradacyjna nie jest radykalnie nowym kierunkiem eksploracyjnej analizy danych, lecz raczej próbą uogólnień znanych koncepcji. Jej podstawowym walorem jest jednak nacisk na wizualną interpretację pojęć, zwyczajowo reprezentowanych przez podsumowujące liczby, takich jak wartość oczekiwana, zmienność, współczynnik polowy (analog indeksu Giniego) itp. Ponadto wgląd w dane umożliwia krzywa koncentracji (analog krzywej Lorenza) oraz mapa nadreprezentacji.

Aby przybliżyć tę wartość dodaną, wykreślono te reprezentacje graficzne dla wybranych wyników egzaminów dla próbki danych z systemu USOS. Ilustracje zawierają załączniki, opis zaś zawarty jest w rozdziale 5. Z wykresów można m.in. odczytać, jakie przedmioty łatwiej zaliczyć, czy też który z prowadzących ten sam przedmiot jest surowszy w ocenianiu studentów, jak uczą się poszczególnych przedmiotów studentki w porównaniu ze studentami, jaka jest łatwość zaliczania w kolejnych terminach egzaminów, na ile dobre oceny z jednych przedmiotów zależą od innych, czy uprawianie sportu wpływa na uzyskiwane wyniki nauczania.

W celu realizacji eksperymentów wykorzystujących analizę gradacyjną do eksploracji danych nt. wyników egzaminów zaimplementowano wybrane metody gradacyjne w języku

programowania R. M.in. można za ich pomocą uzyskać obszarową reprezentację wartości oczekiwanej, zmienności, współczynnika polowego, indeksu Giniego-Simpsona, a także krzywe koncentracji i mapy nadreprezentacji.

Podsumowując, wydaje się, że w fazie eksploracji danych gradacyjna analiza danych może ułatwić dostrzeżenie pewnych tendencji poprzez dotarcie do obrazowej wyobraźni analityka i przez to stymulować jego prace badawcze.

Bibliografia

[But] Tony Butterfield. Statistical data visualization, chemical engineering, university of utah.

[Cio00] A. Ciok. Double versus optimal grade clusterings. In Groenen P.J.F. Schader M. Kiers H.A.L., Rasson J.-P., editor, *Data Analysis, Classification, and Related Methods*, pages 41–46. Springer-Verlag, 2000.

[Cio02] A. Ciok. Grade correspondence-cluster analysis applied to separate components of reversely regular mixtures. In Jajuga K., Sokołowski A., Bock H.-H., editors, *Classification, Clustering and Data Analysis, Recent Advances and Applications*, pages 211–218. Springer-Verlag, 2002.

[Cio05] A. Ciok. Grade analysis of repeated multivariate measurements. In P. Grzegorzewski, O. Hryniewicz, M.A. Gil, editors, *Soft Methods in Probability, Statistics and Data Analysis*, pages 266–273. Physica-Verlag, May 2005.

[FM00] Luisa T. Fernholz, Stephan Morgenthaler. A conversation with John W. Tukey and Elizabeth Tukey. *Statist. Sci.*, 15(1):79–94, 2000.

[Grz10] P. Grzegorzewski. Wykłady ze statystyki i, wydział matematyki i nauk informacyjnych, rok akademicki 2009/2010., 2010.

[JJvEC02] Brower JE, Zar JH, von Ende CN. *Field and laboratory techniques for general ecology.* William. C. Brown Publishers, 2002.

[Kle05] Christian Kleiber. The lorenz curve in economics and econometrics. In *Gini-Lorenz Centennial Conference, Siena, May 23–26, 2005.* May 2005.

[KMPW05] J. Książyk, O. Matyja, E. Pleszczyńska, M. Wiech. *Analiza danych medycznych i demograficznych przy użyciu programu GradeStat.* Intytut Podstaw Informatyki PAN, Instytut "Pomnik - Centrum Zdrowia Dziecka", Warszawa, 2005.

[KNB02] T. Kowalczyk, M. Niewiadomska-Bugaj. A new grade measure of monotone multivariate separability. In C. Cuadras, Fortiana J., J. Rodriquez-Lallena, editors, *Distributions with Given Marginals and Statistical Modelling*, pages 143–151. Kluwer Academic Publishers, 2002.

[KPR04] T. Kowalczyk, E. Pleszczyńska, F. (eds) Ruland. *Grade Models and Methods for Data Analysis: With Applications for the Analysis of Data Populations*, volume 151 of *Studies in Fuzziness and Soft Computing.* Springer-Verlag, 2004.

[Tuk77] J. W. Tukey. *Exploratory Data Analysis.* Addison-Wesley, 1977.

Indeks rzeczowy

Wykaz rysunków

Warszawa, dnia .....

Oświadczenie

Oświadczam, że pracę magisterską pod tytułem: "*Analiza gradacyjna – teoria i zastosowania*" , której promotorem jest *dr inż. Przemysław Biecek* wykonałam samodzielnie, co poświadczam własnoręcznym podpisem.

.....
Elżbieta Anna Kłopotek