Analiza wielowymiarowa

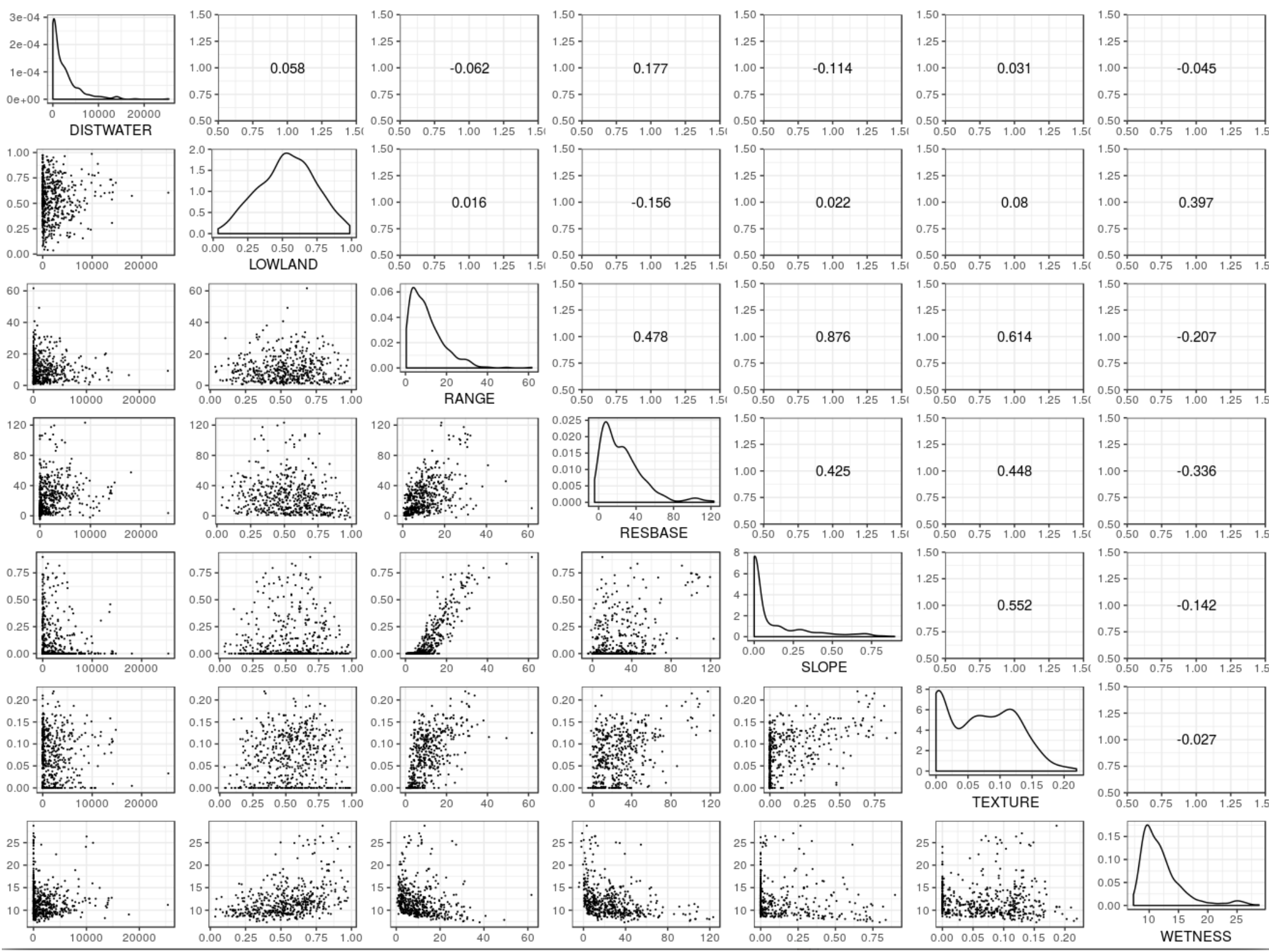
Jarosław Jasiewicz

Eksploracja danych i Uczenie maszynowe

Geoinformacja program magisterski
Specjalność Geoinformatyka

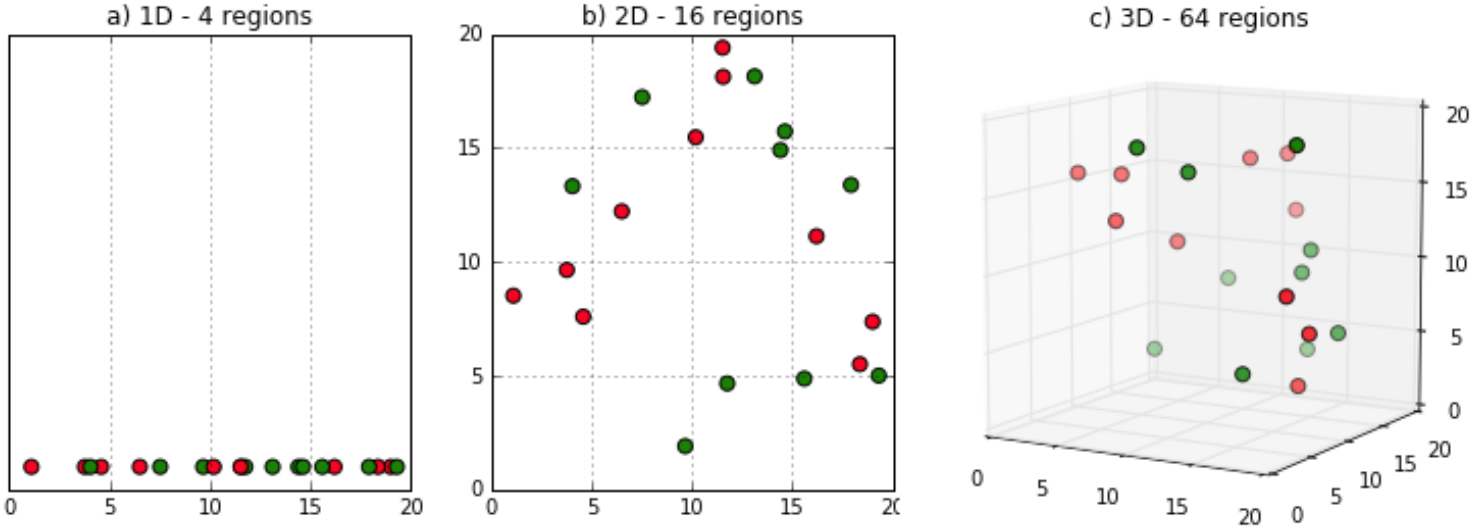
Wymiary

- Wymiar to zbiór informacji na temat mierzalnego zjawiska
- Odpowiednikiem wymiaru jest **zmienna**, **atrybut** albo **kolumna**
- Jeżeli wszystkie zjawiska są opisane wartościami numerycznymi, dającymi się przedstawić na skali ilorazowej, każdemu obiektowi odpowiada wtedy punkt w wielowymiarowej przestrzeni
- Zbiory dwu- i trójwymiarowe można przedstawić odpowiednio przy pomocy wykresu punktowego i wykresu trójwymiarowego rozproszonego
- W przypadku większej ilości wymiarów stosuje się wykresy „pairplots” w których zestawia się wymiary każdy z każdym w postaci macierzy. Przekątna macierzy najczęściej wykorzystywana jest do prezentacji rozkładów. Górny trójkąt może być wykorzystany do prezentacji/wizualizacji współczynników korelacji


DISTWATER
LOWLAND
RANGE
RESBASE
SLOPE
TEXTURE
WETNESS
0.058
-0.062
0.177
-0.114
0.031
-0.045
0.016
-0.156
0.022
0.08
0.397
0.478
0.876
0.614
-0.207
0.425
0.448
-0.336
0.552
-0.142
-0.027

Problem z wielowymiarowością

- Poza oczywistymi problemami z wizualizacją danych nadmiar wymiarów powoduje dodatkowe, mniej oczywiste problemy:
  - Wiele informacji staje się współdzielona pomiędzy zmiennymi (zmienne są częściowo skorelowane)
  - Wiele zmiennych jest bez znaczenia (nic nie wnosi)
  - Przy rosnącej liczbie wymiarów w przestrzeni wartości zaczynają dominować pustki, a dane stają się rozproszone



Przekleństwo wymiarowości

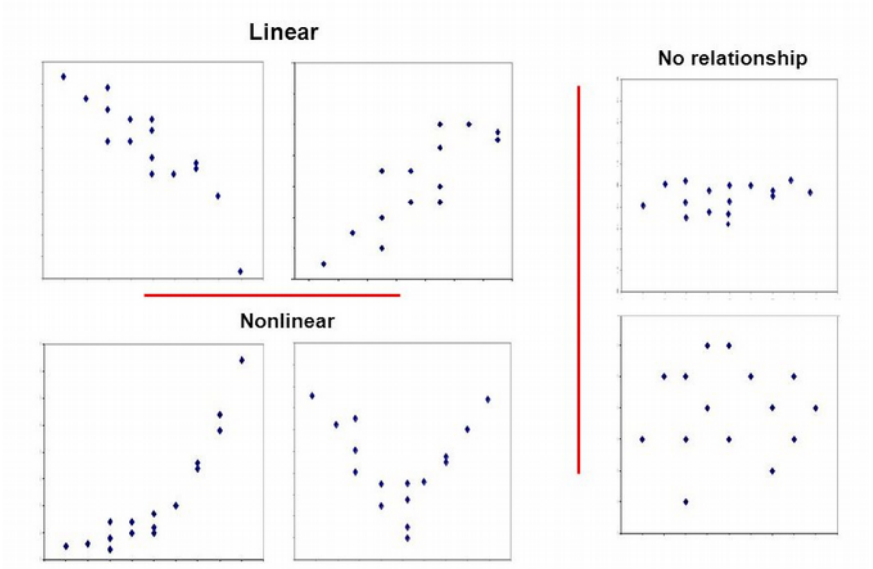
- **Kombinatoryka:** wraz ze wzrostem liczby wymiarów rośnie liczba możliwych kombinacji ich wartości. Prowadzi to do eksplozji kombinatorycznej. Jeżeli liczba wymiarów jest równa liczbie przypadków, nie ma możliwości zrobienia podsumowania
- **Próbkowanie:** wraz ze wzrostem wymiarów spada możliwość wyboru reprezentatywnej próby
- **Uczenie maszynowe/klasyfikacje:** przy dużej liczbie wymiarów wymagana jest bardzo duża liczba danych uczących aby mieć pewność, że każda klasa jest objęta wystarczającą zmiennością cech
- **Grupowanie:** Wraz ze wzrostem ilości zmiennych, wpływ pojedynczych zmiennych na niepodobieństwo zmiennych staje się coraz mniejszy
- **Wykrywanie anomalii:** przy nadmiernej liczbie wymiarów każdy przypadek staje się anomalią względem cech referencyjnych

Skąd się bierze nadmiar wymiarów

- Nie każda cecha, która jest łatwo mierzalna, jest cechą wpływającą **bezpośrednio na decyzję**. Na przykład na zakup urządzenia AGD wpływa jakość. Nie jest to pojęcie mierzalne ale raczej zestawienie kilku cech: funkcjonalność, trwałość, wyposażenie itp.
- Łatwiej jest zbierać informacje mierzalne niż niemierzalne
- Przed rozpoczęciem zbierania danych nie wiemy jakie informacje okażą się istotne
- Łatwiej jest zrezygnować ze zbędnej zmiennej niż zdobyć brakującą
- Często łatwiej jest zwiększyć liczbę mierzonych cech niż liczbę obiektów. Na przykład wykonanie wielu analiz z pobranej próby z dna oceanu jest prostsze/tańsze niż pobranie kolejnej próby

Relacje pomiędzy dwoma zmiennymi

- Relacja pomiędzy dwoma zmiennymi numerycznymi najprościej określić przy pomocy wykresu dwóch zmiennych. Wyróżniamy relacje
  - liniowe gdzie występuje zależność wyrażona w postaci równania liniowego y=ax+b
  - Nieliniowe, gdzie jest zależność wyrażona bardziej złożoną formułą lub ma charakter empiryczny
  - Brak relacji
- Do oceny relacji liniowej służą korelacja i kowariancja



Kowariancja

- Kowariancja – miara wspólnej zmienności dwóch zmiennych losowych, im bardziej dwie zmienne (wymiary) korespondują ze sobą tym wyższa wartość bezwzględna kowariancji. Znak pokazuje tendencję czy zależność jest w tym samym kierunku czy w kierunkach przeciwnych. Wartość kowariancji nie jest standaryzowana i wynosi (-∞,+∞), zależy od zakresu wartości zmiennych i z tego powodu jest trudna do interpretacji. Macierz kowariancji pokazuje zróżnicowanie pomiędzy zmiennymi

$$cov(X,Y)=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{2}(x_i-x_j)\cdot(y_i-y_j)=\frac{1}{n^2}\sum_{i}\sum_{j>i}(x_i-x_j)\cdot(y_i-y_j)$$

- Standaryzowana kowariancja to korelacja

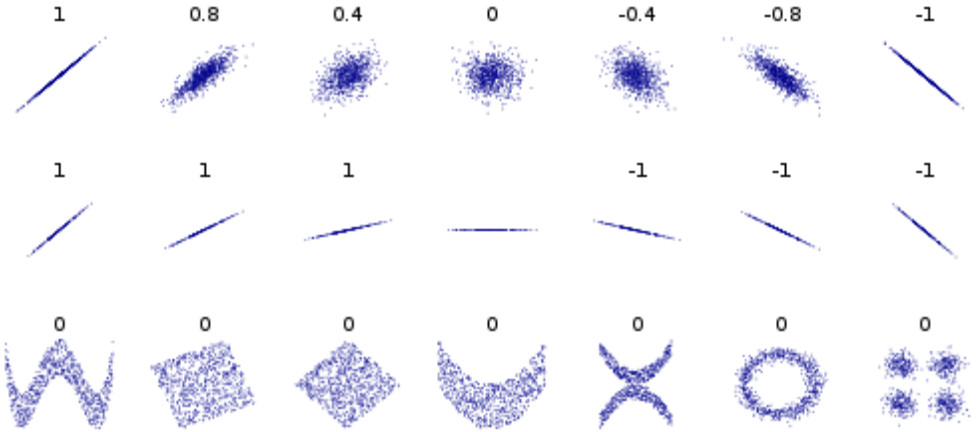
Korelacja

- Korelacja określa relację statystyczną – zależność pomiędzy dwoma zmiennymi, tj w jakim stopniu pomiędzy dwoma zmiennymi zachodzi relacja liniowa. Wartość korelacji zmienia się od [-1,1], gdzie 0 oznacza brak jakiejkolwiek zależności liniowej, natomiast 1 i -1 odpowiednio wskazują czy zmienne są zależne zgodnie, czy odwrotnie (antykorelacja).
- W praktyce korelacja jest bardzo użytecznym narzędziem określania zależności pomiędzy zjawiskami, jednakże w teorii nie implikuje takiej zależności. Dla przykładu wyniki z przedmiotu A na 1 roku studiów korelują z wynikami z przedmiotu B na 2 roku. Może to być jednak efektem tego że osoby dobrze się uczące uzyskują lepsze oceny i na odwrót.

$$cor(X,Y) = \frac{cov(X,Y)}{\sigma_x \sigma_y}$$

Siła relacji a współczynnik korelacji

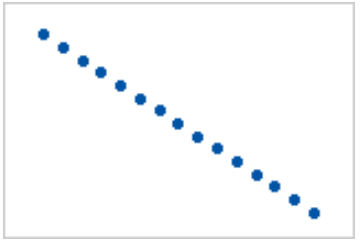
- Współczynnik korelacji nie zależy od nachylenia prostej regresji a jedynie od stopnia skupienia punktów wzdłuż prostej (odchylenia od modelu liniowego). Z tego powodu ani korelacja ani kowariancja nie nadają się do opisu relacji nieliniowych



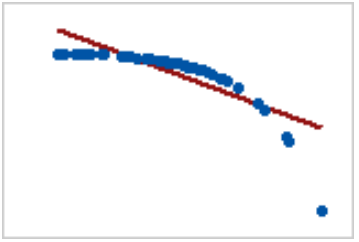
Korelacja Pearsona i Spearmana

- **Korelacja Pearsona** – bada liniową relację pomiędzy wartościami zmiennych, w jakim stopniu wartość jednej zmiennej zmienia się proporcjonalnie do drugiej
- **Korelacja Spearmana** – jest korelacją rang wartości, sprawdza jedynie tendencję zmian a nie dokładne wartości. Z tego powodu pokazuje monotoniczną zgodność tendencji a nie wartości

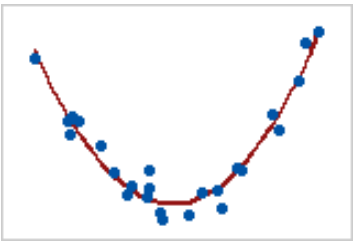
P=-1 S=-1



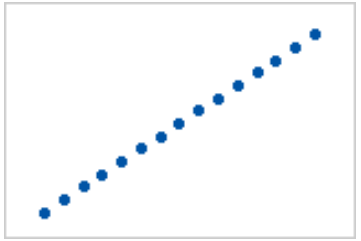
P=-0.8 S=-1



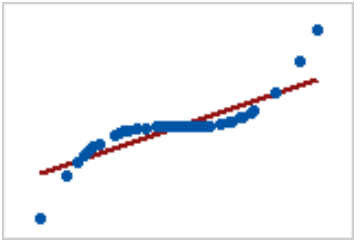
P=0 S=0



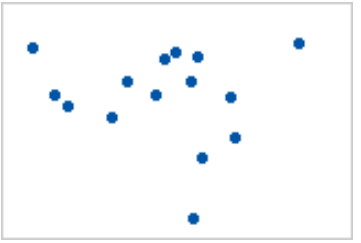
P=1 S=1



P=0.85 S=1

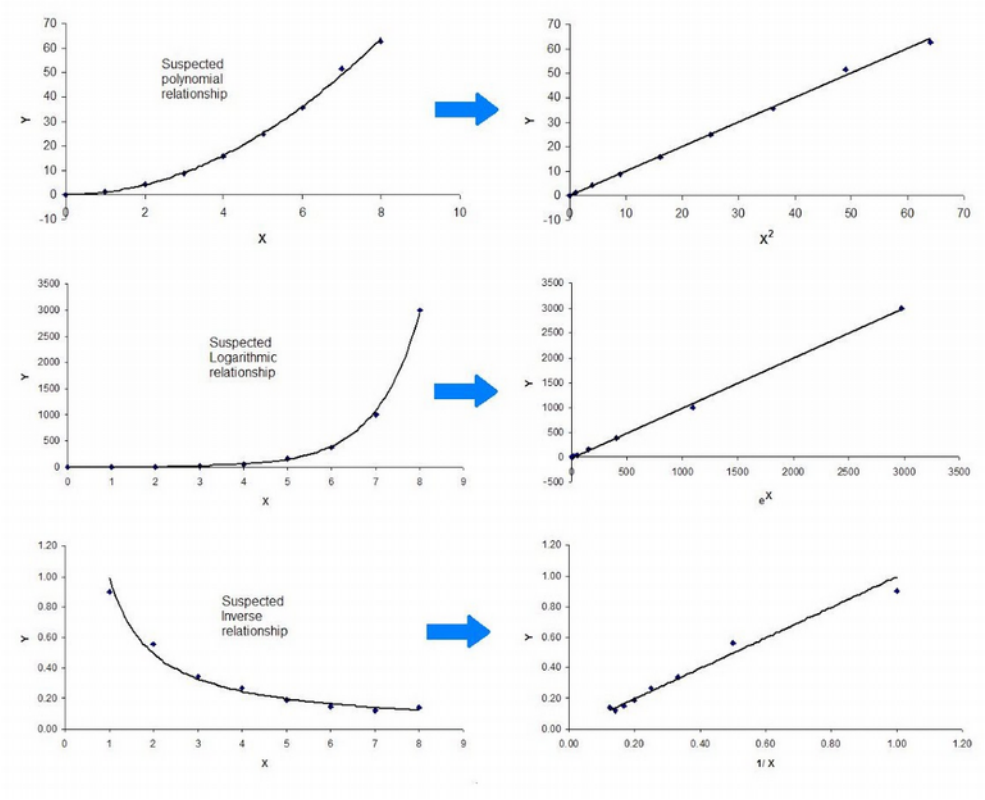


P=0.03 S=0.03



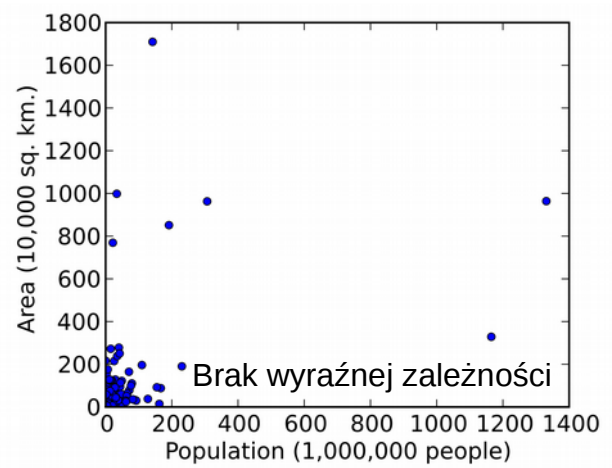
Zależności nieliniowe

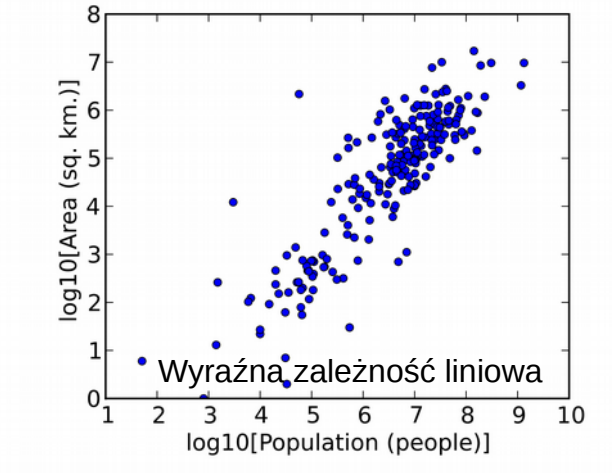
- Pomiędzy zmiennymi mogą zachodzić relacje nieliniowe (tj takie których nie da się opisać linią prostą). Sytuacja jest spowodowana tym że zmienne mają różne rozkłady. Współczynnik korelacji Spearmana pomiędzy takimi zmiennymi będzie miał zaniżone wartości. W takiej sytuacji należy dokonać transformacji rozkładów, aby można było określić korelację



Relacja między zmiennymi a typ rozkładu

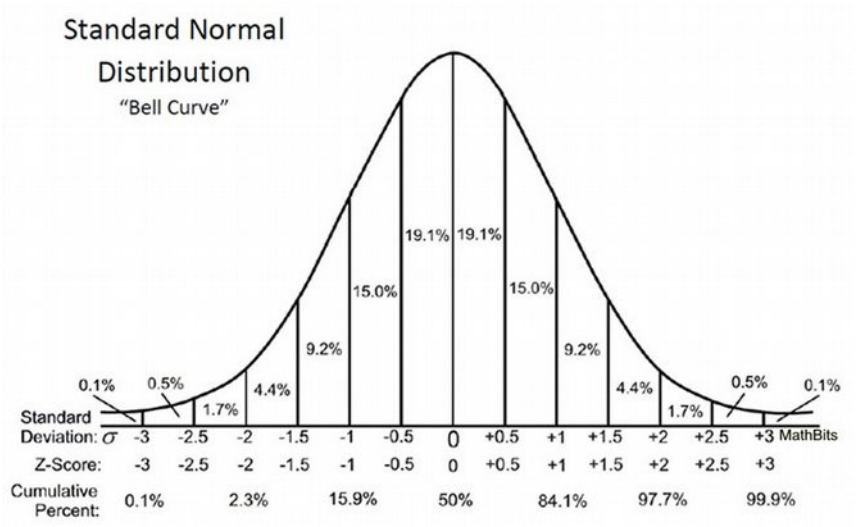
- Idealna sytuacja ma miejsce wtedy, kiedy każda zmienna ma rozkład normalny
- Tylko dla rozkładów normalnych wartość oczekiwana ma interpretację
- Zaletą rozkładu normalnego jest uproszczenie obliczeń matematycznych: momentów centralnych (średnia, odchylenie), współczynnika korelacji, wskaźnika regresji
- Aby prawidłowo badać zależności między zmiennymi należy je sprowadzić do rozkładu normalnego

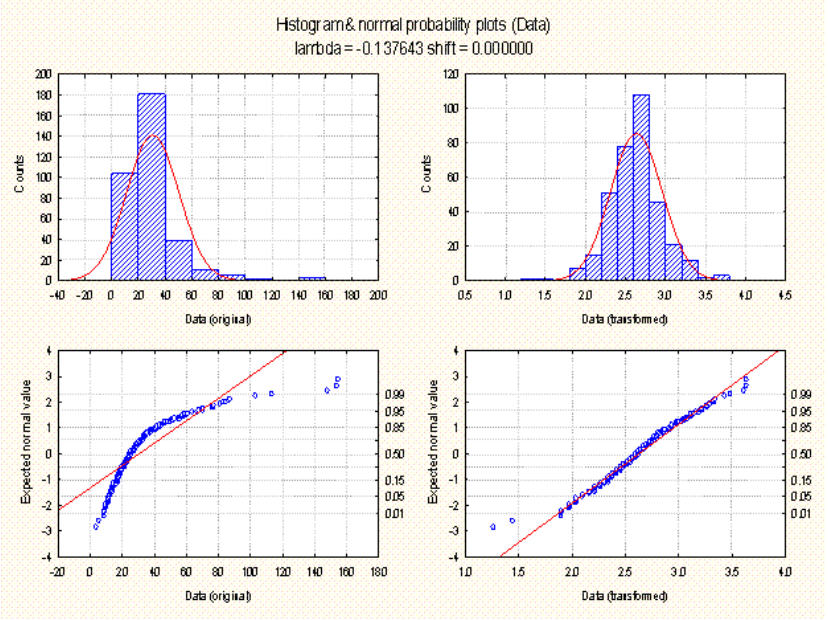




Rozkład normalny a skośność

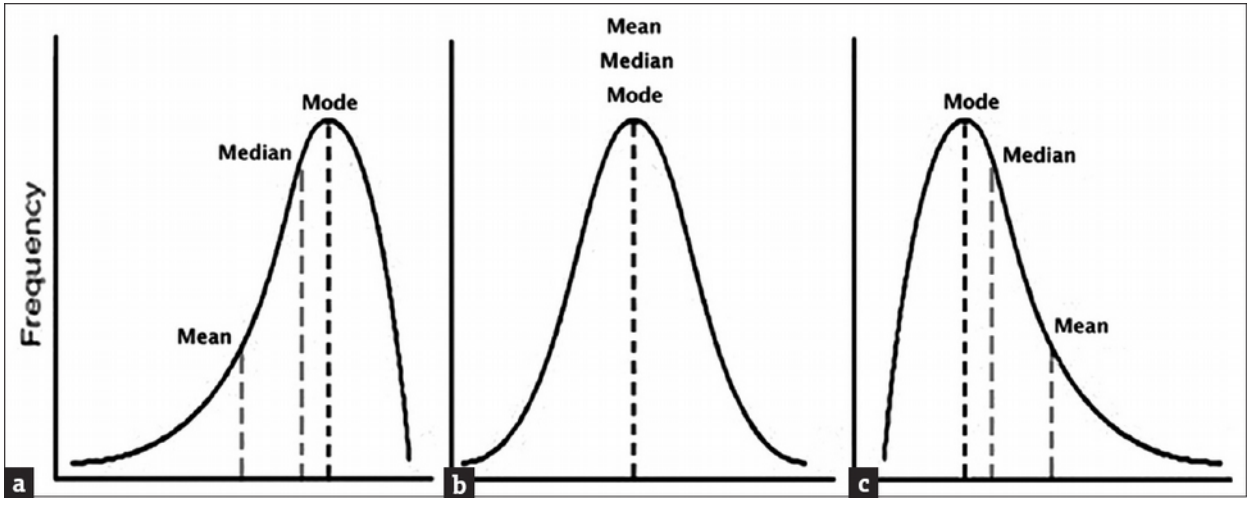
- Rozkład ciągły, używany w wielu dziedzinach aby reprezentować zmienne losowe o nieznanym rozkładzie
- Cechą idealnego rozkładu normalnego jest jednakowa wartość średniej, mediany i mody.
- Dla rozkład standaryzowanego, średnia wynosi 0 odchylenie standardowe 1, skośność 0 a kurtoza 3 (momenty centralne)

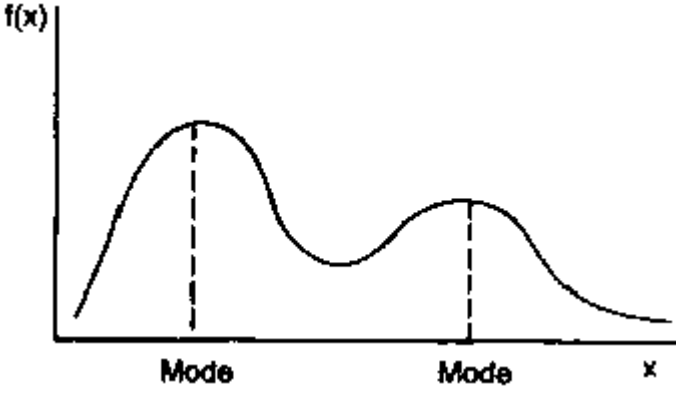




Rozkłady skośne i wielomodalne

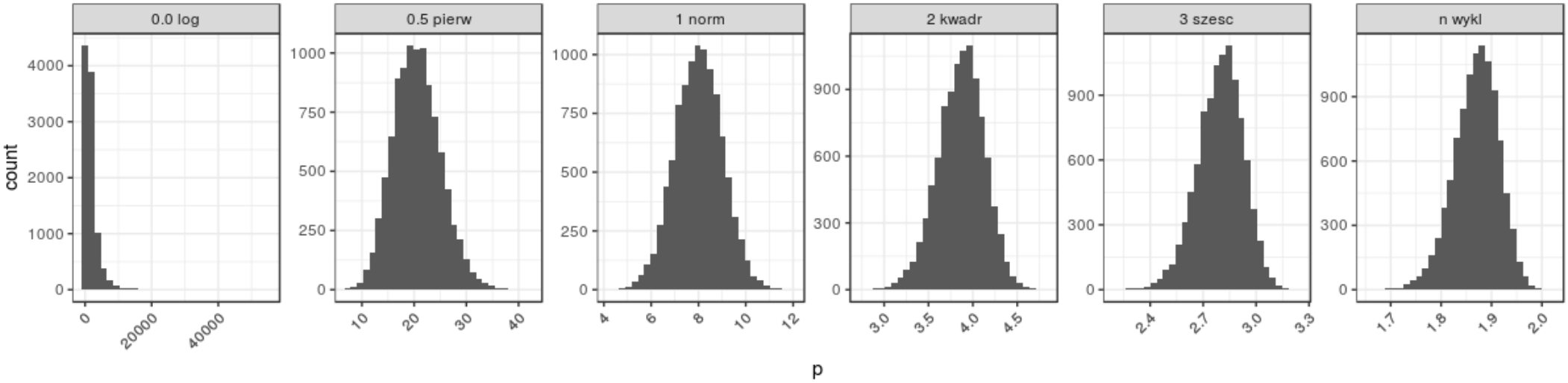
- Rozkłady skośne to takie, gdzie średnia i mediana nie pokrywają się ze sobą.
- Rozkłady wielomodalne mają co najmniej dwa maksima w swoim rozkładzie. Wielomodalność wskazuje że zmienna jest mieszaniną dwóch lub więcej populacji (więcej: modele mieszane)
- Wielomodalność często **jest korzystna** w procesie uczenia maszynowego. Więcej: modele mieszane





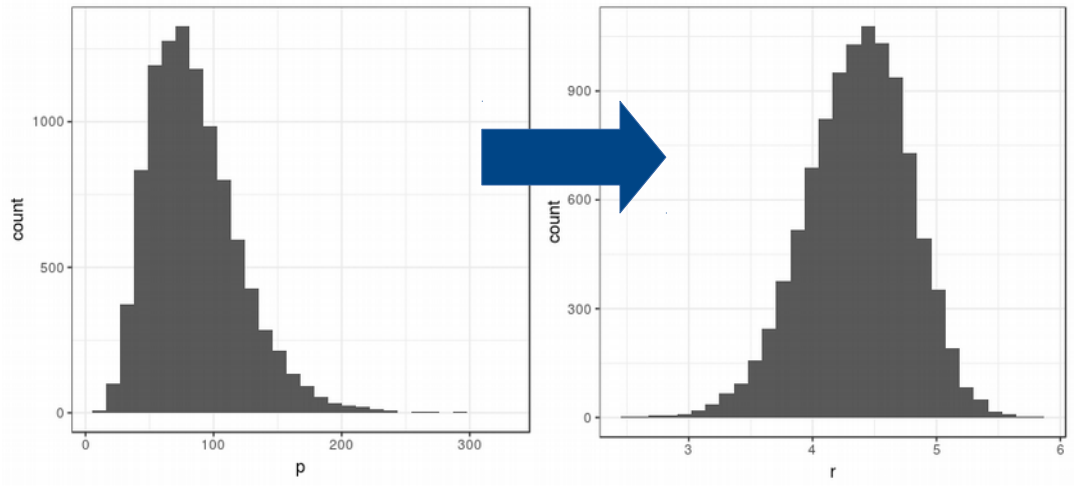
Transformacje rozkładów

- W celu modyfikacji rozkładu do formy zbliżonej do normalnego stosuje się transfromacje. W zależości od kształtu rozkładu są to transformacje:
- Dla lewoskośnych(mn):
  - Logarytmiczno-normalna
  - Pierwiastkowa
- Dla prawoskośnych (mn):
  - Kwadratowa, sześcienna
  - Wykładnicza



Wybór właściwej metody

- Znalezienie właściwej formy transformacji nie jest proste, gdyż rzeczywiste rozkłady rzadko są zgodne z rozkładami teoretycznymi i mają formę pośrednią np. pomiędzy rozkładem logarytmicznym a pierwiastkowym. W takiej sytuacji transformacja logarytmiczna zamieni nam jedynie rozkład lewoskośny na prawoskośny
- W takiej sytuacji stosuje się transformacje uniwersalne: BoxCox i Yeo_Johnson
- Transformacje uniwersalne stosuje się w celu automatyzacji procesu, na przykład przy dużej ilości zmiennych



Transformacje uniwersalne

- Transofmracja BoxCox – Jest to transformacja potęgowa, gdzie orginalna zmienna podnoszona jest do wybranej potęgi λ, w praktyce λ = (-4,4) w zależności od wartości odpowiada transformacjom:
  - -2 odwrotna kwadratowa
  - -1 odwrotna
  - 0 – logarytmiczna
  - 0.5 pierwiastkowa
  - 1 brak
  - Kwadratowa itp..

$$T(x)=\frac{x^{\lambda}-1}{\lambda};\lambda\neq 0$$
$$T(x)=\ln(x);\lambda=0$$

- Bardziej skuteczna jest transformacja Yeo-Johnsona, która może transformować również wartości ujemne

$$y_i^{(\lambda)}=\begin{cases}((y_i+1)^{\lambda}-1)/\lambda & \text{if } \lambda\neq 0, y\geq 0\\ \log(y_i+1) & \text{if } \lambda=0, y\geq 0\\ -[(-y_i+1)^{(2-\lambda)}-1]/(2-\lambda) & \text{if } \lambda\neq 2, y<0\\ -\log(-y_i+1) & \text{if } \lambda=2, y<0\end{cases}$$

Skalowanie

- Skalowanie to przekształcenie rozkładu z jednego zakresu zmienności w drugi, bez zmiany kształtu rozkładu
- Ogólna formuła skalowania przyjmuje postać:

$$\frac{x - shift}{scaling}$$

Gdzie: shift nowe położenie zera, a scaling parametr skalujący rozkład. W praktyce stosuje się dwa rodzaje skalowania: skalowanie oraz standaryzację albo z-score

Przeskalowanie i standaryzacja

- **Skalowanie** do przedziału [0,1] lub innego przedziału (rescaling) stosuje się głównie w calu bezwzględnego ujednolicenia przedziałów – na przykład w celu wizualizacji jednolitą skalą. Określa się formułą, gdzie przesunięcie to minimalna wartość zmiennej a współczynnik skalujący to różnica wartości.

$$\frac{x - min(x)}{max(x) - min(x)} \times (max_{new} - min_{new}) + min_{new}$$

- **Standaryzację** stosuje się w celu ujednolicenia danych znajdujących się w różnych zakresach i skalach wartości, tak aby zmienne o największej rozpiętości nie dominowały wyniku. Wartość przesunięcia to średnia z rozkładu a współczynnik skalujący to odchylenie standardowe. Po standaryzacji wszyskie zmienne mają średnią = 0 i odchylenie standardowe =1. Nie zmienia się geometria rozkładu

$$\frac{x - \bar{x}}{\sigma}$$

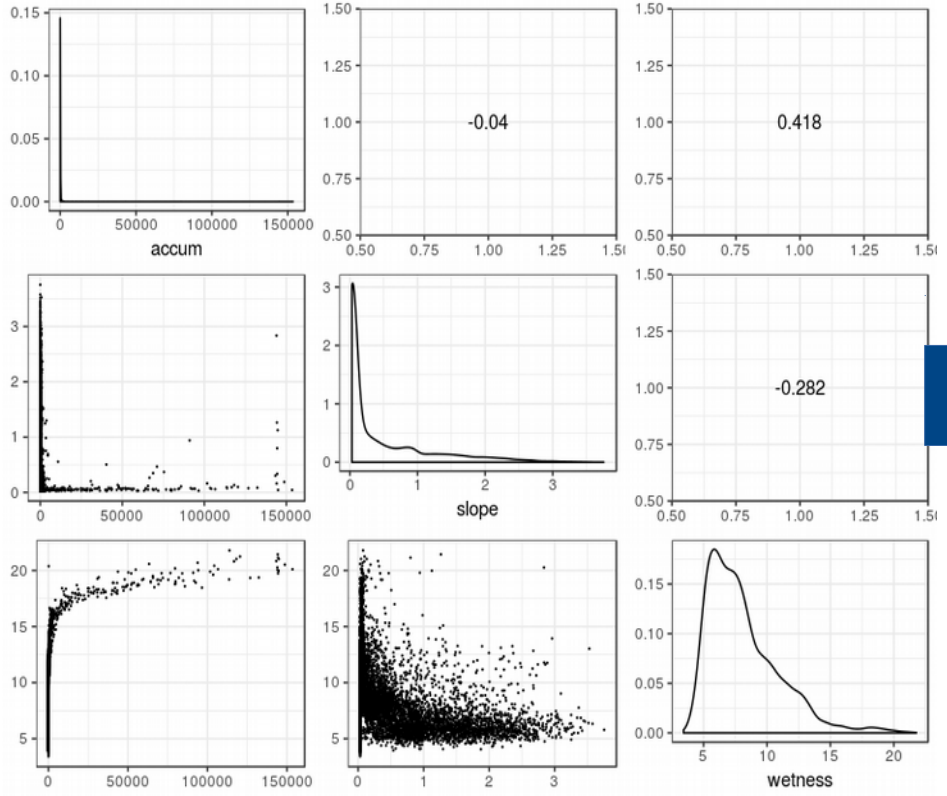
Relacje pomiędzy zmiennymi

- Jednym z celów transformacji danych jest ujawnienie relacji liniowych pomiędzy zmiennymi
- W rzeczywistym świecie, zwłaszcza w naukach przyrodniczych zmienne są ze sobą w relacji, zależą wzajemnie od siebie
- Na przykład:

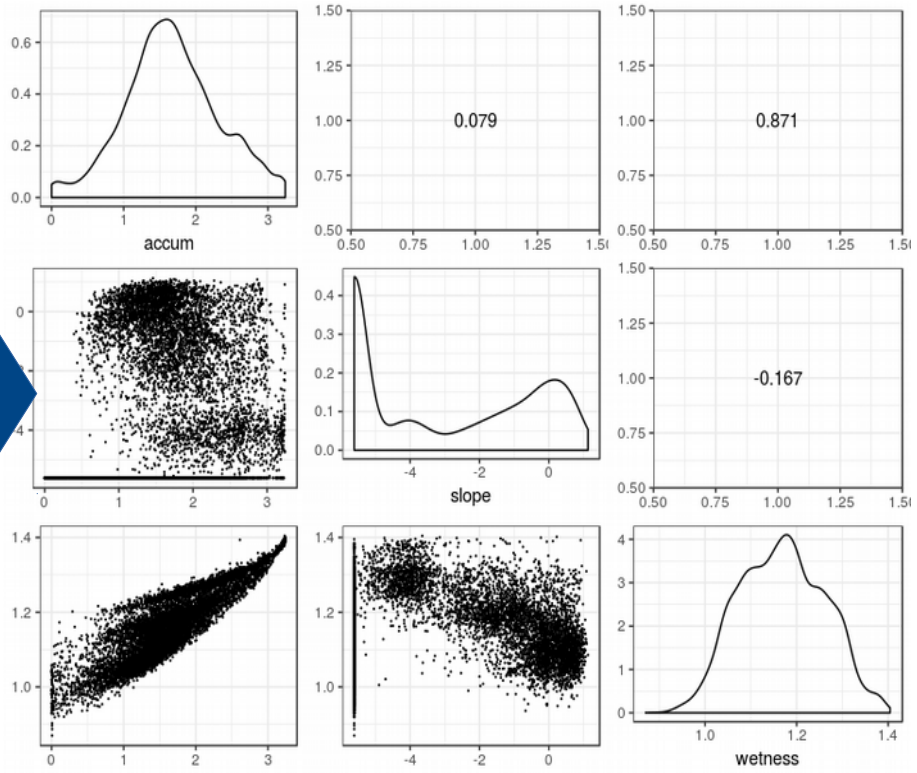
  ilość_opadów ~ odległość_od oceanu + wys_nad_poziom_morza

- Relacja nie zakłada czy jest to sprzężenie dodatnie czy ujemne

Ukryte relacje

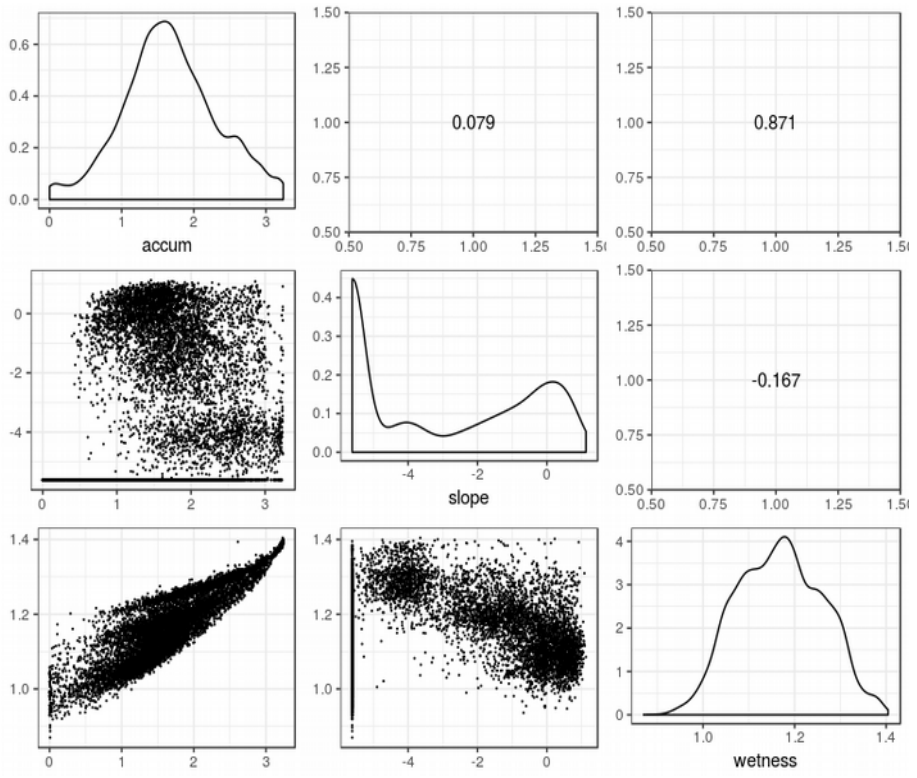


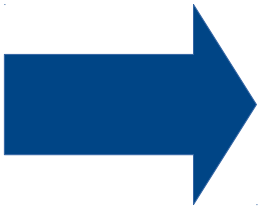
BoxCox

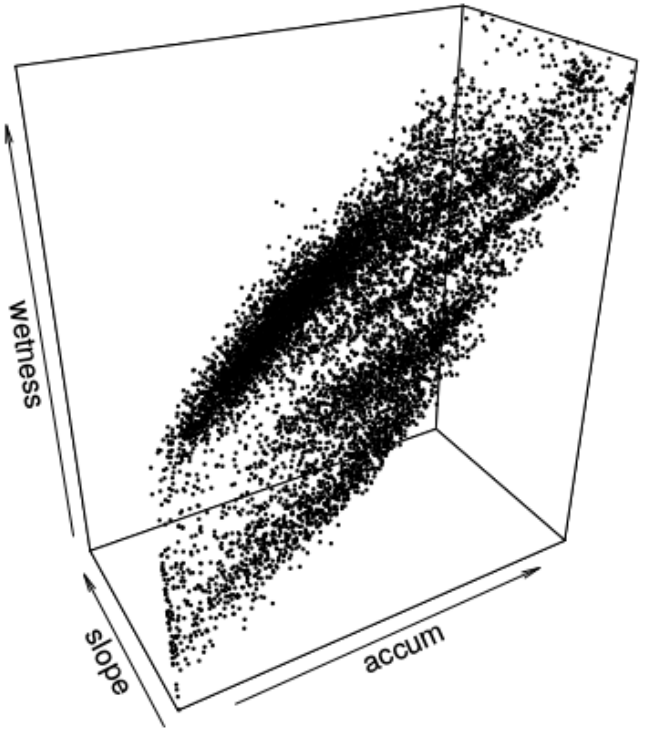


$$wetness = \ln \frac{accum}{slope}$$

Mapowanie do 3 wymiarów

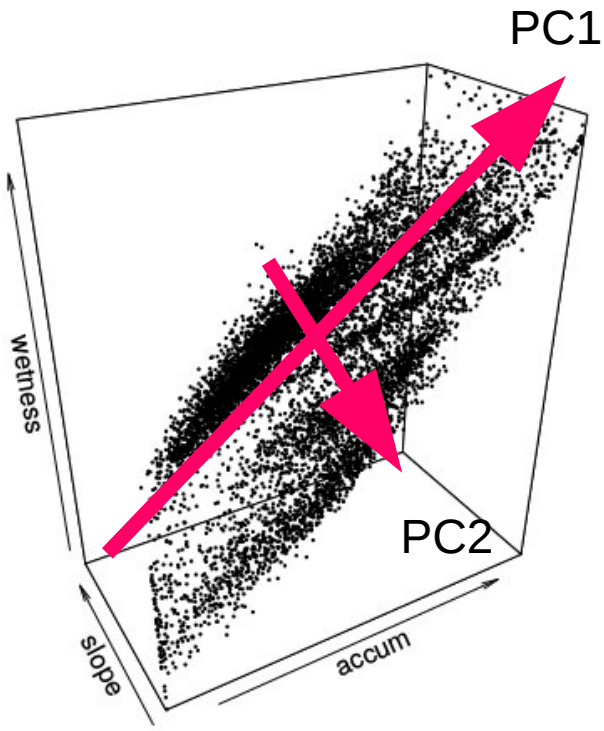






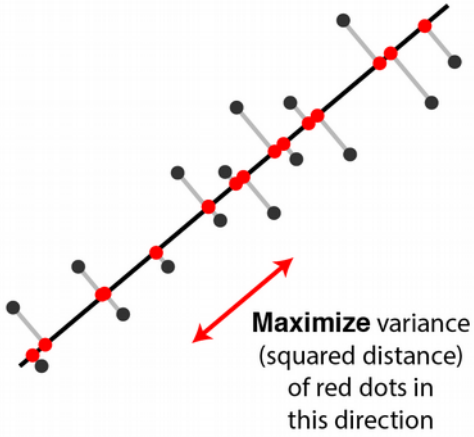
Składowe główne

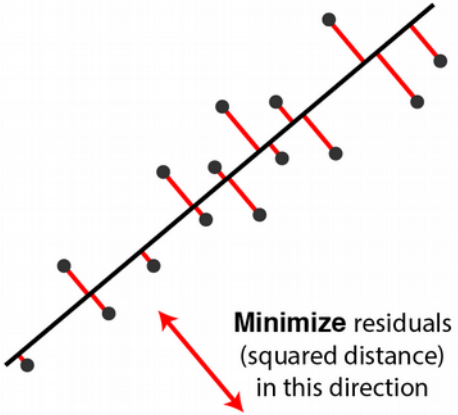
- Analiza składowych głównych ma dwa cele:
  - Redukcja liczby wymiarów (zmiennych) opisujących zjawisko, poprzez odrzucenie ostatnich, najmniej istotnych składowych
  - Połączenie zmiennych wzajemnie od siebie zależnych w nowe niezależne byty, dla których można przeprowadzać nową interpretację

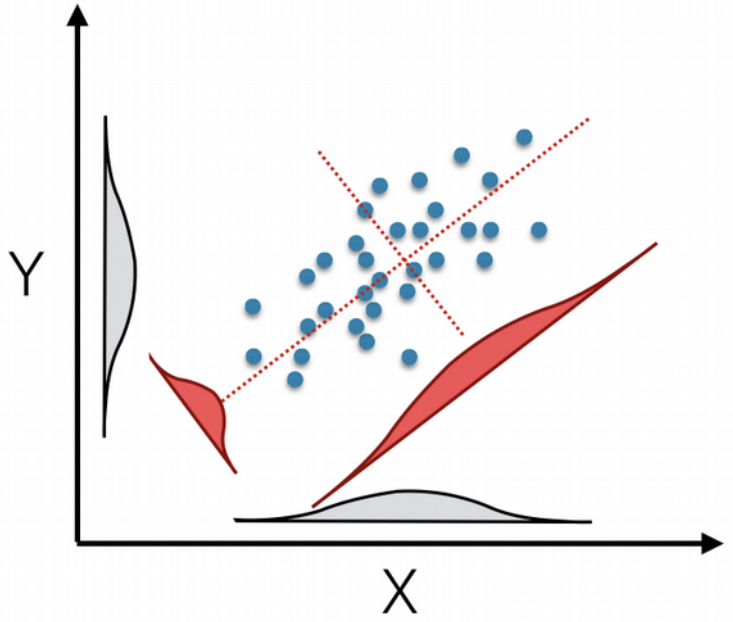


Składowe główne

- Zakładając że zbiór danych tworzy chmurę punktów w n- wymiarowej przestrzeni analiza składowych głównych ma na celu taki obrót układu współrzędnych, aby maksymalizować w pierwszej kolejności wariancję 1 współrzędnej, potem 2 współrzędnej itp. Nowe współrzędne nazywamy ładunkami, które tworzą nową przestrzeń, gdzie najwięcej zmienności wyjaśniają pierwsze czynniki

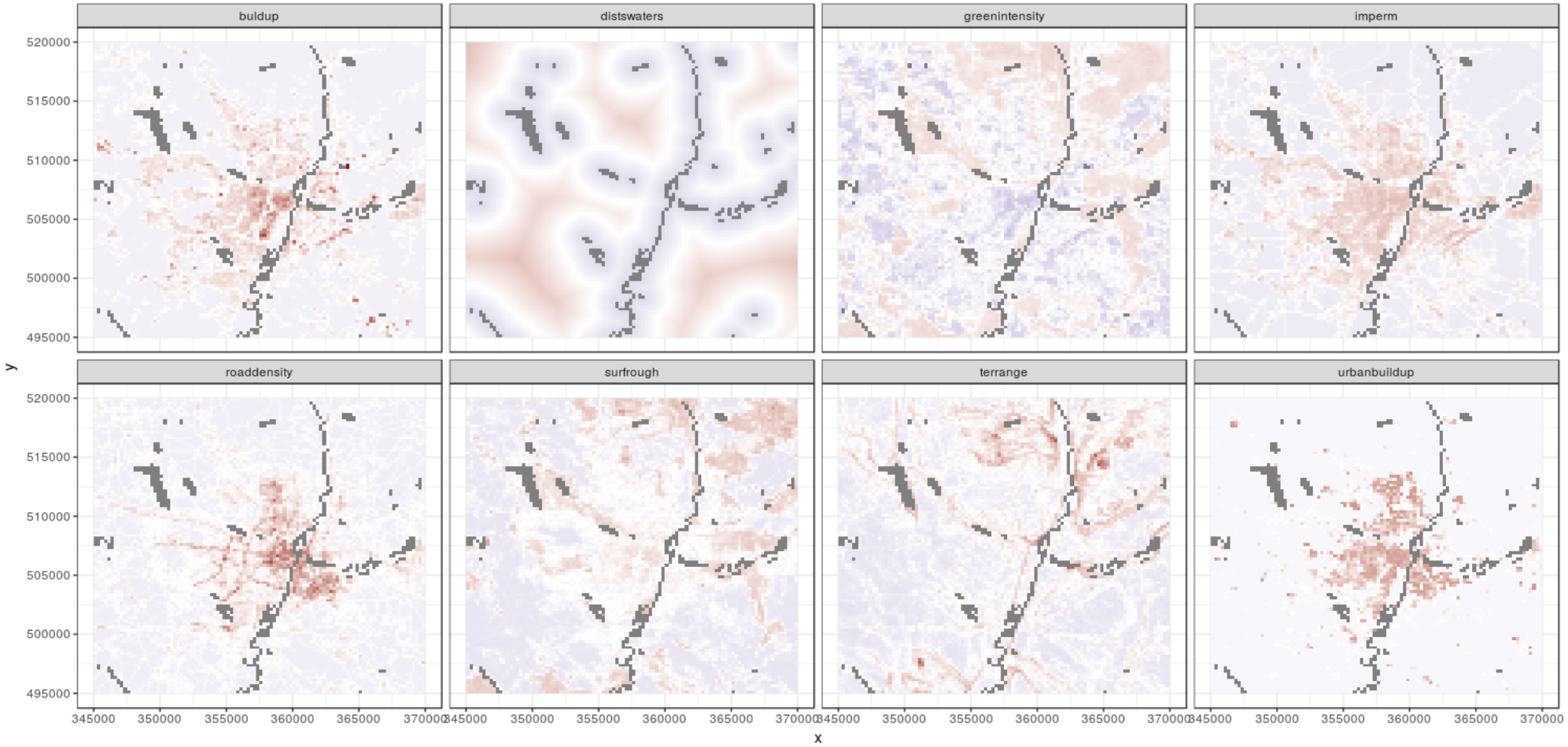


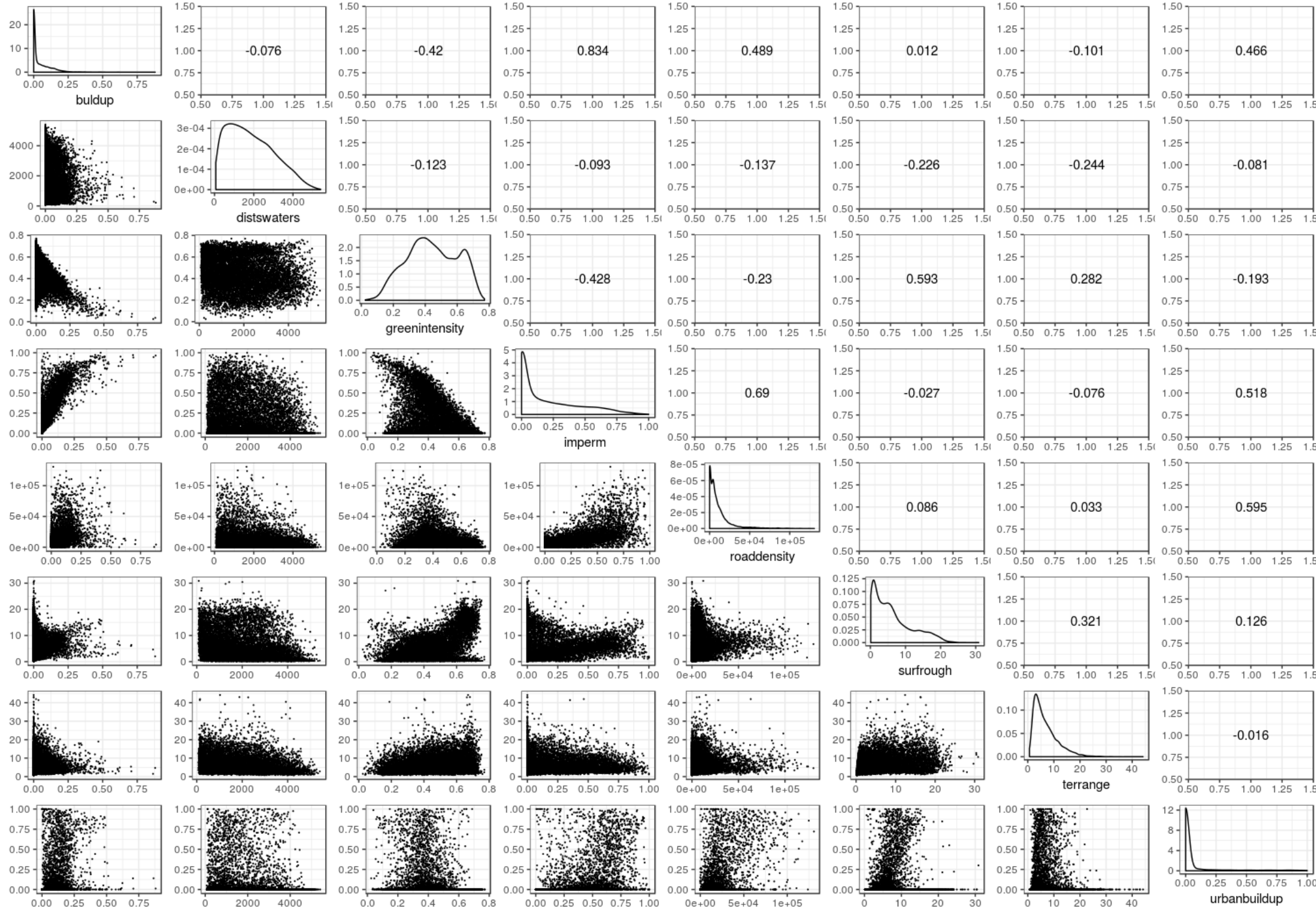




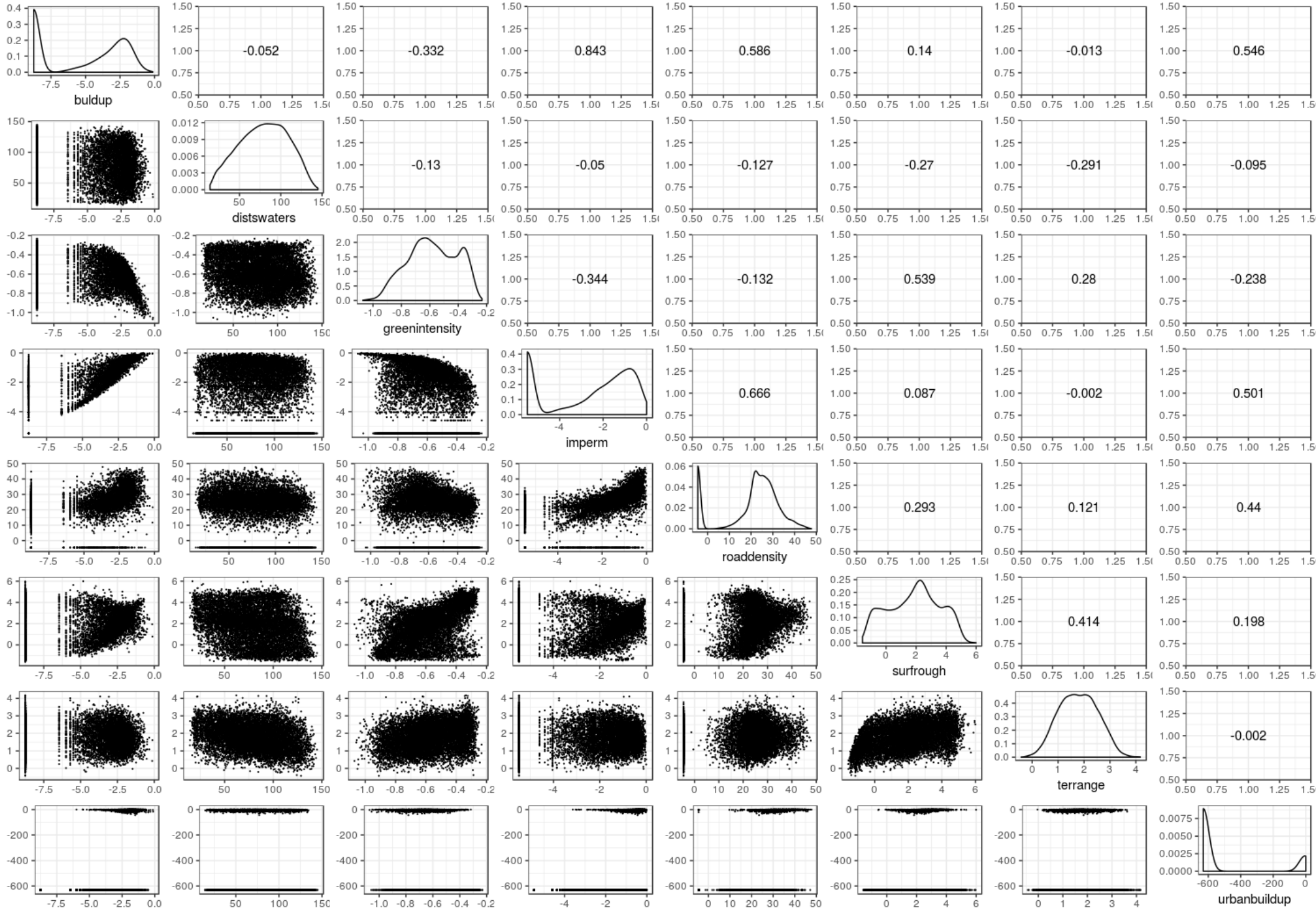
Analiza danych geoprzestrzennych

- Przykład obejmuje 8 zmiennych opisujących pokrycie terenu dla miasta Poznania





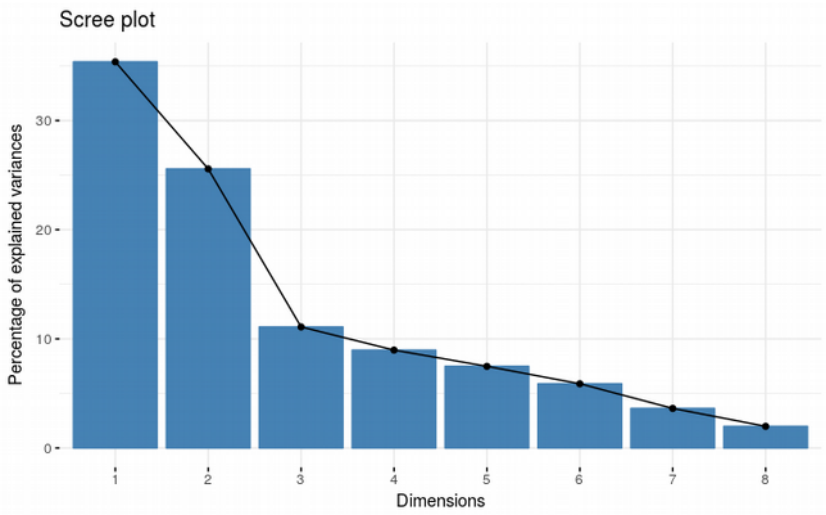
Dane przed transformacją

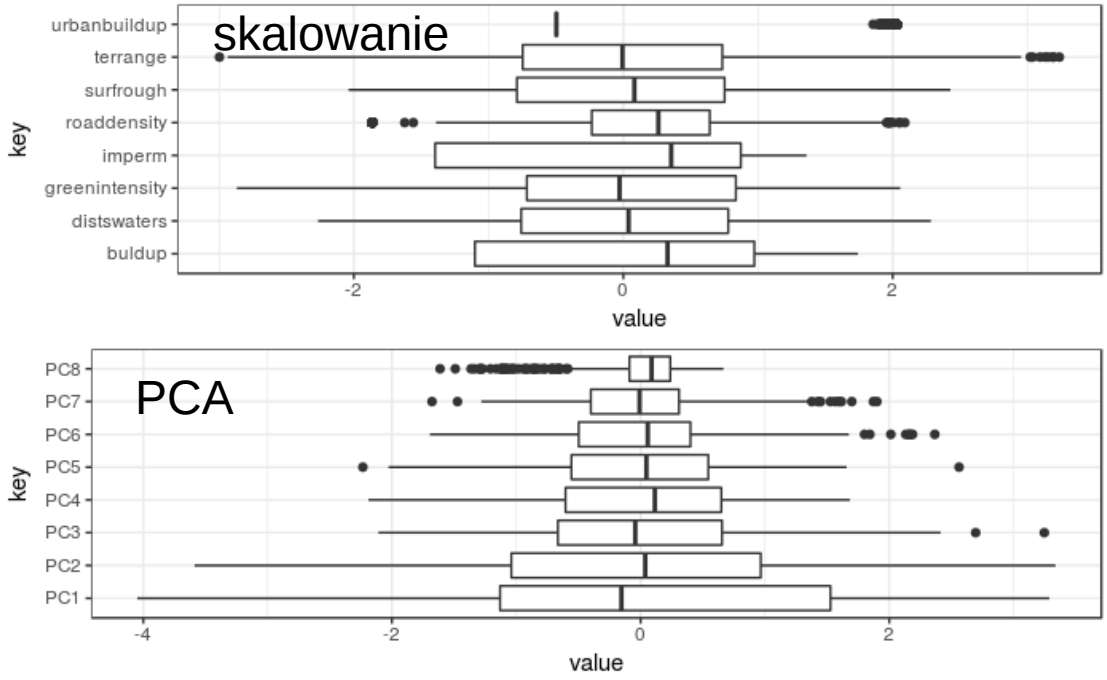


Dane po transformacji

Ważność zmiennych

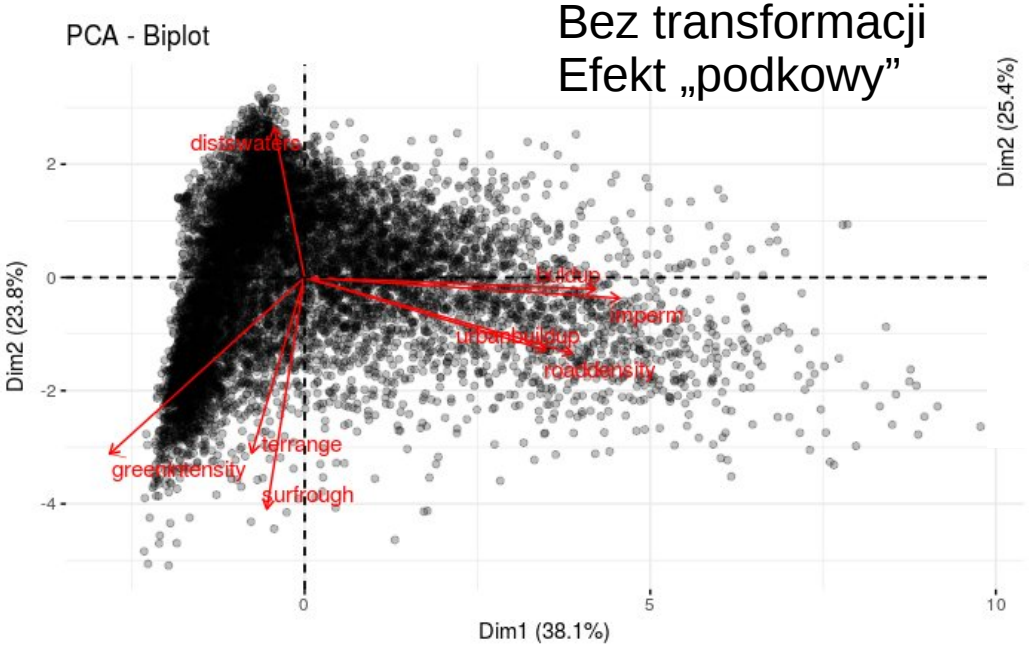
- W przypadku składowych głównych, pierwsze zmienne mają znacząco wyższy udział w wyjaśnianiu zmienności, niż w przypadku zmiennych skalowanych, gdzie udział każdej ze zmiennych jest taki sam. Analiza udziału wariancji pozwala wybrać ile zmiennych zostanie użytych do dalszej analizy (reguła kciuka – 80% wyjaśnienia)



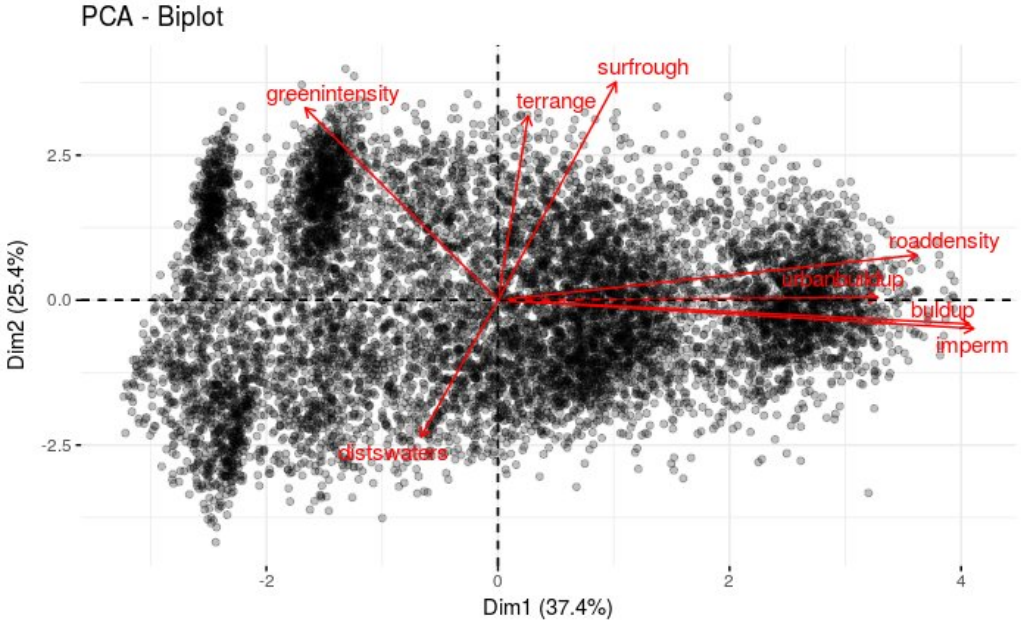


Gradienty składowych

- Oryginalne zmienne – zjawiska które możemy mierzyć nie zawsze są czynnikiem wyjaśniającym badany proces. Na przykład przy decyzji o zakupie produktu, konsument rzadko kieruje się ceną ani jakością, najczęściej jest to złożony parametr – gradient: stosunek jakości do ceny



Bez transformacji
Efekt „podkowy”



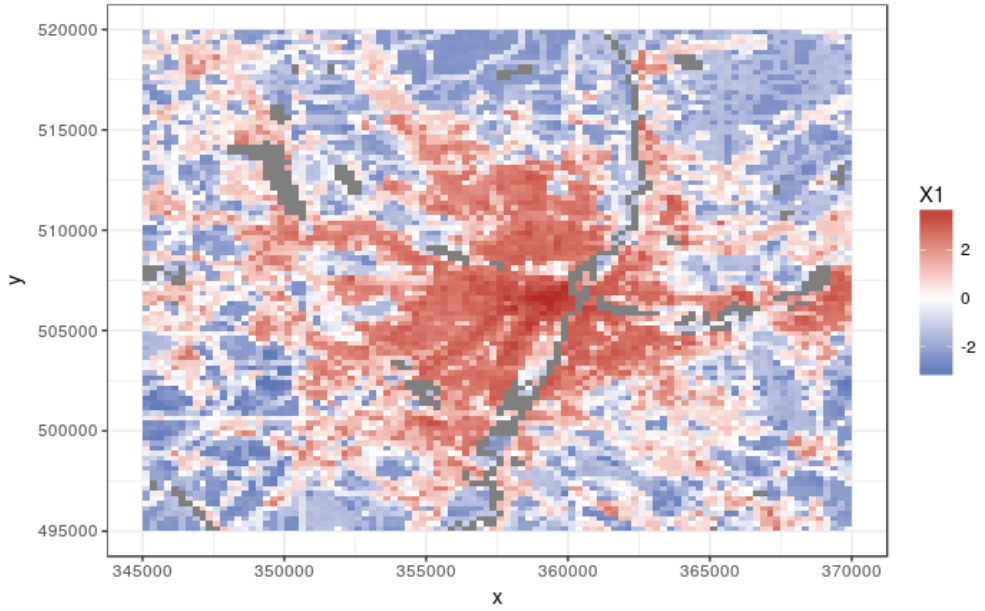
Po transformacji
Wyraźne skupienia

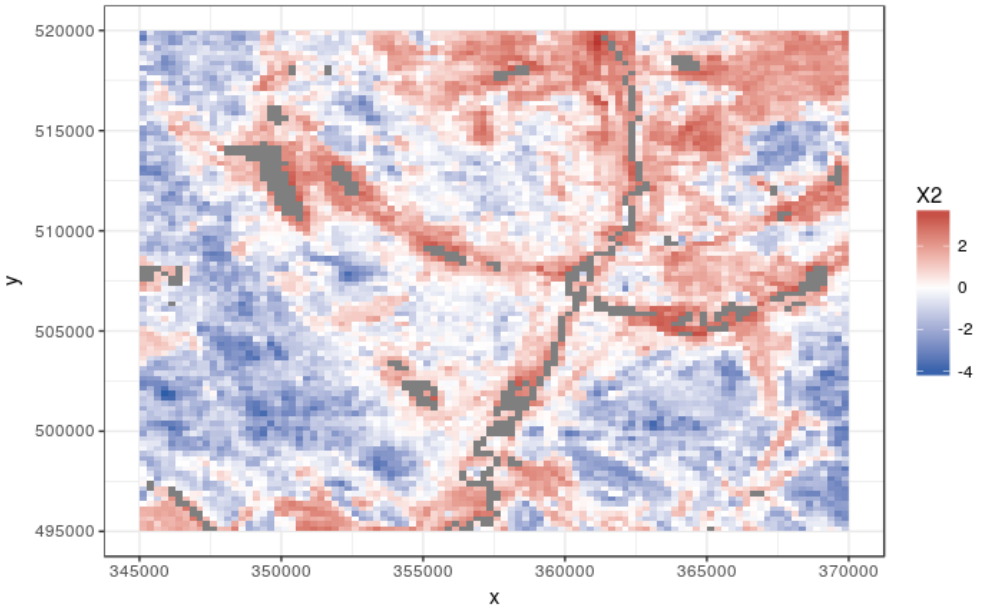
Ocena Gradientów

- Analiza składowych głównych ujawnia takie gradienty, które w wielowymiarowej przestrzeni są trudne lub wręcz nie są możliwe do rozpoznania
- Po wykonaniu transformacji przy pomocy wykresów dwuwarstwowych (bi-plots) można analizować relację oryginalnych zmiennych w stosunku do wybranych składowych i na tej podstawie określać rodzaj gradientu jaki dana zmienna reprezentuje
- Obecność lub brak transformacji zmiennych do rozkładu normalnego ma znaczenie dla możliwości oceny i rozpoznania gradientów

Interpretacja kartograficzna

- Składowe główne dla danych przestrzennych można wizualizaować na mapie
- Okazuje się że za 60% zmienności cech pokrycia terenu na obszarze Poznania może być wyjaśnione przy pomocy dwóch zmiennych: zabudowy i zróżnicowania naturalnego pokrycia terenu

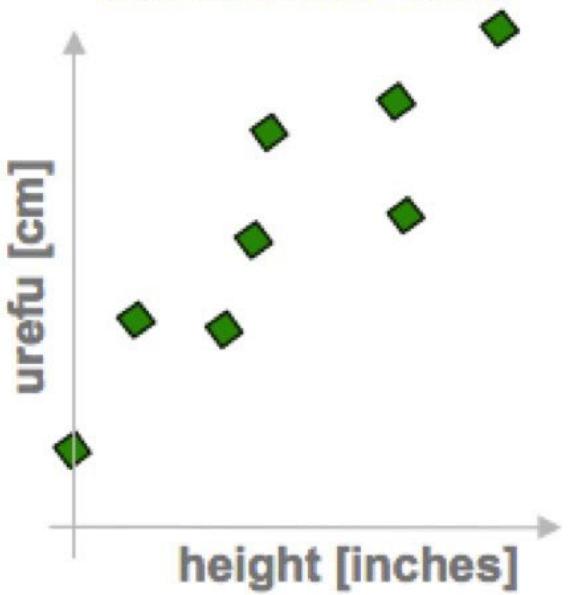




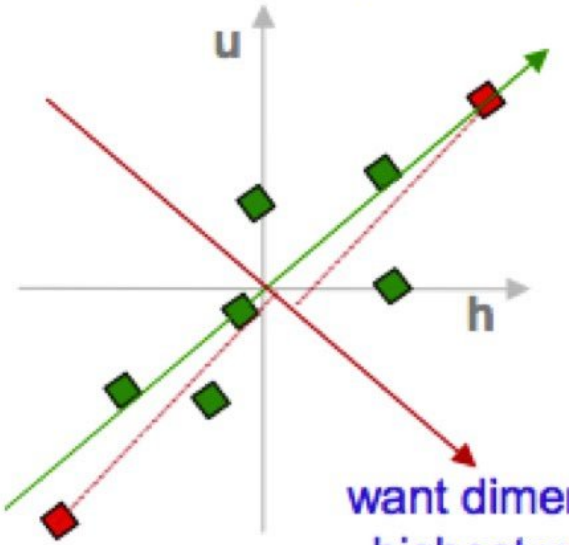
PCA in a nutshell

1. correlated hi-d data

("urefu" means "height" in Swahili)



2. center the points



want dimension of highest variance

3. compute covariance matrix

$$\begin{array}{c} \\ h \\ u \end{array}\begin{array}{c} \begin{array}{cc} h & u \end{array} \\ \begin{pmatrix} 2.0 & 0.8 \\ 0.8 & 0.6 \end{pmatrix} \end{array} \rightarrow \mathrm{cov}(h,u) = \frac{1}{n}\sum_{i=1}^{n} h_i u_i$$

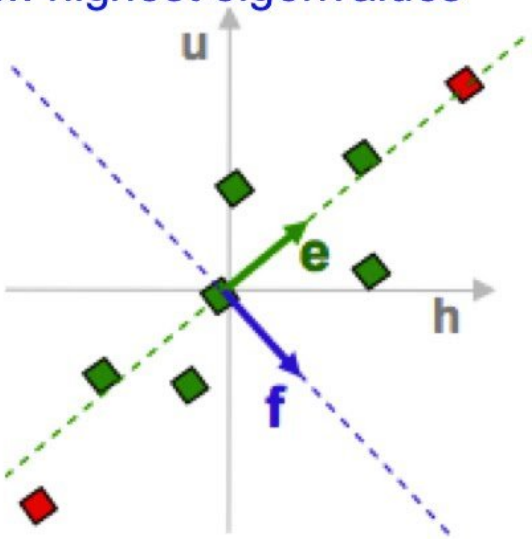
4. eigenvectors + eigenvalues

$$\begin{pmatrix} 2.0 & 0.8 \\ 0.8 & 0.6 \end{pmatrix} \begin{pmatrix} e_h \\ e_u \end{pmatrix} = \lambda_e \begin{pmatrix} e_h \\ e_u \end{pmatrix}$$

$$\begin{pmatrix} 2.0 & 0.8 \\ 0.8 & 0.6 \end{pmatrix} \begin{pmatrix} f_h \\ f_u \end{pmatrix} = \lambda_f \begin{pmatrix} f_h \\ f_u \end{pmatrix}$$

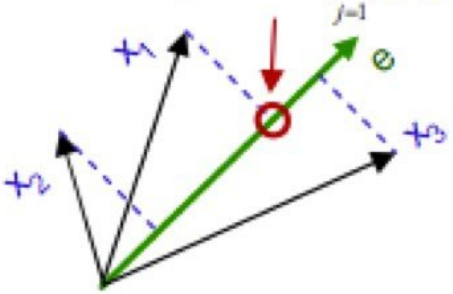
`eig(cov(data))`

5. pick m<d eigenvectors w. highest eigenvalues

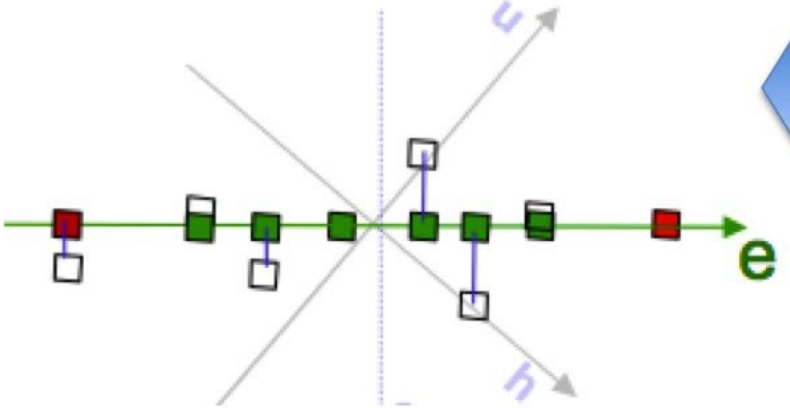


6. project data points to those eigenvectors

$$x_e' = x^T e = \sum_{j=1}^{a} x_{ij} e_j$$

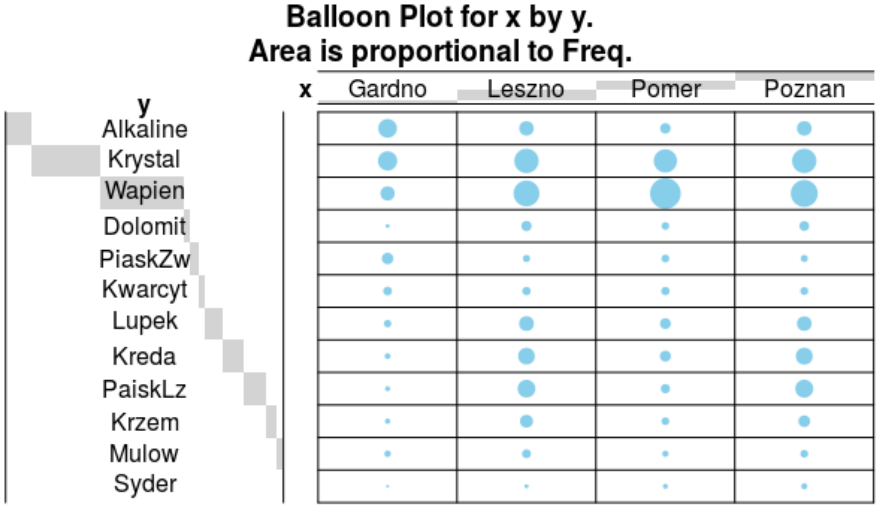


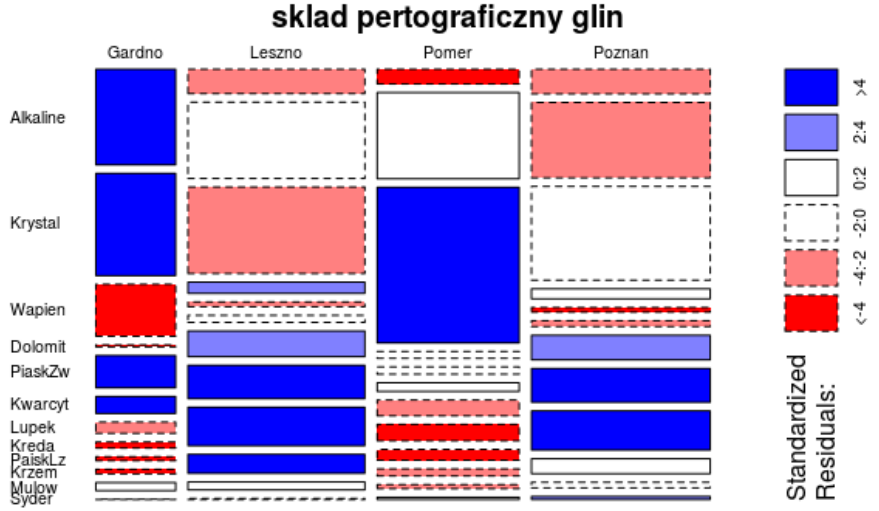
7. uncorrelated low-d data



Analiza korespondencji

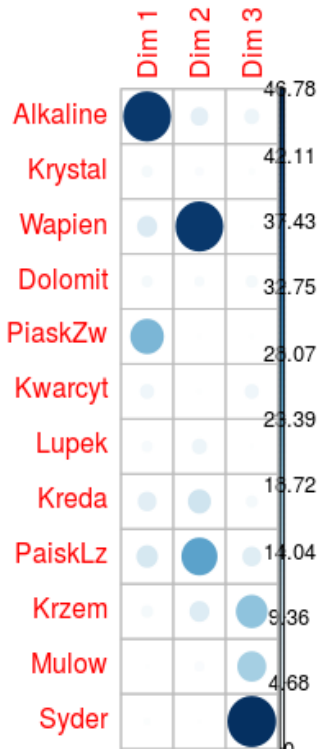
- Analiza korespondencji jest odpowiednikiem analizy składowych głównych stosowaną dla danych kategoryzacyjnych, gdzie dane są zorganizowane w postaci tzw tabeli kontyngencji (przypadkowości) i pozwala wizualizować takie dane w formie dwuwymiarowej
- Jako przykład posłużą typy gazików zebranych dla różnych faz ostatniego zlodowacenia



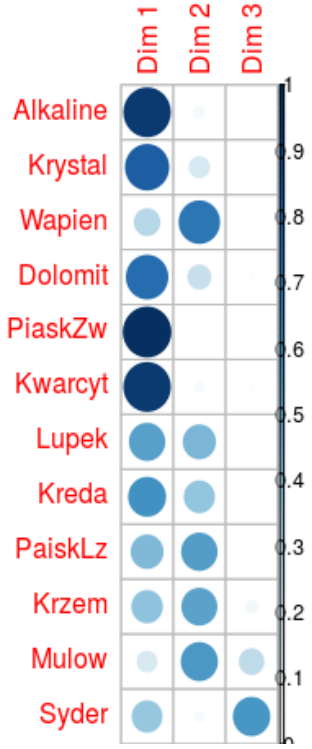


Składowe analizy korespondencji

- Podobnie jak PCA analiza zwraca składowe (wymiary) udział w wyjaśnianiu u jakość reprezentacji składowej przez cechę



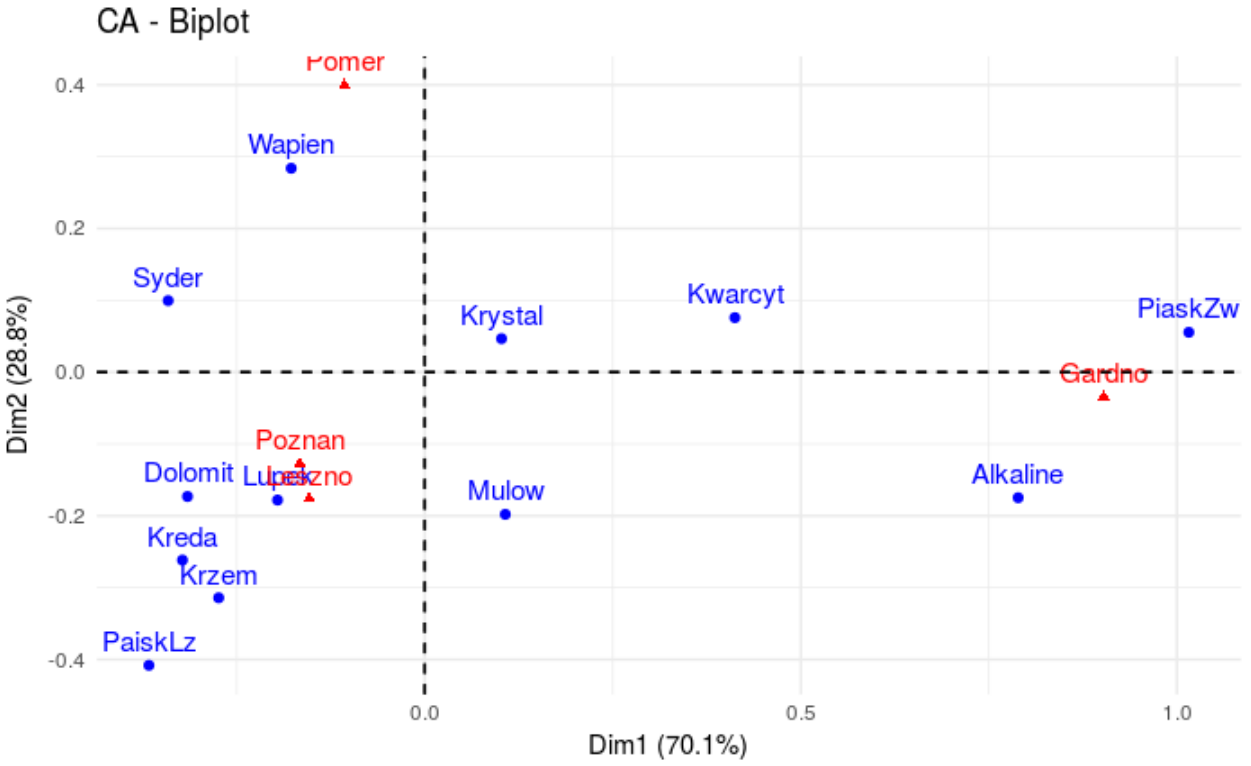
Udział składowej w wyjaśnianiu



Reprezentacja składowej przez cechę

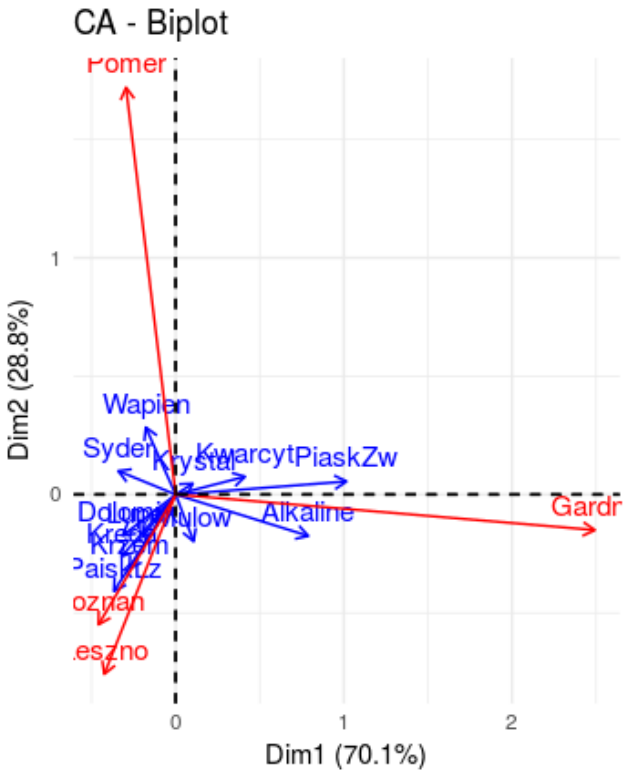
Wykres dwuwarstwowy

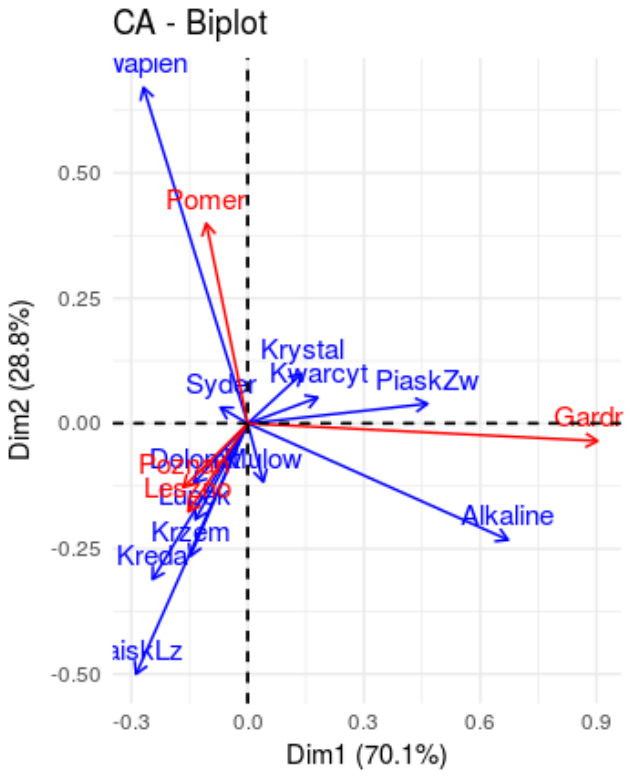
- pozwala określić relację pomiędzy cechami a składowymi zarówno dla cech jak i przypadków tu czy można wyjaśniać czy występowanie określonych typów skał wiąże się z określonym zlodowaceniem



Wzajemne relacje pomiędzy zmiennymi

- Wykresy biplot pozwalają określić relację pomiędzy zmiennymi:
  1) Jakie typy skał identyfikują poszczególne nasunięcia
  2) Jakich skał spodziewać się w glinach poszczególnych nasunięć





Pre-processing - podsumowanie

- Aby wydobyć ze zmiennych istotne informacje należy dane prawidłowo przygotować do analizy. Standardowy pre-processing obejmuje następujące kroki:
  1) Organizację danych do postaci płaskiej, ew. agregację (**wrangling**)
  2) Uzupełnianie braków w danych (**imputation**)
  3) Transformacje do rozkładu (zbliżonego do) normalnego (**normalisation**)
  4) Skalowanie do porównywalnych zakresów (**standarisation**)
  5) Redukcja wymiarowości: PCA lub inne (**dimemsionality reduction**)
  6) Wybór istotnych cech (składowych) do dalszej analizy (**feature selection**)


BIG
DATA
INFORMATION
TECHNOLOGIES
MANAGEMENT
ANALYSIS
STORAGE
LARGE
SETS
MILLION
SEARCH
PETABYTES
INTERNET
RESEARCH
LEARNING
SOFTWARE
VOLUME
PROCESS
GOVERNMENT
SAN
NAS
SIZE
EXABYTES
DEVELOPMENT
TECHNOLOGY
SCIENCE
FUTURE
NEW
BILLION
PARALLEL
INITIATIVE
TERABYTES
BASED
EVERY
WORLD
AMOUNT
REAL
STRUCTURE
WORK
APPLICATIONS
INFRASTRUCTURE
DATABASES
PROCESSING
ANALYTICS
SIMULATIONS
FRAMEWORK
BUSINESS
BETWEEN
EFFECTIVE
VISUALIZATION
DEFINITION
STATISTICS
USING
MASSIVE
FACTOR
MANAGE
ONLINE
DAY
WORLDWIDE
SYSTEMS
MAKING
INTEGRATION
SOCIAL
YEARS
RESULTS
DRIVE
REPLICATION
HANDLE
SOURCES
CHALLENGES
LIMITS
COMPLEX
ANNOUNCED
PAST
HAND
CAPACITY
ONE
PRIVACY
INTERNATIONAL
IMPLEMENTATION
BECOMES
SECOND
NATIONAL
FLOW
INTELLIGENCE
APPROACHES
PROCESSES
PRIVATE
TERMS
YEAR
SUCCESSFUL
NEEDED
USES
COMPARED
CAPABILITIES
LESS
HUGE
TIMES
SHARED
VARIOUS
ARCHITECTURE
PARADIGM
FUNDING
INSIGHT
TOOLS
CENTER
RECORDS
COST
PER
DISTRIBUTED
RELEVANT
USED
PROJECT
ANNUAL
NOW
ECONOMIC
VARIETY
MARKET
TIME
SENSOR
WOULD
RELATED
CURRENTLY
USE
STATE
CRITIQUE
PEOPLE
TRAFFIC
CRITIQUES
MAPREDUCE
NEARLY
MOST
END
HYPOTHESIS
INCLUDE
NETWORKS
OUT
APPROACH
MANY
SET
DEPARTMENT
QUERIES
HIGH
USERS
DETERMINE
WELL
ALGORITHMS
COMPANIES
HIGHER