**INSTYTUT BADAŃ SYSTEMOWYCH**
**POLSKIEJ AKADEMII NAUK**

# TECHNIKI INFORMACYJNE TEORIA I ZASTOSOWANIA

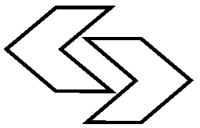
Wybrane problemy
Tom 1(13)

*poprzednio*

**ANALIZA SYSTEMOWA W FINANSACH I ZARZĄDZANIU**

Pod redakcją
Jerzego HOŁUBCA

Warszawa 2011



**INSTYTUT BADAŃ SYSTEMOWYCH**
**POLSKIEJ AKADEMII NAUK**

# TECHNIKI INFORMACYJNE TEORIA I ZASTOSOWANIA

Wybrane problemy
Tom 1(13)

*poprzednio*

**ANALIZA SYSTEMOWA W FINANSACH I ZARZĄDZANIU**

Pod redakcją
Jerzego HOŁUBCA

Warszawa 2011

Wykaz opiniodawców artykułów zamieszczonych
w niniejszym tomie:

Dr hab. inż. Przemysław GRZEGORZEWSKI, prof. PAN

Prof. dr hab. inż. Jerzy HOŁUBIEC

Dr inż. Tatiana JAWORSKA

Dr hab. inż. Wiesław KRAJEWSKI, prof. PAN

Dr hab. inż. Maciej KRAWCZAK, prof. PAN

Dr hab. Michał MAJSTEREK

Dr hab. inż. Andrzej MYŚLIŃSKI, prof. PAN

Prof. dr hab. inż. Witold PEDRYCZ

Dr hab. inż. Ryszard SMARZEWSKI, prof. KUL

Prof. dr hab. inż. Andrzej STRASZAK

Dr Dominik ŚLĘZAK

Prof. dr hab. inż. Stanisław WALUKIEWICZ



**ISBN 9788389475336**

# GRADACYJNE METODY ANALIZY DANYCH NA PRZYKŁADZIE BADANIA SOLIDNOŚCI KIENTÓW BANKU

***Małgorzata Ostrycharz***
*Studia Doktoranckie IBS PAN*

*Gradacyjna analiza danych (ang. Grade Data Analysis, GDA) to nowy kierunek z dziedziny eksploracyjnej analizy danych wielowymiarowych. Gradacyjne podejście do analizy danych bazuje na pomiarze nierównomierności i koncentracji – zamiast analizy rzeczywistych rozkładów zmiennych losowych bada się rozkład ich skoncentrowania, a następnie odpowiednio porządkuje. Siła GDA tkwi w możliwości wykrywania podobieństw i różnic równocześnie pomiędzy wieloma zmiennymi oraz przystępnej prezentacji graficznej owych zależności. W pracy zaprezentowano ideę GDA stosując metodę do oceny wiarygodności kredytowej klientów banku.*

*Grade Methods for Data Analysis (GDA) is a new approach to Exploratory Data Analysis. Grade approach to data analysis is based on the measure of inequality and concentration – analysis of many random variables' distributions is replaced by studying concentration distributions, which are then ordered. The potential of GDA lies behind the capacity of investigate similarities and differencies between many variables simultaneously as well as the intelligible graphical presentation. The paper shows the way of modeling credit worthiness involving GDA.*

***Słowa kluczowe**: gradacyjna analiza danych, krzywa koncentracji, krzywa Lorenza, indeks Giniego, kopuła, mapa nadreprezentacji, korelacja, korelacja gradacyjna, gradacyjna funkcja regresji*

## 1. Wstęp

Dokonując badania i opisu otaczających nas zjawisk niejednokrotnie stajemy przed koniecznością uwzględnienia wielu cech badanego obiektu. Owa potrzeba uogólnienia metod statystyki jednej zmiennej na przypadek większej liczby zmiennych spowodowała i dalej powoduje rozwój różnorodnych metod wielowymiarowej analizy danych. Tradycyjne podejście modelowe zakłada

istnienie pewnego matematycznego modelu, który opisuje dane zjawisko upraszczając rzeczywistość. Istotną niedogodnością jest tu konieczność spełnienia pewnych założeń leżących u podstaw budowy modelu. Pojawia się pytanie, jak należy postępować w sytuacji niespełnienia założeń przyjętych przy konstrukcji modelu, gdyż pominięcie weryfikacji założeń może prowadzić do błędnych wniosków. Próbą ominięcia niedogodności związanych z modelowaniem i weryfikacją założeń jest eksploracja danych, propagująca narzędzia oparte na prostych metodach obliczeniowych i graficznych. Swego rodzaju modyfikacją eksploracji danych jest data mining – metody skierowane przede wszystkim na praktyczne zastosowania - poszukiwanie użytecznych rozwiązań - często bez wnikania w istotę zjawiska.

Gradacyjna analiza danych (GDA), zapoczątkowana w Instytucie Podstaw Informatyki Polskiej Akademii Nauk przez zespół prof. Elżbiety Pleszczyńskiej, może być postrzegana zarówno jako metoda eksploracji danych, jak i jako technika data mining. Istotą metody jest odpowiedni podział danych na grupy o zbliżonych cechach poprzez **rangowanie**, czyli wprowadzenie porządku według konkretnej cechy zbiorczej. Dzięki temu zabiegowi dane stają się bardziej przejrzyste, a struktury danych bardziej widoczne. Obecnie analiza oparta na metodach gradacyjnych wykorzystuje pomiar koncentracji rozkładów do analizy skupień, analizy odpowiedniości oraz analizy regularności rozkładu cechy w populacji. Algorytm szeregujący dane pod względem stopnia podobieństwa odzwierciedla pewne probabilistyczne własności rozkładów. Dlatego GDA jest nie tylko kombinatoryczną techniką zgłębiania zbiorów danych, ale zespołem metod posiadających solidne fundamenty matematyczne (por. [7], [16]).

W artykule zajmiemy się badaniem solidności klientów banku - ten prosty przykład pozwoli w naturalny sposób wprowadzić formalizm GDA oraz wskazać obszary potencjalnych zastosowań.

## 2. Ocena wiarygodności kredytowej klienta banku

Wyobraźmy sobie sytuację oceny klientów przez bank w procesie przyznawania produktów kredytowych. Konkretniej, mamy daną pewną grupę osób oraz szereg cech. Cechy tworzą przestrzeń atrybutów, przykładowo o liczności $k$, natomiast respondenci - przestrzeń $n$ obserwacji. W kontekście matematycznym mamy do czynienia z przestrzenią $n$ obserwacji, z których każda jest wektorem $k$ atrybutów, każdy będący zmienną losową o pewnym rozkładzie.

Oczywistym jest, że informacją najcenniejszą dla banku jest zarówno zdolność kredytowa klienta - wyznaczana w oparciu o dochody i wydatki klienta - jak i jego „wiarygodność" kredytowa. Dokładniej mówiąc, **punktowa ocena wiarygodności kredytowej klienta**, jest to estymator przyszłej intencji klienta do spłaty wnioskowanego kredytu. Ocenia się wiarygodność kredytową klienta pod kątem prawdopodobieństwa wystąpienia zdarzenia niewykonania zobowiązań w horyzoncie 1 roku - określane jako tzw. ***zdarzenie default***. Dzięki określeniu takiej szansy, bank może w łatwy sposób identyfikować osoby potencjalnie niezdolne do spłaty i w konsekwencji odmówić im udzielenia produktu finansowego. Badanie wiarygodności kredytowej może być instrumentem kształtowania indywidualnych decyzji kredytowych, wsparciem procesu podejmowania decyzji kredytowych. Badanie może pomóc w określeniu, którym osobom (o jakich cechach) opłaca się zaproponować oferty według specjalnych algorytmów, na korzystniejszych warunkach itd. Pozwala na określenie grupy docelowej, do której skierowanie konkretnej informacji przyniesie z dużym prawdopodobieństwem spodziewany rezultat finansowy. Sensowność przeprowadzenia takiego badania wydaje się oczywista.

Podstawą stworzenia dobrego modelu oceny wiarygodności jest identyfikacja najbardziej istotnych czynników i ich odpowiednich kombinacji (wag) dla zapewnienia maksymalnej zdolności dyskryminacyjnej, tzn. zdolności różnicowania „dobrych"' i „złych" klientów. Systemy dyskryminacyjne tego typu mogą być (i z reguły są) oparte na wyznaczaniu pewnej punktacji klientów, która w monotoniczny sposób określa ich wiarygodność kredytową.

## 3. Przykład ilustrujący ideę GDA

Załóżmy, że mamy dostatecznie dużo obserwacji, z których każda określa jednego klienta indywidualnego, a dokładniej kilka jego cech, które opisane są za pomocą poniższych zmiennych:

- *historia konta*,
- *wiek*,
- *miesięczne obciążenie kredytem*,
- *default,* czyli zmienna binarna identyfikująca zajście zdarzenia nie wywiązania się ze spłaty w ciągu roku od uruchomienia kredytu (zdarzenie default). Zdarzenie to może dotyczyć nie tylko samego faktu niespłacenia kredytu, ale również długotrwałego przeterminowania

w spłatach (jest to określone osobną definicją regulowaną przez odpowiednie uchwały i ustalenia).

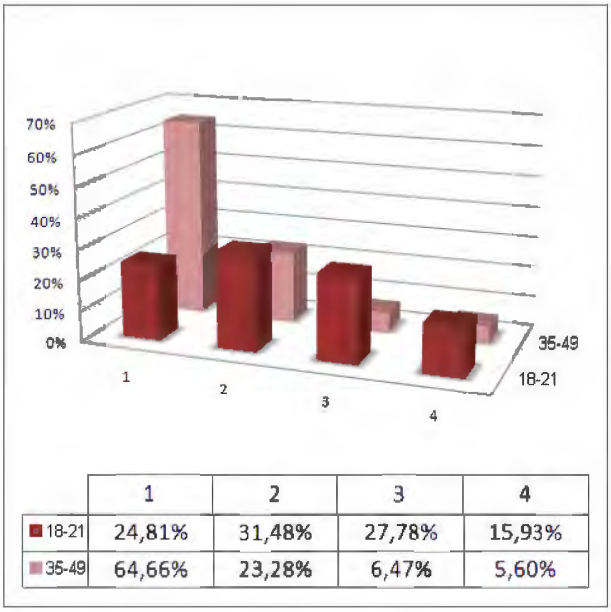
- *solidność*, czyli zmienna nominalna określającą jak często w ciągu roku zachodziło zdarzenie default i czy było to zdarzenie o dużej wadze (typu duże opóźnienie w spłacie lub/i duża kwota przeterminowana) czy małej (małe opóźnienie w spłacie lub mała kwota przeterminowana).

W procesie oceny spłacalności kluczowa jest odpowiedź na pytanie, jakie cechy posiada klient „dobry", a jakie klient „gorszy". Dobrym rozwiązaniem problemu wydaje się być odpowiednie uszeregowanie wartości cech determinujące cechę objaśnianą, jaką jest prawdopodobieństwo spłaty. Przykładowo, mając dane dotyczące wieku i historii konta klienta - dokładniej liczby lat istnienia konta osobistego - poszukujemy takiego porządku wśród wartości atrybutów, takiej „kolejki" wartości danej cechy (lub wielu cech jednocześnie), która powoduje, że osoby o wartościach z jej początku cechują się wysoką wiarygodnością kredytową, a osoby z jej końca - niską. Po takim uszeregowaniu następuje odrzucenie klientów, będących poniżej pewnego progu odcięcia - czyli pewnej wartości cechy w omawianej „kolejce" - i uznanie ich za klientów o niedopuszczalnie niskiej wiarygodności. Wybór progu odcięcia jest już kwestią optymalizacyjną pod względem zysków i strat, jakie bank ponosi z tytułu akceptacji pewnej grupy klientów, zawierającej określoną frakcję osób „niesolidnych".

Załóżmy, że zebrano dane dotyczące pewnej grupy klientów, następnie zaczęto analizować relacje cech klientów oraz cechy *solidność*. Przyjrzyjmy się charakterystyce *wiek*. Dla zmiennej nominalnej *solidność* wyróżniono cztery kategorie, począwszy od klienta bardzo solidnego (etykieta 1) skończywszy na niesolidnym (etykieta 4) w pięciu grupach wiekowych (wiek mierzony w latach): *18 -21*, *22 -25*, *26 -34*, *35 -49* oraz *50* (50 lat i więcej) (patrz tab. 1).

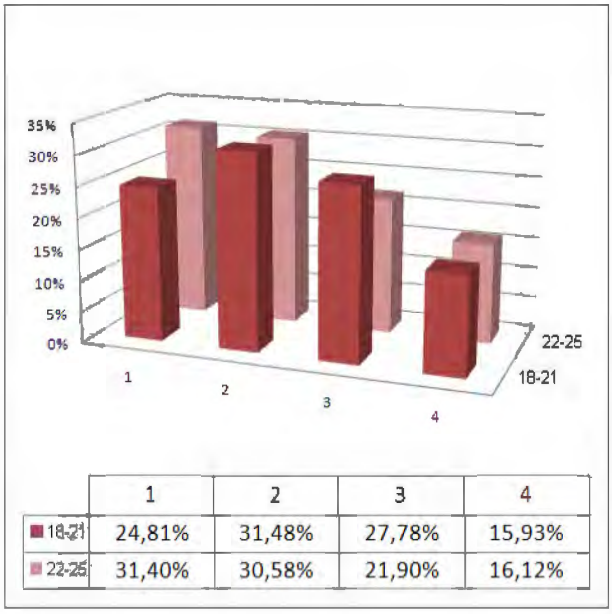
Tabela 1 - Tablica kontyngencji dla cechy *solidność* w różnych grupach wiekowych.

| Solidność (etykieta) | *18-21* | *22-25* | *26-34* | *35-49* | *50+* | suma |
|---|---|---|---|---|---|---|
| Bardzo solidny (1) | 67 | 76 | 103 | 150 | 87 | 483 |
| Średnio solidny (2) | 85 | 74 | 34 | 54 | 61 | 308 |
| Mało solidny (3) | 75 | 53 | 32 | 15 | 38 | 213 |
| Niesolidny (4) | 43 | 39 | 17 | 13 | 21 | 133 |
| suma | 270 | 242 | 186 | 232 | 207 | 1137 |




| | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 18-21 | 24,81% | 31,48% | 27,78% | 15,93% |
| 35-49 | 64,66% | 23,28% | 6,47% | 5,60% |

Rysunek 1. Rozkład cechy *solidność* w grupach wiekowych *18-21* i *35-49*.




| | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 18-21 | 24,81% | 31,48% | 27,78% | 15,93% |
| 22-25 | 31,40% | 30,58% | 21,90% | 16,12% |

Rysunek 2. Rozkład cechy *solidność* w grupach wiekowych *18-21* i *22-25*.

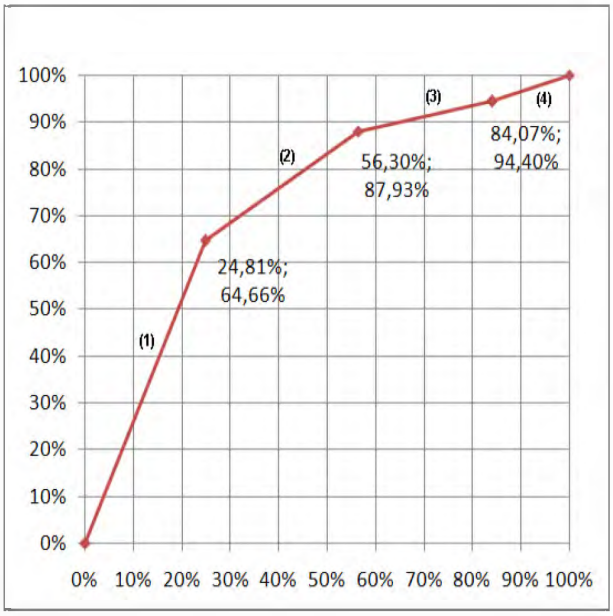
W tabeli 1 wyznaczono dodatkowo sumy w wierszach i kolumnach. Przykładowo, najmniej liczną grupę stanowią klienci niesolidni, jest ich 133 na 1137 wszystkich klientów, co stanowi blisko 12% badanej grupy. Jednak w klasie klientów młodych, poniżej 22 lat, klienci niesolidni stanowią już prawie 16%, natomiast w grupie wiekowej *35-49* jedynie 5,6%. Istnieje podejrzenie, że solidność klientów jest w pewnym stopniu zależna od ich wieku. Można zatem przyjrzeć się liczbie klientów solidnych w różnych grupach wiekowych oraz zaobserwować, na ile rozkłady te różnią się między sobą (por. rys. 1 oraz rys. 2).

Badamy zróżnicowanie rozkładu solidności pod względem wieku. Wydzielimy zatem pięć zmiennych losowych odpowiadających solidności w poszczególnych kategoriach wiekowych: $S_{18\text{-}21}$, $S_{22\text{-}25}$, $S_{26\text{-}34}$, $S_{35\text{-}49}$, $S_{50+}$. Będziemy mówić o kategoriach zmiennej *wiek* lub o zmiennych losowych, odpowiadających określonemu przedziałowi wiekowemu. Zmienną *wiek* nazywać będziemy dalej **zmienną grupującą** lub **wymiarem**, natomiast zmienną *solidność* - **cechą badaną**. Na rysunkach 1 oraz 2 przedstawiono graficznie dwie pary rozkładów. Zarówno rozkłady $S_{18\text{-}21}$ i $S_{22\text{-}25}$, oraz $S_{18\text{-}21}$ i $S_{35\text{-}49}$, różnią się w widoczny sposób. Postaramy się ujawnić koncentrację dwóch zmiennych losowych za pomocą pewnej krzywej. Na jednym wykresie umieścimy punkty $(u_i,v_i)$, gdzie $u_i$ będzie wartością dystrybuanty pierwszej zmiennej losowej, natomiast $v_i$ - wartością dystrybuanty drugiej zmiennej losowej, dla kategorii $i$, przy pewnym początkowym uszeregowaniu kategorii. Sąsiednie punkty połączymy odcinkami, przy czym punkt pierwszy połączymy z punktem $(0,0)$. Taką krzywą narysujemy dla dwóch par rozkładów:

- dla rozkładu zmiennej losowej $S_{18\text{-}21}$ oraz $S_{35\text{-}49}$ (por. rys. 3),
- dla rozkładu zmiennej losowej $S_{18\text{-}21}$ oraz $S_{22\text{-}25}$ (por. rys. 4).

Krzywa z rysunków 3 i 4 nazywana jest ***krzywą koncentracji*** dwóch rozkładów i składa się z punktów oraz odcinków łączących te punkty. Wykres z rysunku 4 jest zbliżony do przekątnej kwadratu jednostkowego - odpowiednie wartości dystrybuant są podobne, drugi wykres (rysunek 3) jest bardziej odchylony od przekątnej kwadratu jednostkowego - oznacza to większe zróżnicowanie rozkładów prezentowanych na tym wykresie. Kształt krzywej koncentracji jest zatem wyznacznikiem zróżnicowania dwóch rozkładów. Różnice między rozkładami solidności w grupie *18-21* a pozostałymi grupami przedstawiono na rysunku 5. Im krzywa bardziej oddalona od prostej $y = x$ tym większe jest zróżnicowanie rozkładów, które dana krzywa reprezentuje.

Najbardziej zróżnicowane są rozkłady $S_{18\text{-}21}$ i $S_{35\text{-}49}$ oraz $S_{18\text{-}21}$ i $S_{26\text{-}34}$ - widać w tych grupach wiekowych struktura osób solidnych i niesolidnych prezentuje znaczne odstępstwa.



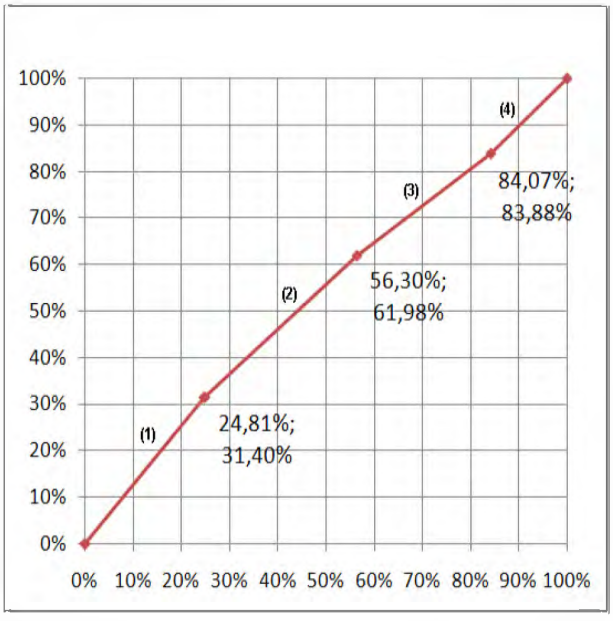

Rysunek 3 - Krzywa koncentracji dla zmiennych $S_{18\text{-}21}$ oraz $S_{35\text{-}49}$.

Zauważmy, że jeśli krzywa koncentracji jest zbliżona do przekątnej kwadratu o wierzchołkach (0,0), (1,0), (0,1), (1,1), to rozkłady cech prezentowanych przez tę krzywą są podobne - można uznać, że nie różnią się między sobą lub różnią się nieznacznie. Jeśli natomiast krzywa koncentracji kształtem odbiega od prostej można uznać, że rozkłady różnią się między sobą.

Na rysunku 5 żadne dwie krzywe nie przecinają się, ani też nie przecinają prostej $y=x$ , jednak taka sytuacja nie jest powszechna w ogólności. Zarówno nieprzecinanie się krzywych jak i wypukłość lub wklęsłość krzywych nie jest gwarantowana. Można to jednak zawsze osiągnąć za pomocą odpowiedniego „przegrupowania" odcinków, z których składa się każda z krzywych. Oznacza to przegrupowanie kategorii dotyczącej solidności klientów. Spójrzmy na rysunek 5 - jeśli, przykładowo, dla krzywej odpowiadającej kategorii *35-49* zamieniono by odcinek odpowiadający wartości cechy *solidność* wynoszącej 3 z ostatnim (*solidność* wynosząca 4) otrzymano by krzywą wklęsłą. Uporządkowana krzywa dla rozkładu zmiennej $S_{18\text{-}21}$ względem zmiennej $S_{35\text{-}49}$

widoczna jest na rysunku 6 - ze względu na zamianę zmiennych względem osi układu otrzymano krzywą wypukłą.

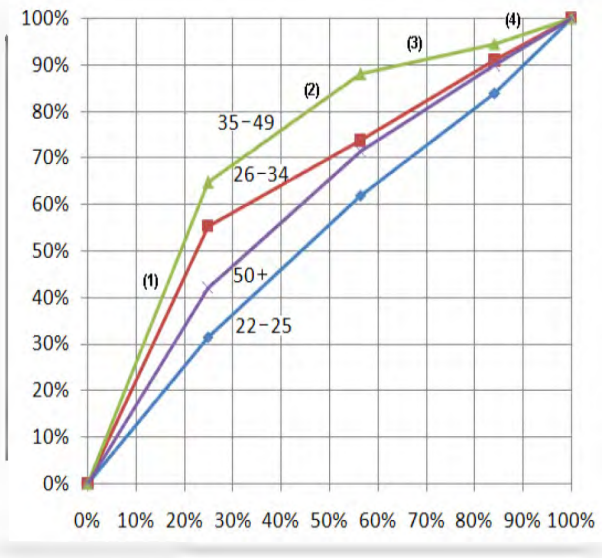
Wspomniane przegrupowanie służy wprowadzeniu takiego porządku wśród kategorii cechy (przyjętej u nas jako *solidność)*, dla którego frakcje w grupach (zmiennych grupujących) są monotoniczne. Otrzymane uporządkowanie kategorii cechy najbardziej różnicuje wymiary pod względem częstości kategorii. Przykładowo dla wymiarów $S_{18\text{-}21}$ oraz $S_{35\text{-}49}$ optymalnym porządkiem *solidności* jest ciąg: 1, 2, 4, 3. Oznacza to, że dla omawianych grup wiekowych różnica w reprezentacji solidności o etykiecie 1 lub 3 jest największa - wartości te są wartościami skrajnymi ciągu. Osób bardzo solidnych (*solidność* = 1) wśród osób w średnim wieku (*35-49*) jest 64,66%, natomiast niesolidnych (*solidność* = 4) jedynie 5,6%. W grupie wiekowej *18-21* frakcje te są już inne: odsetek osób bardzo solidnych wynosi 24,81%, natomiast osób niesolidnych - 15,93%. Jeśli jako $a^{i}_{X/Y}$ oznaczymy iloraz odsetka reprezentantów cechy X i tego samego odsetka dla zmiennej Y w grupie *i*, to przyjmując za X zmienną $S_{35\text{-}49}$, a za Y zmienną $S_{18\text{-}21}$, otrzymujemy ilorazy liczb reprezentantów cechy *solidność* w grupach *18-21* oraz *35-49* (patrz tab. 2).




Rysunek 4 - Krzywa koncentracji dla zmiennych $S_{18\text{-}21}$ oraz $S_{22\text{-}25}$.

Tabela 2 - Odsetki dla cechy *solidność* w różnych grupach wiekowych.

| | | Bardzo solidny (1) | Średnio solidny (2) | Mało solidny (3) | Niesolidny (4) |
|---|---|---|---|---|---|
| X | Odsetek w grupie *35-49* | 64,66% | 23,28% | 6,47% | 5,60% |
| Y | Odsetek w grupie *18-21* | 24,81% | 31,48% | 27,78% | 15,93% |
| $a^i_{X/Y}$ | (Odsetek w grupie *35-49*)/(Odsetek w grupie *18-21*) | 2,61 | 0,74 | 0,23 | 0,35 |

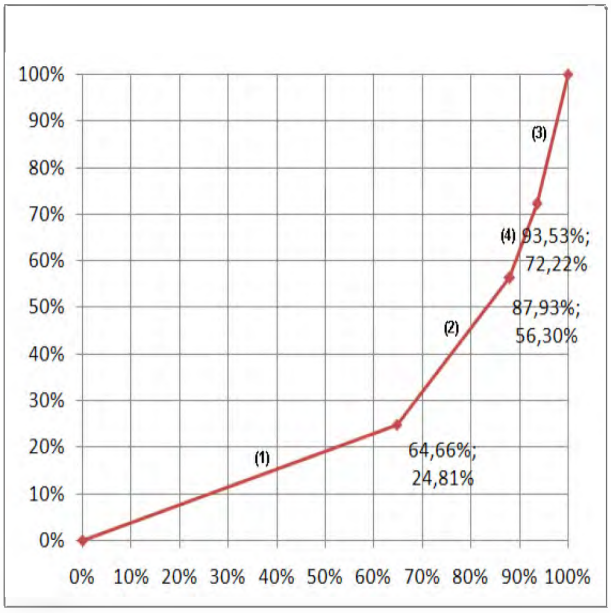



Rysunek 5 - Krzywa koncentracji dla wszystkich grup wiekowych. Za poziom referencyjny przyjęto grupę *18-21* (oś pozioma).

Optymalny porządek uzyskiwany poprzez sortowanie wskaźników $a^i_{X/Y}$ (np. malejąco), określany jest jako **porządek gradacyjny**. Jego wprowadzenie pozwala na ocenę zróżnicowania cechy **wewnątrz kategorii** oraz **pomiędzy kategoriami**. Odsetek reprezentantów grupy *35-49* najbardziej przewyższa odsetek reprezentantów grupy *18-21* w klasie klientów najbardziej solidnych (iloraz $a^i_{X/Y}$ jest największy). Interpretacja dla klientów mało solidnych

(*solidność* = 3) jest odwrotna - w tej grupie to reprezentacja grupy wiekowej *18-21* przewyższa swym odsetkiem reprezentację grupy *35-49*, w dodatku bardziej, niż w klasie klientów niesolidnych (*solidność* = 4). Kategorie cechy 1 oraz 3 są wartościami skrajnymi uporządkowanego ciągu. W przypadku oceny zróżnicowania cechy pomiędzy kategoriami, ilorazy odsetków reprezentantów grupy *35-49* oraz *18-21* najbardziej różnią się pomiędzy grupą klientów bardzo solidnych i mało solidnych.

Zabiegiem, pozwalającym na ocenę różnorodności całego zbioru byłoby idealne przegrupowanie każdej cechy, czyli uzyskanie takiego porządku względem kategorii (u nas stopnia solidności), aby krzywa koncentracji dla każdych dwóch rozkładów była wypukła lub wklęsła. Jednak takie idealne przegrupowanie pod względem jednej charakterystyki nie musi się okazać




Rysunek 6 - Uporządkowana krzywa koncentracji rozkładu zmiennej $S_{18\text{-}21}$ (oś pionowa) względem zmiennej $S_{35\text{-}49}$ (oś pozioma).

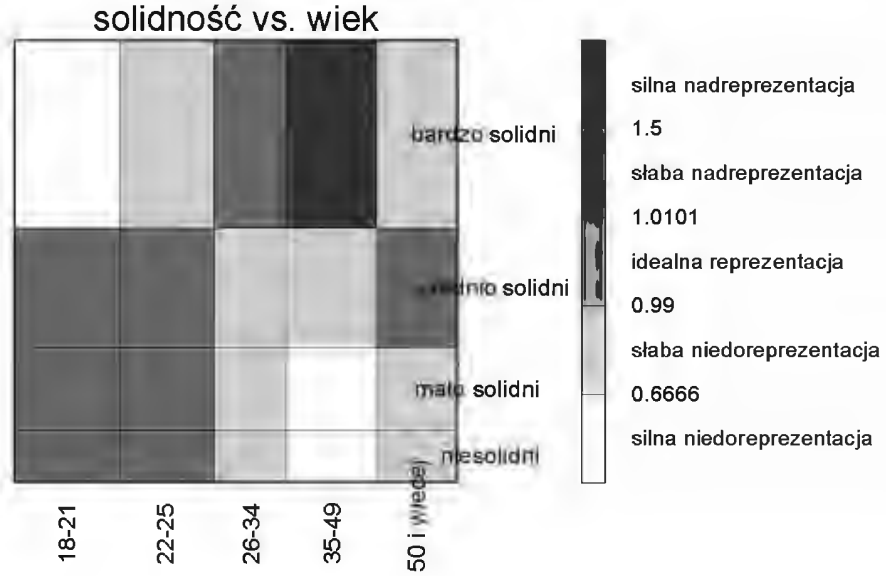
idealnym pod względem drugiej - może się nawet okazać, że druga cecha była bardziej uporządkowana przed wprowadzeniem owego przegrupowania. Przegrupowanie dotyczy również porządku samych charakterystyk, kolejności, przy której charakterystyki najbardziej od siebie oddalone jednocześnie najbardziej się różnią. Rodzi się pytanie, jak mierzyć stopień różnorodności

w całym zbiorze? Optymalnym rozwiązaniem mogłoby być takie pogrupowanie, które porządkuje idealnie możliwie jak najwięcej krzywych koncentracji, nie psując przy tym bardzo porządku pozostałych krzywych. Jednak czy najlepszym rozwiązaniem będzie wprowadzenie gradacyjnego porządku dla jak największej liczby wymiarów? Czy może lepszym rozwiązaniem okaże się uporządkowanie gradacyjne mniejszej liczby zmiennych, ale kosztem mniejszego „popsucia" porządku pozostałych zmiennych? Należałoby tak postawione zagadnienie sprowadzić do zadania optymalizacji - minimalizacji lub maksymalizacji pewnej wielkości, określającej stopień uporządkowania całego zbioru. W gradacyjnej analizie danych ową wielkością jest współczynnik $\varrho^*$ **Spearmana** lub $\tau$ **Kendalla** będący miarą uwzględniającą różnorodność wszystkich par zmiennych grupujących. Dla każdej pary - czyli jednej krzywej koncentracji - miarą pozwalającą na ocenę różnorodności i optymalności permutacji jest współczynnik ***AR***, który jest związany z krzywą koncentracji (por. [7]).

Proces poszukiwania optymalnej permutacji kategorii odbywa się według ***algorytmu GCA*** (ang. Grade Correspondence Analysis), maksymalizującego współczynniki - $\varrho^*$ lub $\tau$ (por. [5]). Optymalny porządek w badanym zbiorze klientów został odnaleziony za pomocą algorytmu GCA zaimplementowanego w programie GradeStat (por. [9]).

Wynik działania algorytmu - optymalne uporządkowanie zmiennych grupujących oraz kategorii cechy *solidność* są w gradacyjnej analizie danych prezentowane w postaci tzw. ***mapy nadreprezentacji***, czyli tablicy dwuwymiarowej o różnych odcieniach szarości (lub kolorach) w komórkach (por. [7]). Odcienie odpowiadają wielkości tzw. **reprezentacji** w danej komórce. O silnej **nadreprezentacji** mówimy jeśli dana komórka zawiera więcej obserwacji, niż miałoby to miejsce, w przypadku, gdyby badane cechy były od siebie niezależne. Podobnie definiuje się silną **niedoreprezentację** - wtedy w danej komórce jest dużo mniej obserwacji, niż powinno być, przy założeniu niezależności cech. Mapa nadreprezentacji dla zbioru klientów została przedstawiona na rysunku 7.

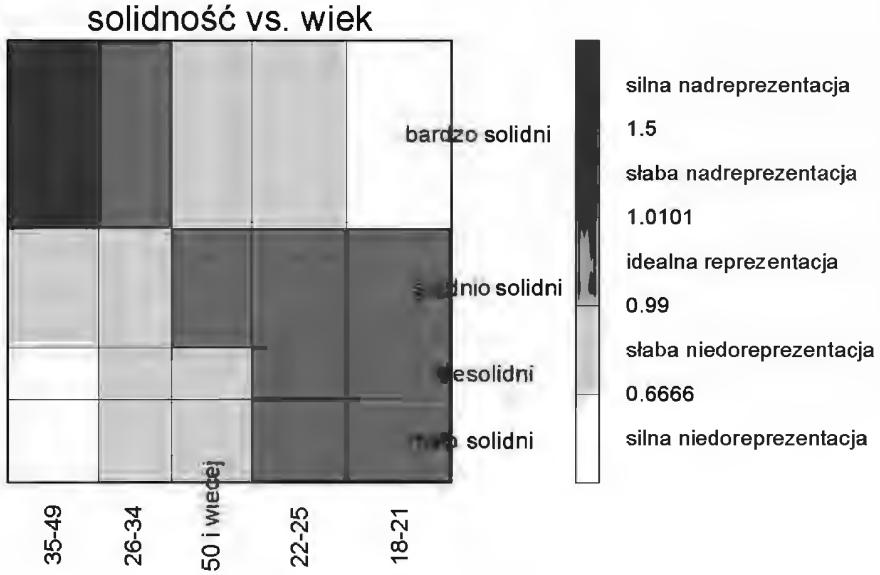
Mapę nadreprezentacji należy interpretować jako rozkład dwuwymiarowy, uwzględniający solidność klientów (zmienna wierszowa) i ich wiek (zmienna kolumnowa). Mapa jest uproszczeniem hiperpowierzchni, jej pewnego rodzaju kwantyzacją. Szerokość kolumn odpowiada odsetkowi w poszczególnych grupach wiekowych, natomiast szerokość wierszy - odsetkowi klientów w danej grupie zmiennej *solidność*.




Rysunek 7 - Mapa nadreprezentacji

Szerokości można zatem utożsamić z rozkładami brzegowymi poszczególnych cech *solidność* i *wiek*. Wnętrze mapy zawiera już więcej informacji - tych dotyczących reprezentacji ilościowej danej cechy (*solidność*) w grupach innej cechy (*wiek*). Mapę nadreprezentacji należałoby traktować jako wizualizację tablicy kontyngencji, w której każda komórka *(i,j)* reprezentuje frakcję charakteryzującą się *i*-tą wartością zmiennej wierszowej oraz *j*-tą wartością zmiennej kolumnowej. Odpowiednie odcienie odzwierciedlają stosunek liczby klientów w danej komórce w odniesieniu do liczby, jaka powinna się tam znaleźć gdyby rozkład łączny był jednostajny. Jeśli w danej komórce prezentowana jest silna nadreprezentacja oznacza to, że liczbę klientów przydzielonych do danej komórki można uznać za wyjątkowo wysoką w stosunku do liczby klientów w przypadku jednakowego rozkładu, proporcjonalnego do rozkładów brzegowych obu cech. Natomiast komórka w odcieniu białym jest niedoreprezentowana, czyli liczba osób reprezentująca tę komórkę jest mniejsza niż miałoby to miejsce w przypadku braku jakiejkolwiek koncentracji. Spróbujemy przegrupować wiersze i kolumny w taki sposób, aby wprowadzić optymalny gradacyjny porządek dla zmiennych *wiek* oraz *solidność*. Po uszeregowaniu algorytmem GCA (patrz rysunek 8) -

grupy zostały ułożone tak, aby zmaksymalizować różnorodność między wierszami i kolumnami.



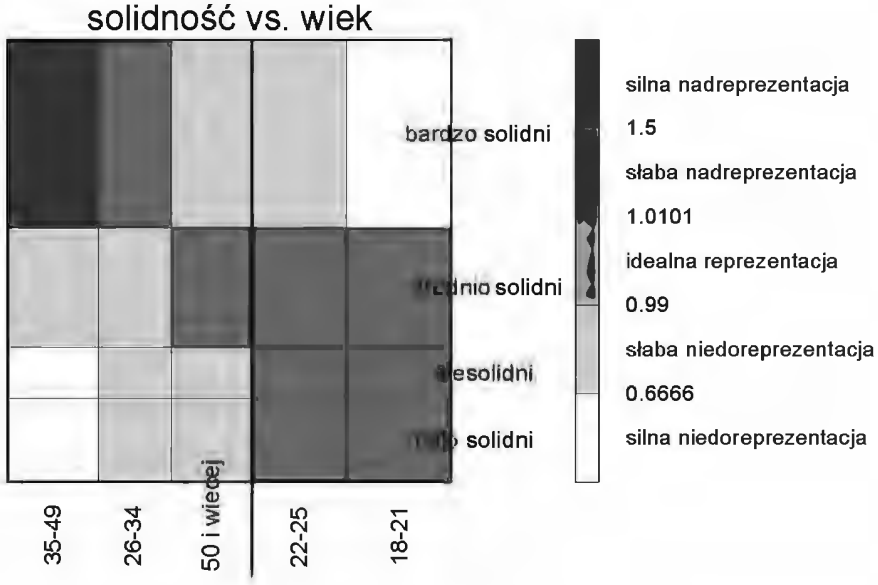

Rysunek 8 - Mapa nadreprezentacji po wykonaniu algorytmu GCA

Mapa nadreprezentacji po zastosowaniu algorytmu grupującego pozwala na wyciągnięcie ciekawych wniosków. Zauważmy, że grupa klientów bardzo solidnych reprezentowana jest wyjątkowo w klasie osób od 35 do 49 roku życia oraz w dużym stopniu w grupie młodszej (*26-34*). W grupie powyżej 50 lat przeważają klienci średnio solidni, natomiast w dwóch młodszych grupach wiekowych również klienci średnio solidni i niesolidni. Grupy wiekowe *22-25* oraz *18-21* są bardzo do siebie podobne, jednak różnią się reprezentacją osób bardzo solidnych - na niekorzyść grupy *18-21* - tam jest ich najmniej w stosunku do wszystkich klientów w tej grupie (*18-21*) oraz uwzględniając całkowitą liczbę osób bardzo solidnych. Taka analiza, z punktu widzenia biznesu, daje sygnał do dyskusji na temat przyznawania kredytu dla osób z grupy wiekowej *35-49* na mniej restrykcyjnych zasadach oraz korzystniejszych warunkach. Natomiast zalecana może być szczególna ostrożność w przypadku grupy wiekowej *18-22* odnośnie udostępnionych im produktów oraz warunków kredytowania. Aby obniżyć straty, jakie generuje

w tej grupie dość duża frakcja klientów niesolidnych, być może należałoby podwyższyć marżę dla klientów z całej omawianej grupy wiekowej.

Spróbujmy zagregować podane grupy wiekowe i wyróżnić skupienia, w których podobieństwo jest duże między grupami wewnątrz skupienia oraz małe między grupami z różnych skupień. Próbę podziału na dwa zbiory przedstawia rysunek 9.

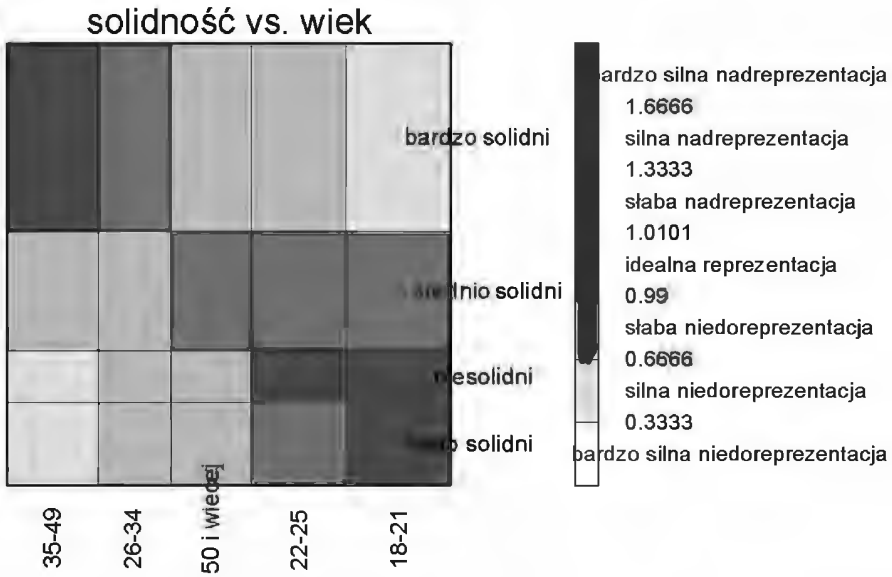
1. Skupienie pierwsze zawiera klasy wiekowe odpowiadające osobom w wieku 26 lat i większym - osoby te uznane są za potencjalnie lepszych klientów.
2. Skupienie drugie zawiera osoby poniżej 26-go roku życia - wśród tej grupy zaobserwowano wyższy niż normalnie odsetek osób niesolidnych lub mało solidnych.




Rysunek 9 - Mapa nadreprezentacji - wydzielenie dwóch skupień

Przy ogólnym spojrzeniu na mapę wyróżniają się klasy wiekowe *35-49* i *26-34* jako grupy zawierające największy udział osób bardzo solidnych w stosunku do wszystkich klientów w danej grupie i jednocześnie próbie. Dlaczego zatem te dwie grupy nie zostały wydzielone jako jedna? Dlaczego

do grupy klientów „lepszych" zaklasyfikowano również grupę *50+*? Przyczyną takiej sytuacji może być większa bliskość tej grupy do grup *26-34* oraz *35-49* niż do pozostałych dwóch „młodszych" grup. Należy spojrzeć również na reprezentację osób mało solidnych i niesolidnych - ta z kolei jest bliższa dla grupy *50+* grupom „starszym" niż „młodszym"'. Podejrzenie spróbujemy uzasadnić wykreślając mapę z większą szczegółowością dotyczącą wartości nadreprezentacji (patrz rys. 10).




Rysunek 10 - Mapa nadreprezentacji - bardziej szczegółowe ujęcie

Przy większej precyzji rysunku zauważyć można, że grupy *22-25* oraz *18-21* są podobne pod względem reprezentacji osób niesolidnych - to właśnie było powodem przydzielenia osób z klasy *50+* do klas *26-34* i *35-49*. Pomimo mniejszego podobieństwa w klasie osób bardzo solidnych, to podobieństwo w klasie osób niesolidnych i mało solidnych okazało się być decyzyjne w kwestii omawianego przydziału grupy najstarszej do dwóch wyraźnie „najlepszych". Większe znaczenie miało wydzielenie osób gorszych niż agregacja zdecydowanie lepszych.

## Podsumowanie

W artykule starano się w intuicyjny sposób przedstawić podstawy metod gradacyjnych oraz wszechstronność ich zastosowania ze zwróceniem uwagi na prostotę prezentacji wyników. W jaki sposób powyższą wizualizację wytłumaczyć na gruncie teorii matematyki i rachunku prawdopodobieństwa, jakie przedstawienie analityczne ma mapa nadreprezentacji oraz czym jest wskaźnik mówiący o stopniu maksymalnego skoncentrowania cechy – wszystko to zostało obszernie opisane w [7] oraz [16]. Dzięki silnemu akcentowi na wizualizację, metody gradacyjne mogą znaleźć zwolenników wśród osób pragnących w prosty i szybki sposób dokonać użytecznych analiz, tudzież tych, którym przeprawa przez meandry tradycyjnej statystyki nastręcza trudności lub których całkowicie zniechęca. Metody gradacyjne są z powodzeniem stosowane w przy wyznaczaniu trendów w danych ([4], [12]), w analizie skupień ([1], [2], [3], [6], [10], [14], [15]), wyszukiwaniu elementów odstających ([11]), są pomocne w lokalizacji błędów oraz przy uzupełnianiu brakujacych danych ([9]).

## Literatura

[1] Ciok A., (2000), *Double versus optimal grade clusterings*, in: Kiers H.A.L., Rasson J.-P., Groenen P.J.F., Schader M. (Eds.), Data Analysis, Classification, and Related Methods, Springer, pp. 41-46.

[2] Ciok A. (2004), *On the number of clusters - a grade approach*, Instytut Podstaw Informatyki PAN, Warszawa.

[3] Ciok A. (1998), *Discretization as a tool in cluster analysis*, in: Rizzi A., Vichi M., Bock H.-H. (Eds.), Advances in Data Science and Classification, Proceedings of the 6th Conference of the International Federation of Classification Societies (IFCS-98), Springer, Rome, July 21-24, 1998, pp. 349-354.

[4] Ciok A., Bulhak W., Skoczylas R. (1997), Exploration of control-experimental data by means of grade correspondence analysis, *Biocybernetics and Biomedical Engineering*, 17, pp. 101-113.

[5] Ciok A., Kowalczyk T., Pleszczyńska E., Szczesny W. (1995), Algorithms of grade correspondence cluster analysis, *Archiwum Informatyki Teoretycznej i Stosowanej* (The Collected Papers on Theoretical and Applied Computer Science), 7(1-4), pp. 5-22.

[6] Jarochowska E., Ciesielski K. (2006), *Grade clustering and seriation of words based on their co-occurrences*, in: Guerrero Bote V. P. (Eds.), Current Research in Information Sciences and Technologies. Multidisciplinary approaches to global information systems, Proceedings of the 1st International Conference on Multidisciplinary Information Sciences and Technologies, InSciT2006, 25-28 October 2006, Merida, Spain (in print).

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

[8] Kowalczyk T. (2000), Link between grade measures of dependence and of separability in pairs of conditional distributions, *Statistics and Probability Letters* 46 (2000), 371-379.

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

[10] Szczesny W., Ciok A., Pleszczyńska E. (1998), Clustering land districts according to their farm magnitude repartition, Statistics in Transition, 3, pp. 757-768.

[11] Szczesny W. (2000), Detecting rows and columns of contingency table, which outlie from a total positivity pattern, *Control and Cybernetics* 19(4), pp. 31-40.

[12] Szczesny W. (2002), Grade correspondence analysis applied to contingency tables and questionnaire data, *Intelligent Data Analysis*, 6(1), pp. 17-51.

[13] Szczesny W., Pleszczyńska E. (1997): A grade statistics approach to exploratory analysis of the HSV data, *Biocybernetics and Biomedical Engineering*, 17, pp. 235-245.

[14] Taylor R. (2000), *Guidelines for Identifying Clusters Using Grade Correspondence Analysis: Practical and Technical Issues*, Microsimulation Unit Research Note MU/RN/39.

[15] Taylor R., Gomulka J., Sutherland H. (2000), *Creating order out of chaos? Identifying homogeneous groups of households across multiple datasets*, Proceedings of the 26th General Conference of the International Association for Research in Income and Wealth, Cracow, Poland, August 27..

[16] Wesołowska M. (2008), *Gradacyjne Metody Analizy Danych*, nieopublikowana praca magisterska, Politechnika Warszawska, Warszawa.

ISBN 9788389475336

J. Hołubiec, red. – TECHNIKI INFORMACYJNE-TEORIA I ZASTOSOWANIA, tom 1 (13 )