# Predykcja a kozy - studium przypadku

I.T. Podolak, A. Roman, K. Bartocha

Instytut Informatyki UJ

26 lutego 2010

# Czy kozy się rozmnażają?

# Podstawowe warunki i pytania

- na wiele tysięcy kóz przypada kilkadziesiąt kozłów
- wartość kozy zależy od materiału genetycznego i od środowiska
- każda koza ma w życiu od 7 do 10 okresów laktacji
  - ilość mleka w litrach
  - zawartość tłuszczu i białka
  - długość laktacji

- co chcemy przewidzieć u koźlęcia?
- czy na podstawie rodziców jesteśmy w stanie przewidzieć wartość oczekiwanego koźlęcia?
- czy potomstwo zależy od obojga rodziców czy też istnieją super-kozły?
- jaki jest wpływ środowiska (stada)?

- hodowcom zależy na jak najlepszym potomstwie
  - istnieje baza danych kozłów, które są do "wynajęcia"
  - z założenia hodowcy zapładniają tylko swoje kozy
  - każdy hodowca może mieć inne oczekiwania

# Jakie są atrybuty rodziców?

- co potrzebujemy zbudować?
  - zbiór atrybutów rodziców $\Longrightarrow$ MODEL $\Longrightarrow$ parametry potomstwa
    - ★ model możemy zbudować wykorzystując znane trójki (ojciec, matka, potomek)
    - ★ model powinien dobrze przewidywać przykłady, których wcześniej nie widział
- matka także daje mleko
  - atrybutami matki mogą być parametry jej laktacji
- kozioł nie daje mleka
  - parametry laktacji matki kozła?
  - parametry laktacji wcześniejszego potomstwa tego kozła (half-sisters)?
    - ★ problem z predykcją dla kozłów rozpoczynających pracę

# Opis bazy danych

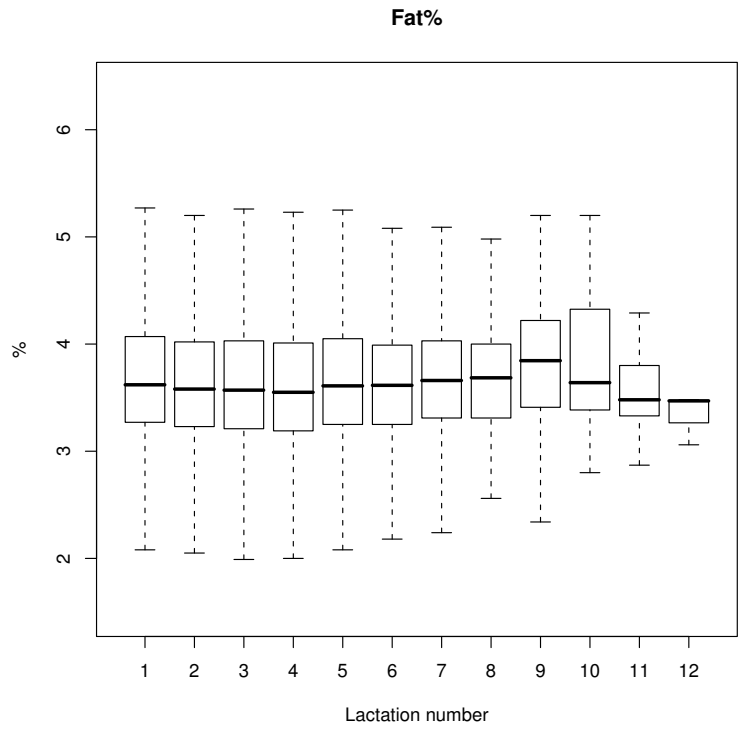
| Table | Records | Field | Desc. |
|---|---|---|---|
| Dam | 7173 | NUMBER | Individual's no. |
| | | HERD | herd's no. |
| | | RACE | Race number |
| Sire | 190 | NUMBER | Individual's no. |
| | | HERD | herd's no. |
| | | RACE | Race number |
| Family | 7372 | SEX | Individual's sex |
| | | NUMBER | Individual's no. |
| | | FATHER_NUMBER | individuals father no. |
| | | MOTHER_NUMBER | individuals mother no. |
| Lact-ation | 16862 | NUMBER | Individual's number |
| | | LACTATION_NUMBER | Lactation's no. |
| | | DAYS | lactation's length (days) |
| | | MILK_OUTPUT | summ. lactation output (kg) |
| | | FAT_PERCENT | Mean fat% |
| | | PROTEIN_PERCENT | Mean protein% |

# Problemy z atrybutami

- dane są bardzo dziurawe
  - niewiele pełnych opisów wszystkich laktacji
    - nie wystarczające do nauczania
  - w jaki sposób wykorzystać istniejące?
    - uzupełniać dane średnimi?
    - uzupełniać dane korzystając z dodatkowego modelu?
    - przetwarzać dane według modelu liniowego?

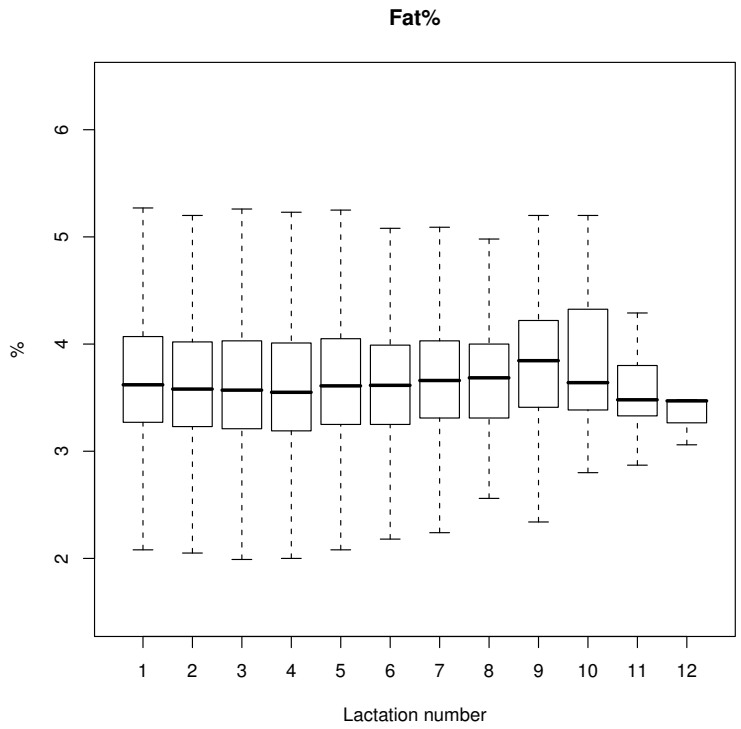
# Indeksy

- jak reprezentować laktacje?
- $$I^k = \frac{1}{L_k} \sum_{l=1}^{L_k} \frac{x_l^k - \mu_l}{3\sigma_l}, \qquad (1)$$
  - $I^k \in \mathbb{R}$ - indeks dla k-tej kozy
  - $x_l^k$ - wartość parametru k-tej kozy w l-tej laktacji
  - $\mu_l$ - średnia wartość parametru dla l-tej laktacji
  - $\sigma_l$ - odchylenie dla l-tej laktacji
  - $L_k$ - liczba laktacji
- indeks obliczany dla każdego atrybutu (mleko, tłuszcz, białko, czas trwania laktacji)
- normalizacja - $\geqslant 99\%$ wartości z przedziału $[-1, 1]$
- indeks:
  - pozwala na wykorzystanie niepełnych danych
  - obrazuje kozę na tle innych



# Indeksy

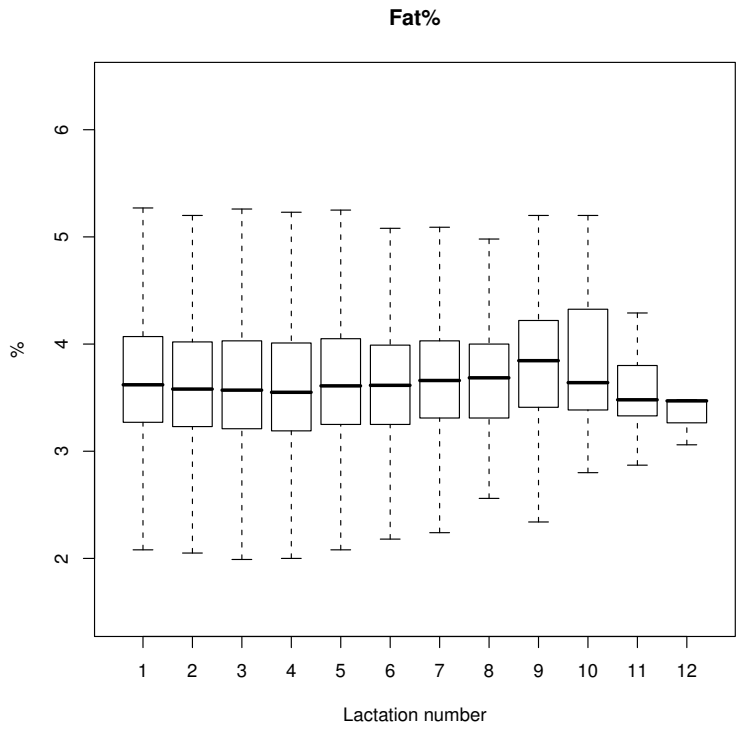
- jak reprezentować laktacje?
- $$I^k = \frac{1}{L_k} \sum_{l=1}^{L_k} \frac{x_l^k - \mu_l}{3\sigma_l}, \qquad (1)$$
  - $I^k \in \mathbb{R}$ - indeks dla k-tej kozy
  - $x_l^k$ - wartość parametru k-tej kozy w l-tej laktacji
  - $\mu_l$ - średnia wartość parametru dla l-tej laktacji
  - $\sigma_l$ - odchylenie dla l-tej laktacji
  - $L_k$ - liczba laktacji
- indeks obliczany dla każdego atrybutu (mleko, tłuszcz, białko, czas trwania laktacji)
- normalizacja - $\geqslant 99\%$ wartości z przedziału $[-1, 1]$
- indeks:
  - pozwala na wykorzystanie niepełnych danych
  - obrazuje kozę na tle innych



# Indeksy

- jak reprezentować laktacje?
- $$I^k = \frac{1}{L_k} \sum_{l=1}^{L_k} \frac{x_l^k - \mu_l}{3\sigma_l}, \qquad (1)$$
  - $I^k \in \mathbb{R}$ - indeks dla k-tej kozy
  - $x_l^k$ - wartość parametru k-tej kozy w l-tej laktacji
  - $\mu_l$ - średnia wartość parametru dla l-tej laktacji
  - $\sigma_l$ - odchylenie dla l-tej laktacji
  - $L_k$ - liczba laktacji
- indeks obliczany dla każdego atrybutu (mleko, tłuszcz, białko, czas trwania laktacji)
- normalizacja - $\geqslant 99\%$ wartości z przedziału $[-1, 1]$

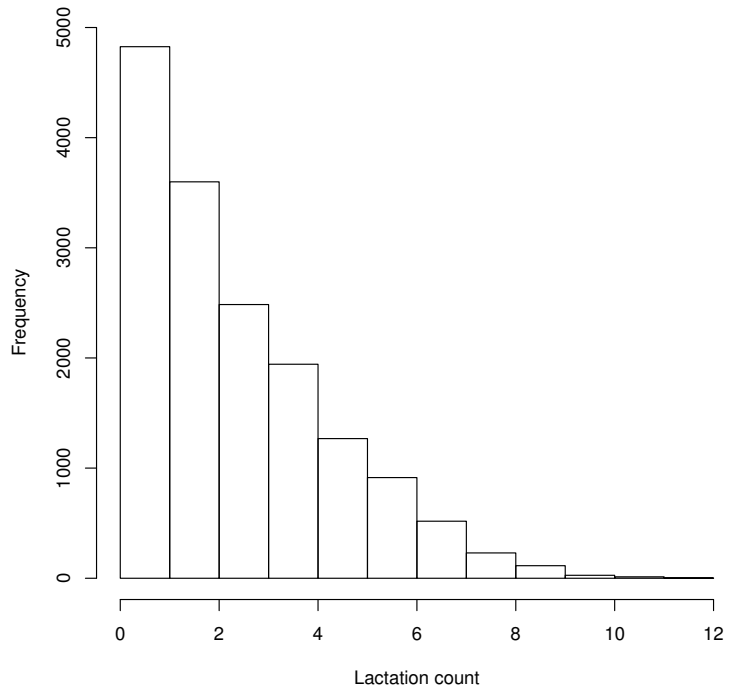
- indeks:
  - pozwala na wykorzystanie niepełnych danych
  - obrazuje kozę na tle innych



# Histogram liczby laktacji i zalety koncepcji indeksu

- nie wszystkie kozy mają taką samą liczbę laktacji
- istnieją kozy z "dziurami" w laktacjach, np. dostępne są dane dla laktacji nr 1,2,3,6
- koncepcja indeksu pozwala obejść niektóre problemy:
  - ponieważ uśrednia, jest niewrażliwa na zmienną liczbę laktacji
  - ponieważ uśrednia, jest niewrażliwa na "dziury"
- niewiele kóz z $> 7$ laktacjami $\Rightarrow$ duże odchylenie std
- zatem rozważamy tylko pierwszych 7 laktacji

- każda koza to element $x \in \mathbb{R}^4$



# Po co dane o środowisku?

- Animal Model:
- $$V = E + G + cov(E, G), \tag{2}$$
  - V - całkowita wartość osobnika
  - E - wpływ środowiska
  - G - wpływ czynników genetycznych
- w indeksie występuje średnia $\mu_l$ ze stada - to jest wpływ środowiska

# Jaki model?

- liniowy (BLUP)
  - najczęściej wykorzystywany (choć świat jest zwykle nieliniowy)
  - przewidywanie wszystkich parametrów potomstwa na podstawie atrybutów rodziców
- nieliniowy
- możliwość wykorzystania modelu sieci neuronowych
- znane algorytmy skutecznego nauczania
- problem klasyfikacji czy aproksymacji funkcji?
  - zbiór uczący składa się z par przykładów (x, t)
  - dla problemu klasyfikacji t należy do ustalonego zbioru etykiet
    - ⋆ aproksymacja atrybutów koźląt jest nierelana
- co tak na prawdę interesuje hodowcę?
  - raczej typ zwierzęcia
  - predykcja będzie obarczona mniejszym błędem

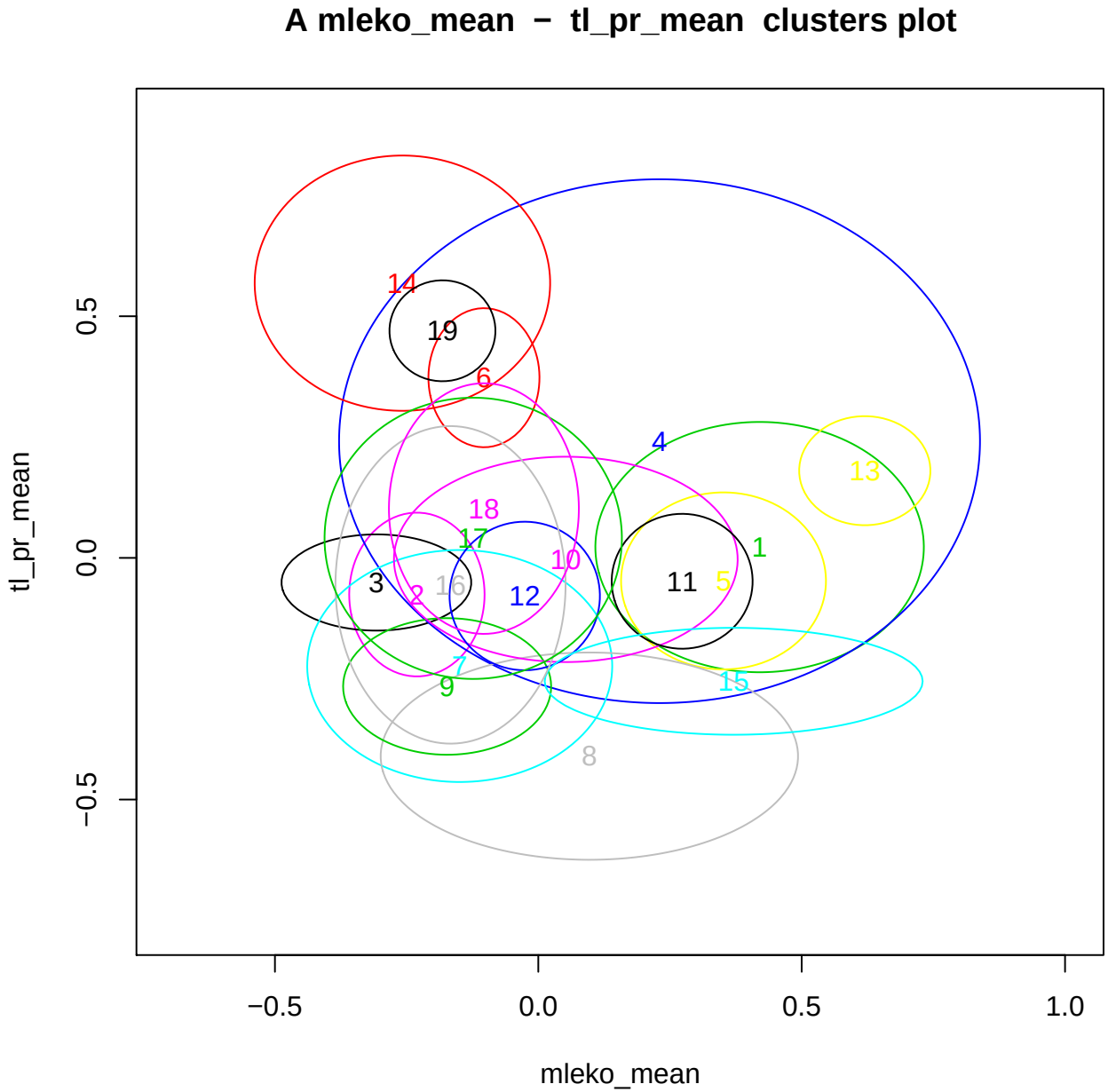
# Kategorie zwierząt

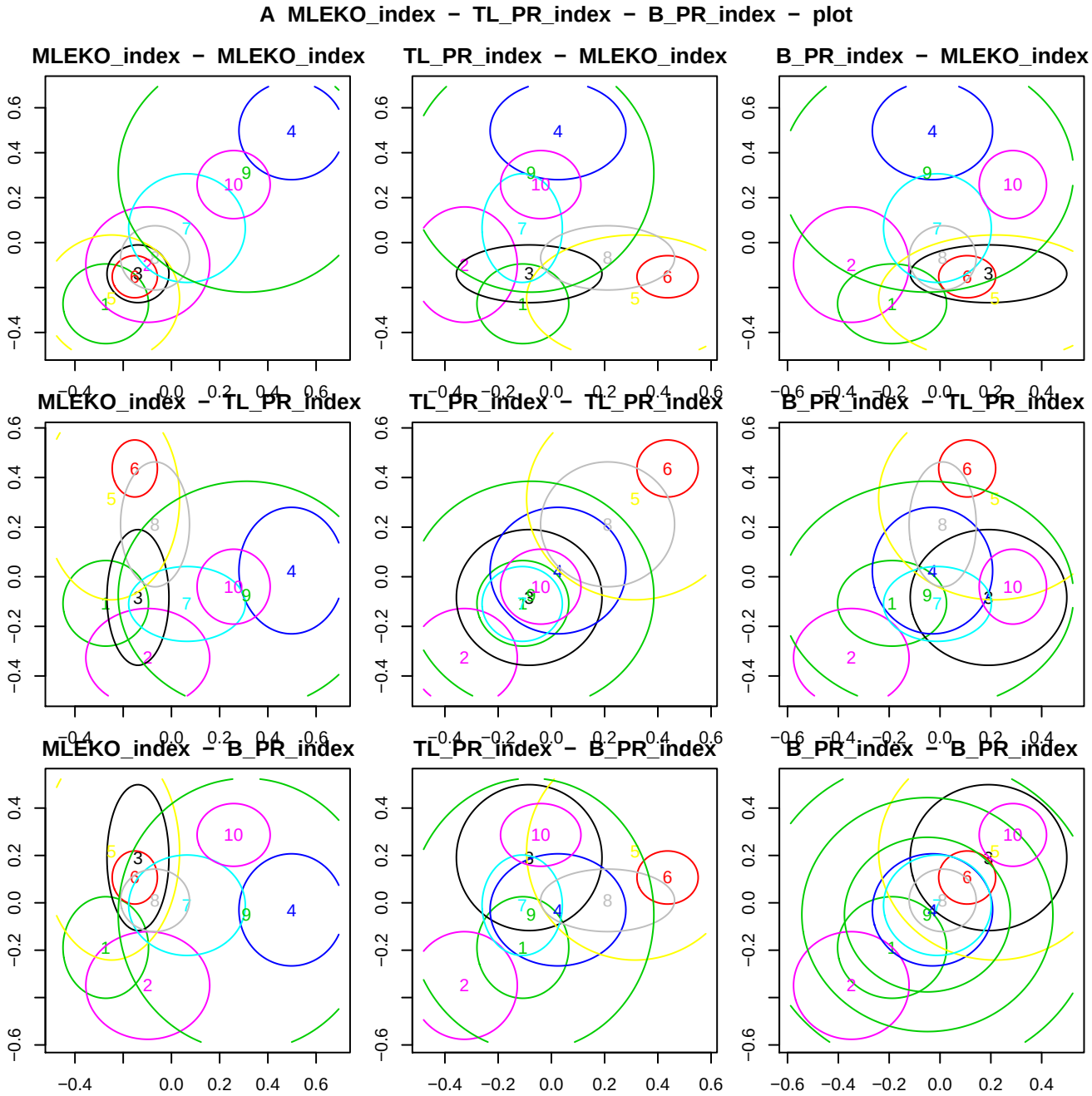
- możemy zapytać hodowców (“ekspertów”)
  - trudne w realizacji: kosztowne, różne odpowiedzi, czasochłonne

- a może by samemu stworzyć grupy zwierząt?
  - pogrupować zwierzęta o podobnych atrybutach?
  - jakie algorytmy?
  - ile grup?

  - klastrowanie!! (zamiana problemu predykcji (”ciągłej”) na problem klasyfikacji)
    - atrybuty zwierzęcia reprezentują punkt w przestrzeni atrybutów
    - grupujmy razem zwierzęta blisko siebie

    - algorytmy K-średnich, rozmyte, hierarchiczne


A mleko_mean − tl_pr_mean clusters plot
tl_pr_mean
mleko_mean
0.5
0.0
−0.5
−0.5
0.0
0.5
1.0
14
19
6
4
13
18
17
10
1
3
16
2
12
11
5
7
9
15
8


A MLEKO_index − TL_PR_index − B_PR_index − plot
MLEKO_index − MLEKO_index
TL_PR_index − MLEKO_index
B_PR_index − MLEKO_index
MLEKO_index − TL_PR_index
TL_PR_index − TL_PR_index
B_PR_index − TL_PR_index
MLEKO_index − B_PR_index
TL_PR_index − B_PR_index
B_PR_index − B_PR_index

# Ile klastrów?

- przykłady w pojedynczych klastrach powinny być wyraźnie do siebie podobne

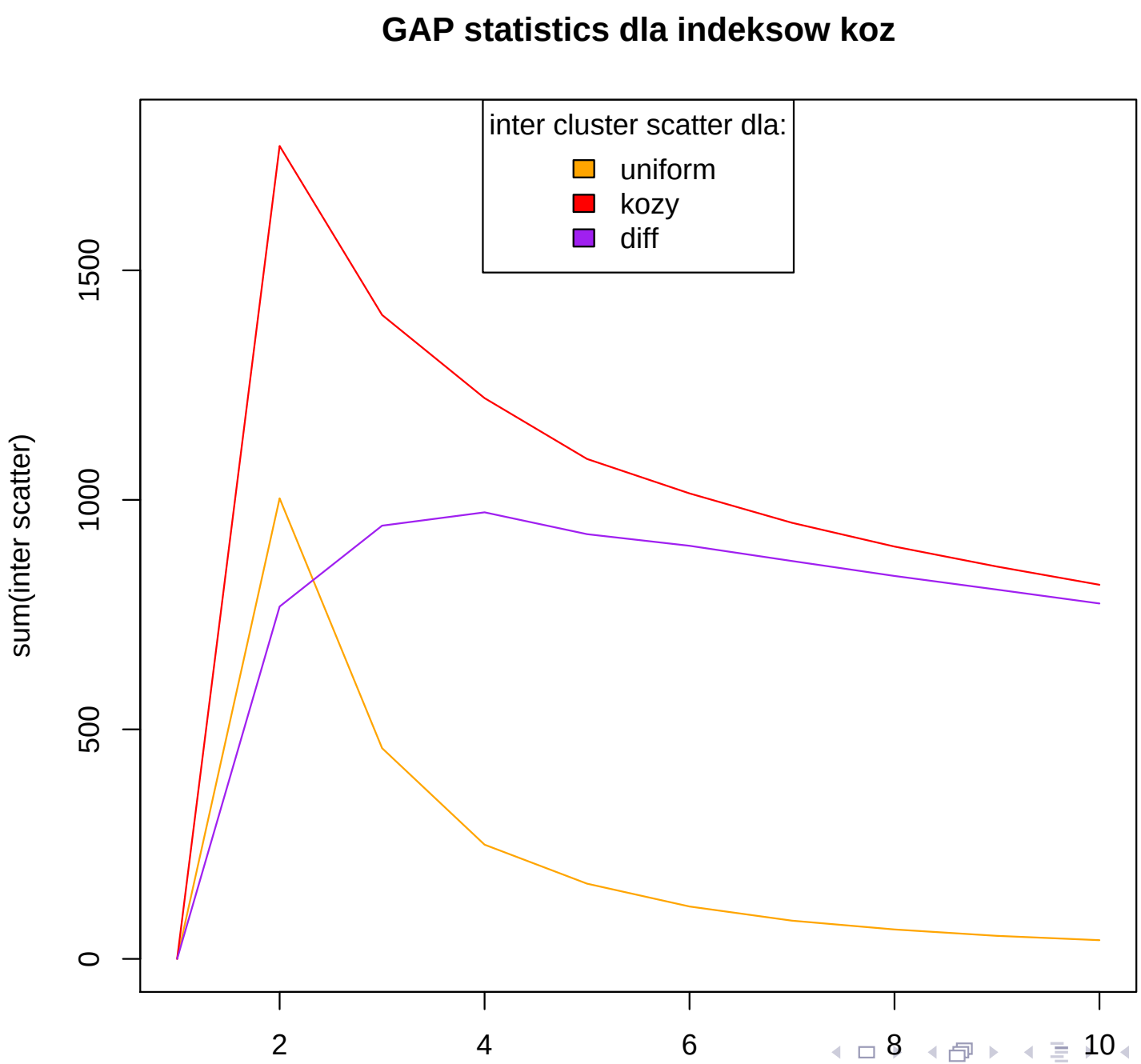
$$\mathrm{wcs}(C) = \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(i')=k} d(x_i, x_{i'})$$

- to w różnych natomiast wyraźnie różne

$$\mathrm{bcs}(C) = \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(i')\neq k} d(x_i, x_{i'})$$

- $\mathrm{wcs}(C) + \mathrm{bcs}(C)$ jest stałe dla danego zbioru danych
- minimalizujemy wcs() lub maksymalizujemy bcs()

- jaka jest optymalna liczba klastrów $K^*$?
  - dla $K < K^*$ klastry będą zawierać kilka "prawdziwych" co da szybki wzrost wcs() przy wzroście K
  - dla $K > K^*$ "prawdziwe" klastry będą dzielone, co da wolniejszy spadek wcs() przy wzroście K
  - porównajmy z rozkładem równomiernym i szukajmy wyraźnej zmiany
  - to jest tzw. gap-statistics


GAP statistics dla indeksow koz
inter cluster scatter dla:
uniform
kozy
diff
sum(inter scatter)
0
500
1000
1500
2
4
6
8
10

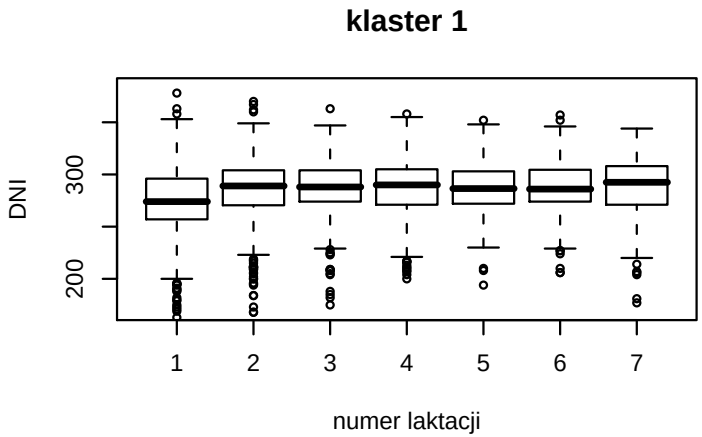
# To jest to!!!!

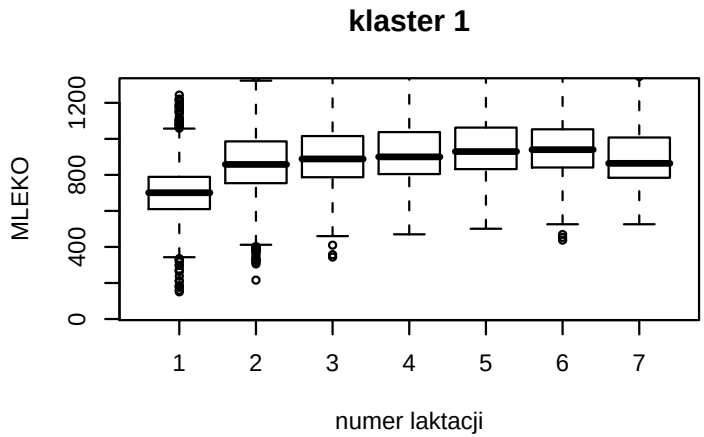
- ekspert stwierdził, że jest to rozsądną liczba grup na które można podzielić zwierzęta
- bliższa analiza znalezionych klastrów pozwoliła opisać klastry jako kozy do
  1. przetwórstwa mleka spożywczego
  2. przetwórstwa na sery miękkie (więcej białka)
  3. przetwórstwa na sery twarde (więcej tłuszczu)
  4. zwierzęta ogólnoużytkowe (tylko 8.2% zwierząt)

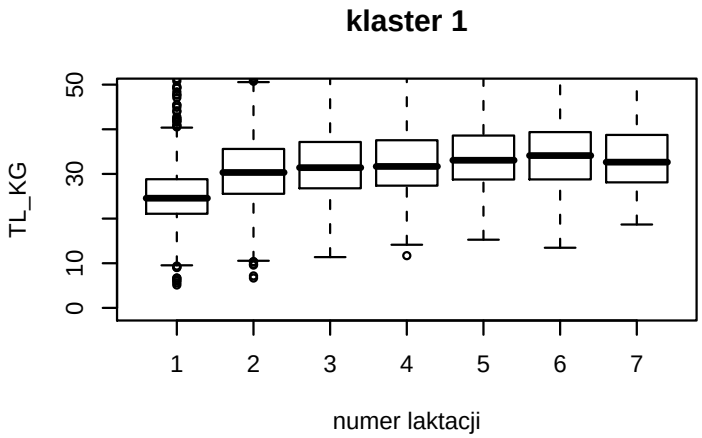
# Tabela kontyngencji wg ras i klastrów

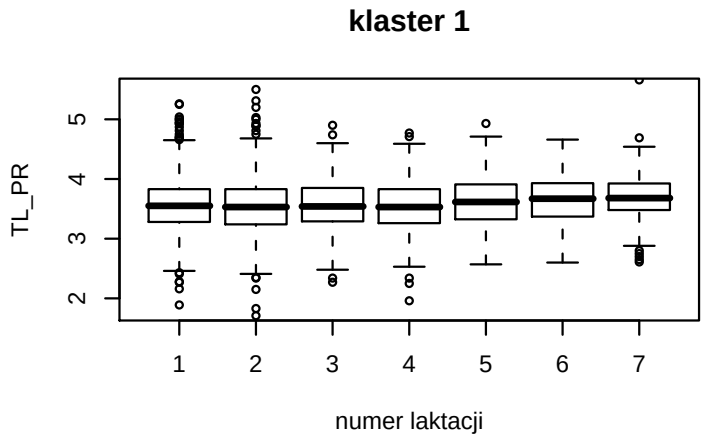
| Kod rasy | Klaster | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| 1 | 501 | 767 | 647 | 171 |
| 2 | 154 | 172 | 233 | 128 |
| 3 | 23 | 38 | 14 | 13 |
| 4 | 74 | 80 | 41 | 10 |
| 8 | 1 | 2 | 5 | 0 |
| 9 | 5 | 0 | 6 | 4 |

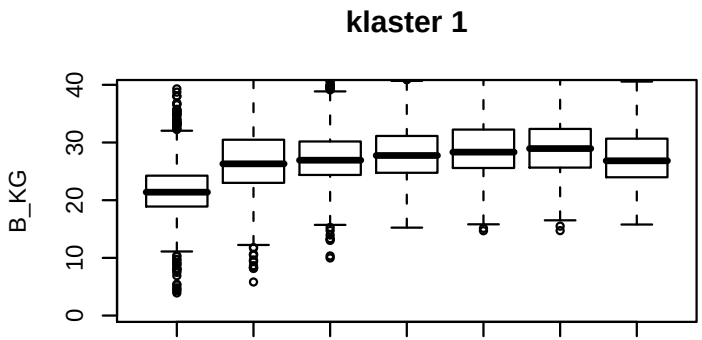
- frakcje poszczególnych ras w klastrach proporcjonalne
- wygląda na to, że sama rasa nie ma istotnego wpływu na klastrowanie

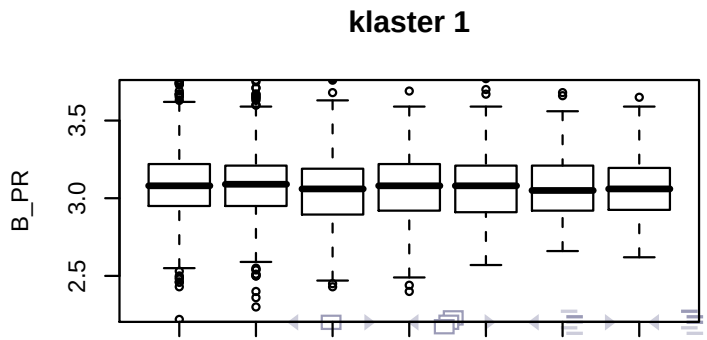

klaster 1
DNI
200
300
1
2
3
4
5
6
7
numer laktacji

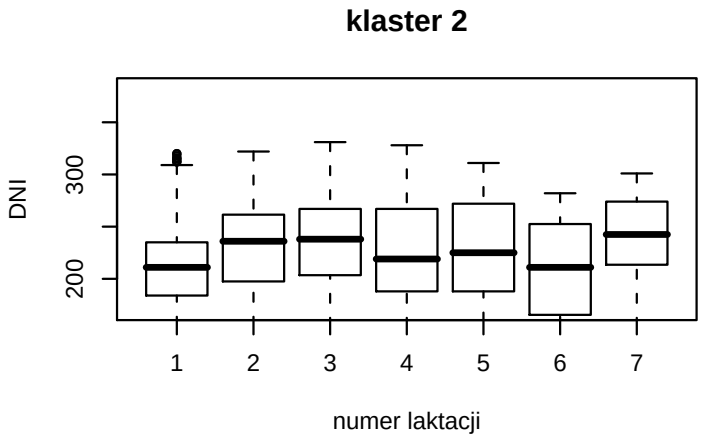

klaster 1
MLEKO
0
400
800
1200
1
2
3
4
5
6
7
numer laktacji

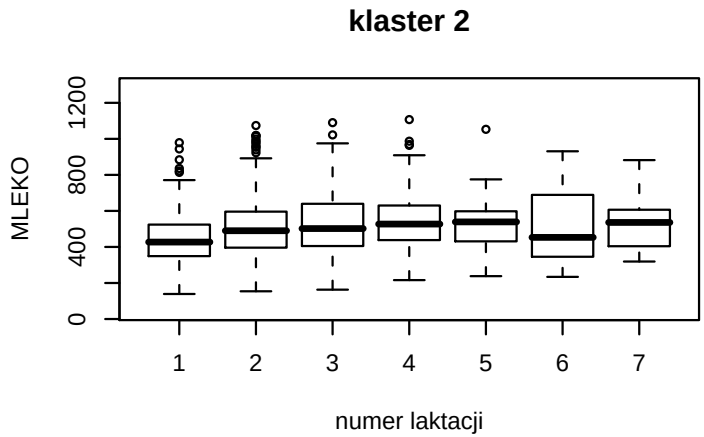

klaster 1
TL_KG
0
10
30
50
1
2
3
4
5
6
7
numer laktacji

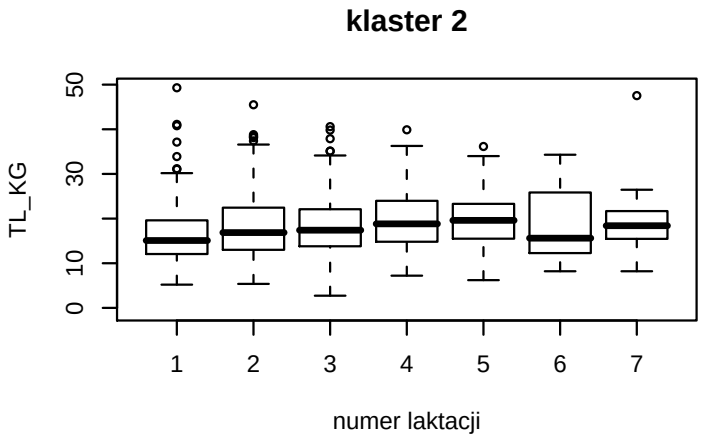

klaster 1
TL_PR
2
3
4
5
1
2
3
4
5
6
7
numer laktacji

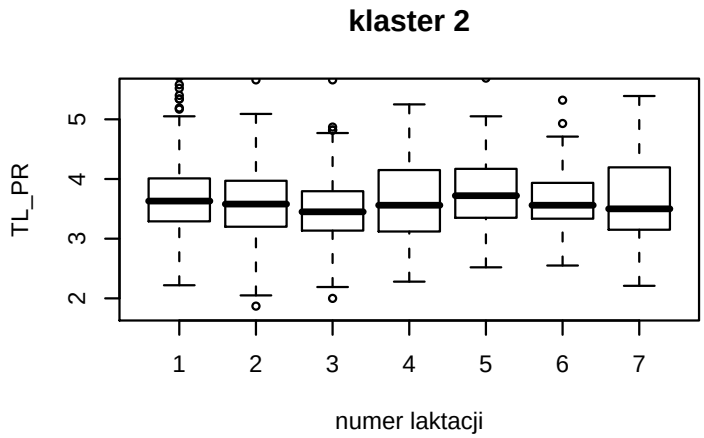

klaster 1
B_KG
0
10
20
30
40

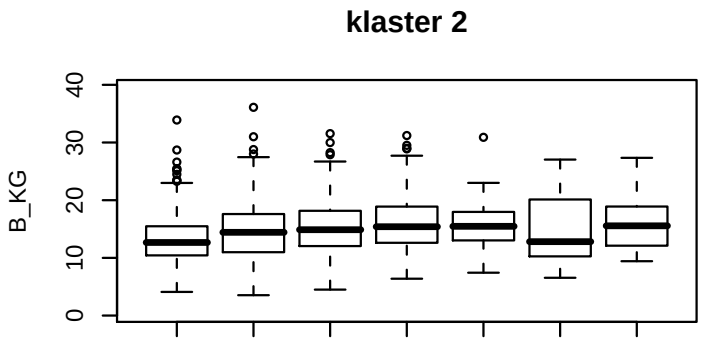

klaster 1
B_PR
2.5
3.0
3.5

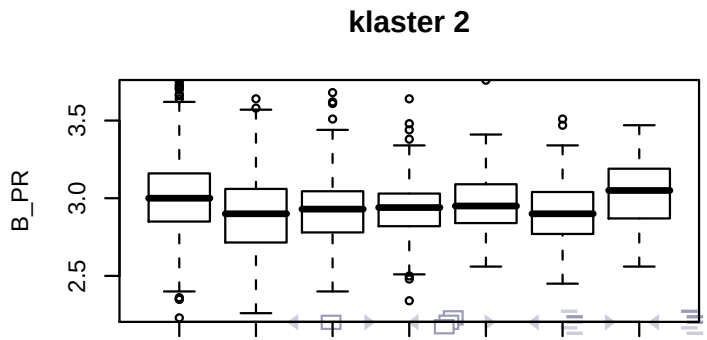

klaster 2
DNI
200
300
1
2
3
4
5
6
7
numer laktacji

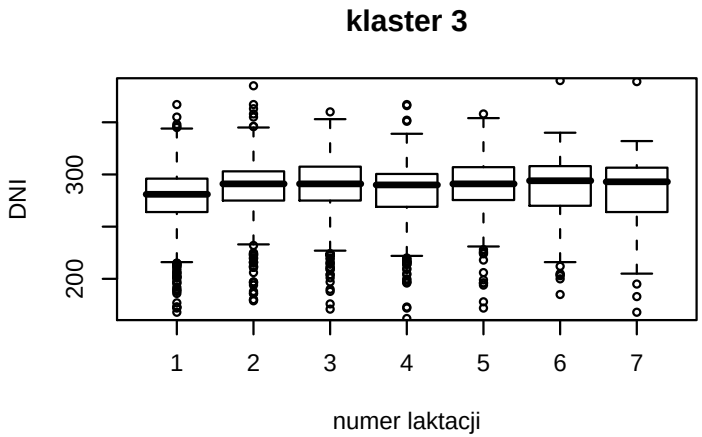

klaster 2
MLEKO
0
400
800
1200
1
2
3
4
5
6
7
numer laktacji

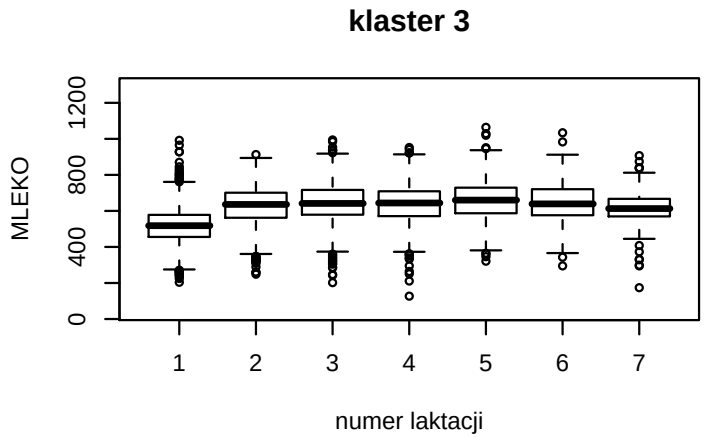

klaster 2
TL_KG
0
10
30
50
1
2
3
4
5
6
7
numer laktacji

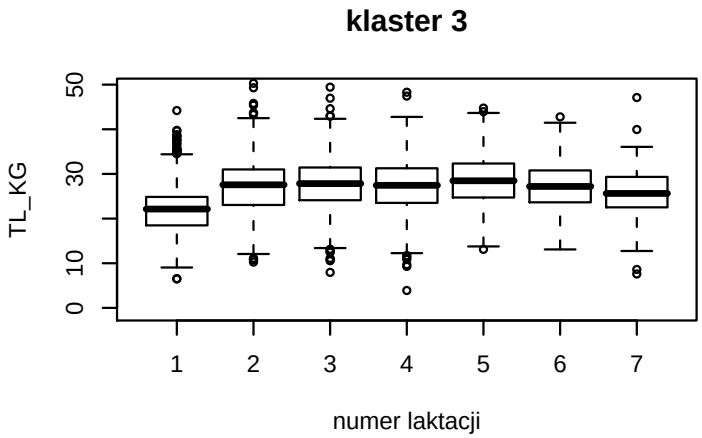

klaster 2
TL_PR
2
3
4
5
1
2
3
4
5
6
7
numer laktacji

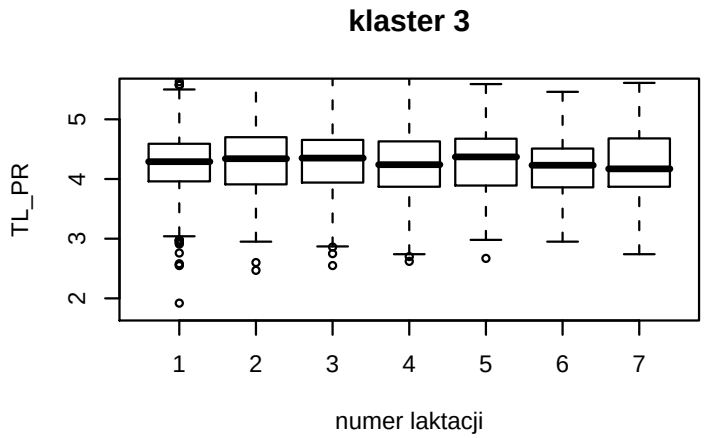

klaster 2
B_KG
0
10
20
30
40

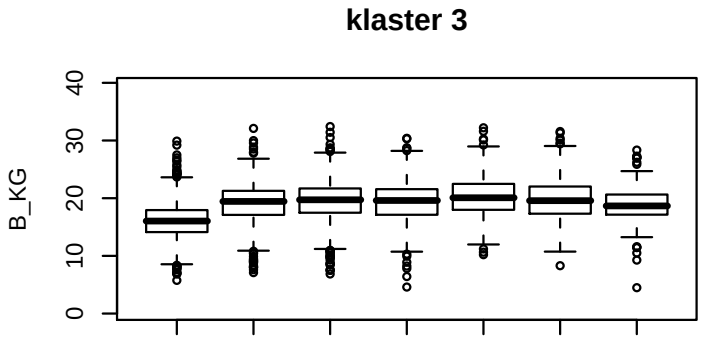

klaster 2
B_PR
2.5
3.0
3.5

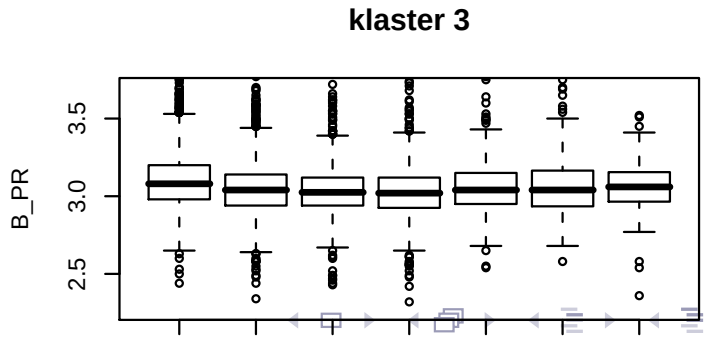

klaster 3
DNI
200
300
1
2
3
4
5
6
7
numer laktacji

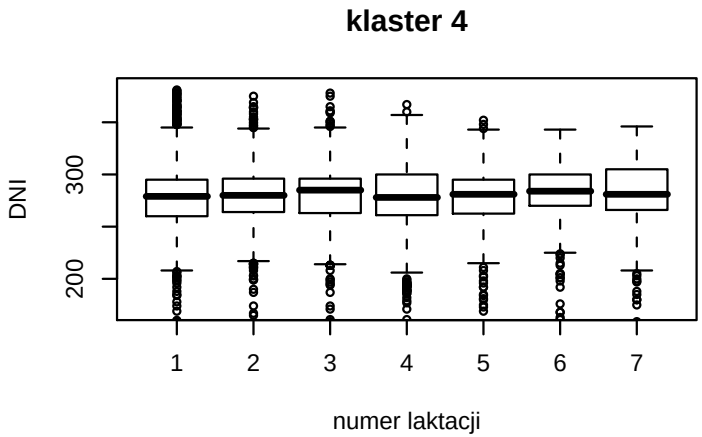

klaster 3
MLEKO
0
400
800
1200
1
2
3
4
5
6
7
numer laktacji

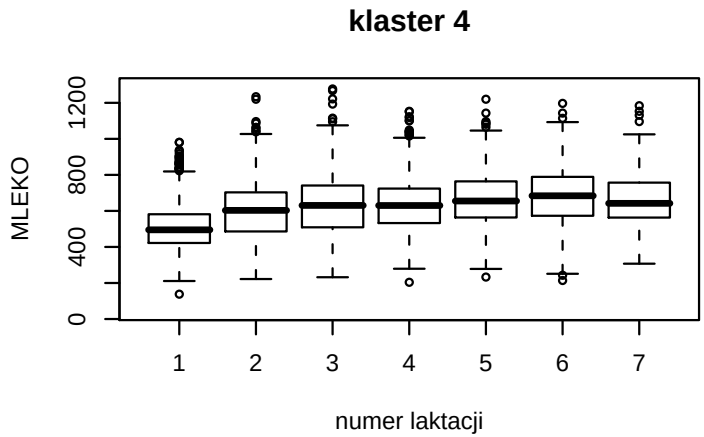

klaster 3
TL_KG
0
10
30
50
1
2
3
4
5
6
7
numer laktacji

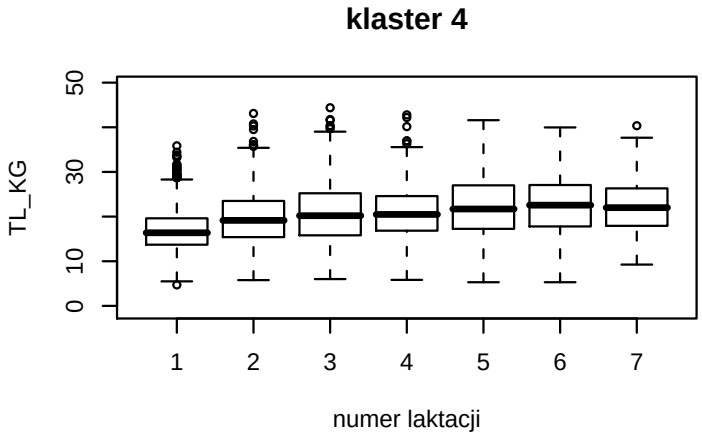

klaster 3
TL_PR
2
3
4
5
1
2
3
4
5
6
7
numer laktacji

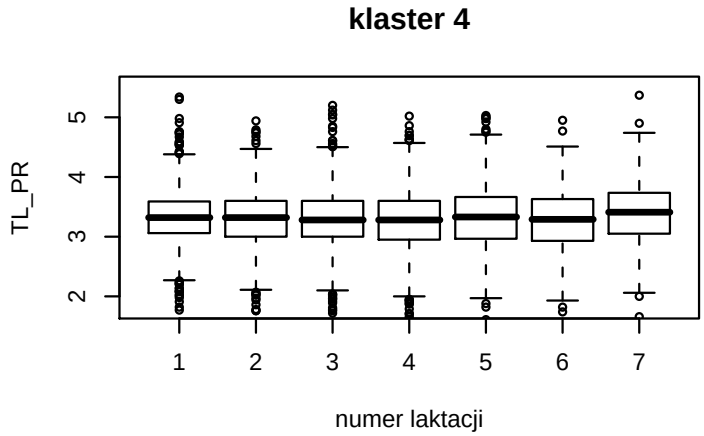

klaster 3
B_KG
0
10
20
30
40

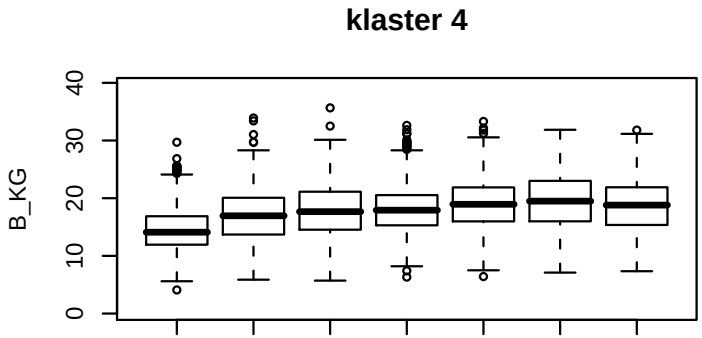

klaster 3
B_PR
2.5
3.0
3.5

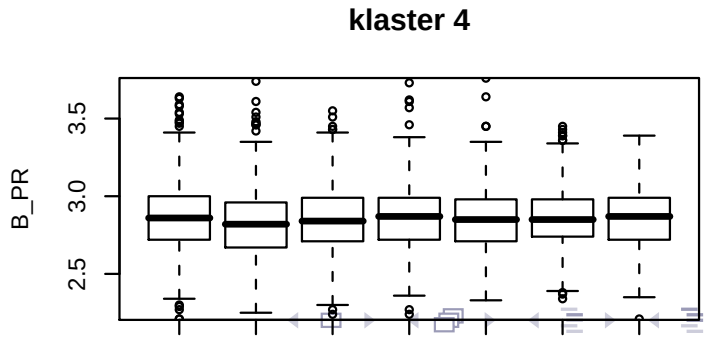

klaster 4
DNI
200
300
1
2
3
4
5
6
7
numer laktacji


klaster 4
MLEKO
0
400
800
1200
1
2
3
4
5
6
7
numer laktacji


klaster 4
TL_KG
0
10
30
50
1
2
3
4
5
6
7
numer laktacji


klaster 4
TL_PR
2
3
4
5
1
2
3
4
5
6
7
numer laktacji


klaster 4
B_KG
0
10
20
30
40


klaster 4
B_PR
2.5
3.0
3.5

# I co teraz?

- predykcja przy pomocy sieci neuronowych warstwowych

| neurony | train | test | best train | best test |
|---|---|---|---|---|
| 2 | 0.73 | 0.71 | 0.74 | 0.76 |
| 4 | 0.74 | 0.72 | 0.74 | 0.77 |
| 7 | 0.74 | 0.73 | 0.74 | 0.79 |
| 12 | 0.75 | 0.725 | 0.76 | 0.80 |
| 16 | 0.75 | 0.71 | 0.758 | 0.76 |
| 24 | 0.758 | 0.73 | 0.77 | 0.77 |
| 40 | 0.769 | 0.728 | 0.77 | 0.77 |

- czy te wyniki są wystarczająco dokładne?
- czy można zastosować inne algorytmy nauczania?
- problem okazuje się, wbrew pozorom, bardzo trudny

# A może niepotrzebnie komplikujemy sprawę? ”One-rule classifier”

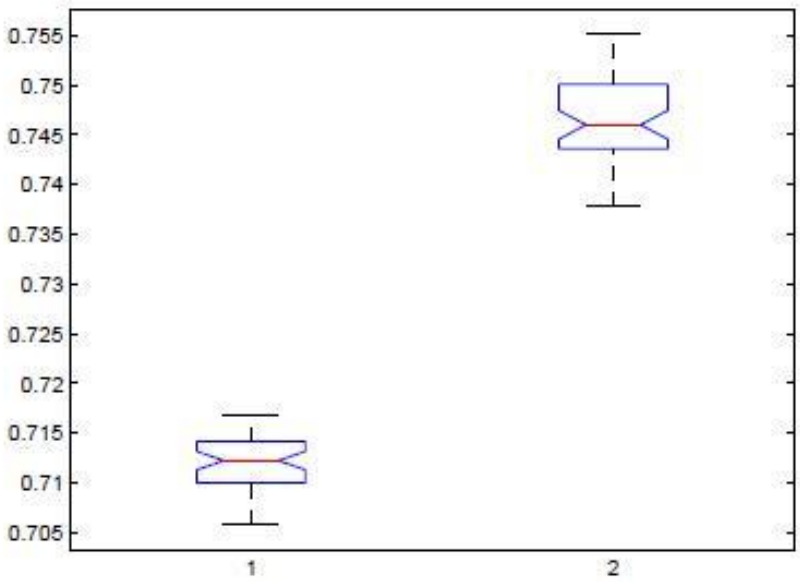
| Klaster matki | Klaster córki | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| 1 | 527 | 60 | 111 | 56 |
| 2 | 389 | 543 | 59 | 61 |
| 3 | 264 | 37 | 617 | 263 |
| 4 | 126 | 70 | 84 | 44 |

- one-rule classifier wykorzystuje ”dziedziczenie” przynależności do klastra
- działanie: jeśli matka jest w klastrze k, to córka też
- skuteczność empiryczna: $\frac{527+543+617+44}{527+60+...+84+44} = \frac{1731}{3311} = 0.5228$
- skuteczność sieci neuronowej: 0.76 - 0.80. Jednak sieć jest lepsza!

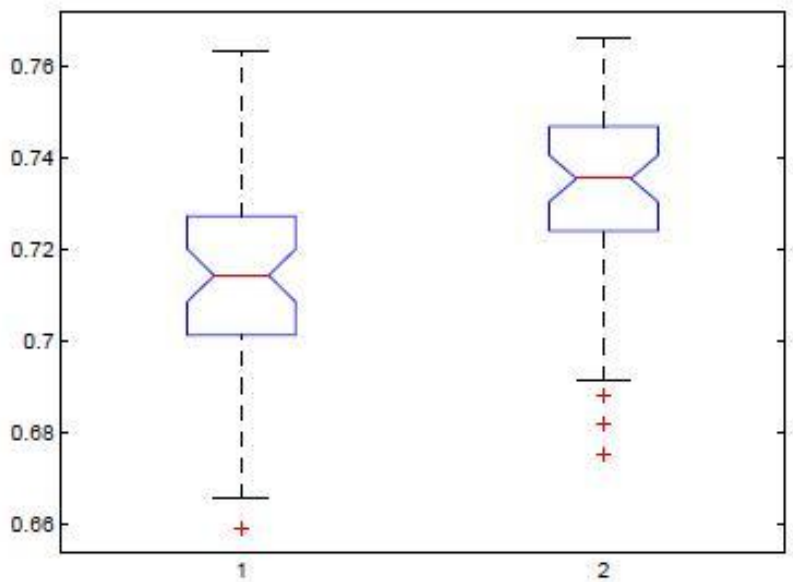
# A może jednak model liniowy jest wystarczający?

- $$Y = AX + b, \tag{3}$$
  - Y - wektor przewidywanych cech
  - X - wektor parametrów (atrybutów)
  - A - macierz współczynników
  - b - wektor współczynników
- wyniki dla modelu liniowego: zbiór uczący 0.71, zbiór testowy: 0.70

# Model liniowy vs. sieć neuronowa




ZBIÓR UCZĄCY




ZBIÓR TESTOWY

**1 model liniowy, 2 sieć neuronowa**

**p=0,00001** **p=0,01**

# Wnioski

- Koncepcja indeksu pozwala reprezentować serię laktacji jako jedną liczbę
- Koza reprezentowana jako $x \in \mathbb{R}^4$
- Transformacja problemu predykcji dokładnej wartości parametru na problem klasyfikacji
- Klastrowanie - klastry mają sens zootechniczny
- Sieć neuronowa (model nieliniowy) daje najlepsze rezultaty

# Pytania i problemy

- Czy możliwe jest inne klastrowanie?
- Czy jest jakiś lepszy model nieliniowy?
- Czy ma sens model przewidujący przebieg laktacji? (problem z brakującymi danymi)
- Może wykorzystać Hierarchiczny Klasyfikator?
  - HC sprawdza się, gdy jest wiele klas - tu mamy tylko 4
  - pomysł: klasa K opisana czwórką $(I_1^K, I_2^K, I_3^K, I_4^K)$, gdzie $I_j^K$ - pewien przedział parametru j
  - jeśli każdy parametr podzielić na 4 przedziały, to można uzyskać $4^4 = 256$ klas (niektóre mogą być puste)
- ...

# ANY QUESTIONS ???