



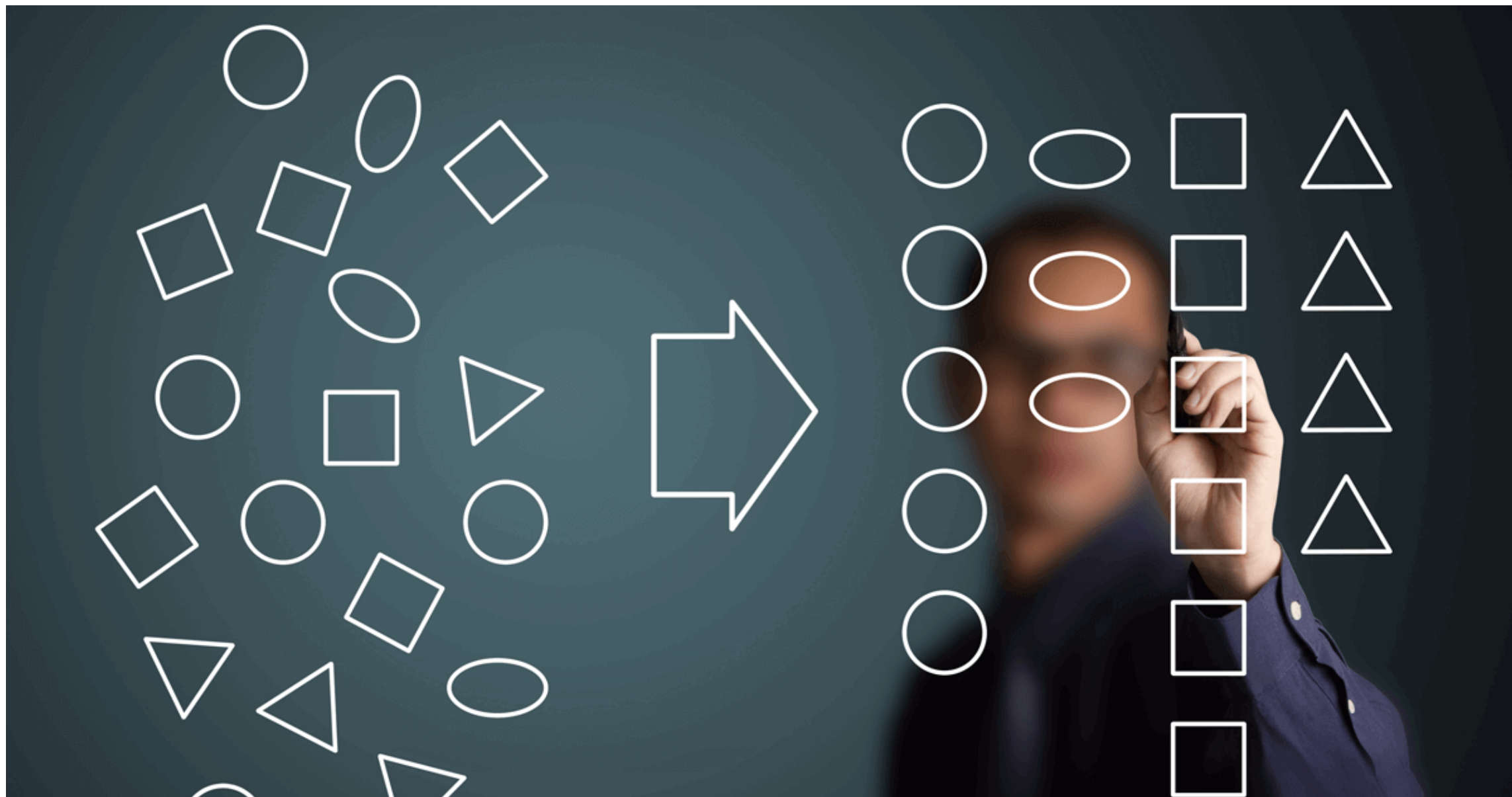
AKADEMIA GÓRNICZO-HUTNICZA  
IM. STANISŁAWA STASZICA W KRAKOWIE  
AGH UNIVERSITY OF KRAKOW

# Geograficzne Systemy Informacyjne

Metody klasyfikacji danych

Tomasz Bartuś  
Wydział Geologii, Geofizyki i Ochrony Środowiska  
Katedra Geologii Ogólnej i Geoturystyki

# Klasyfikacja danych

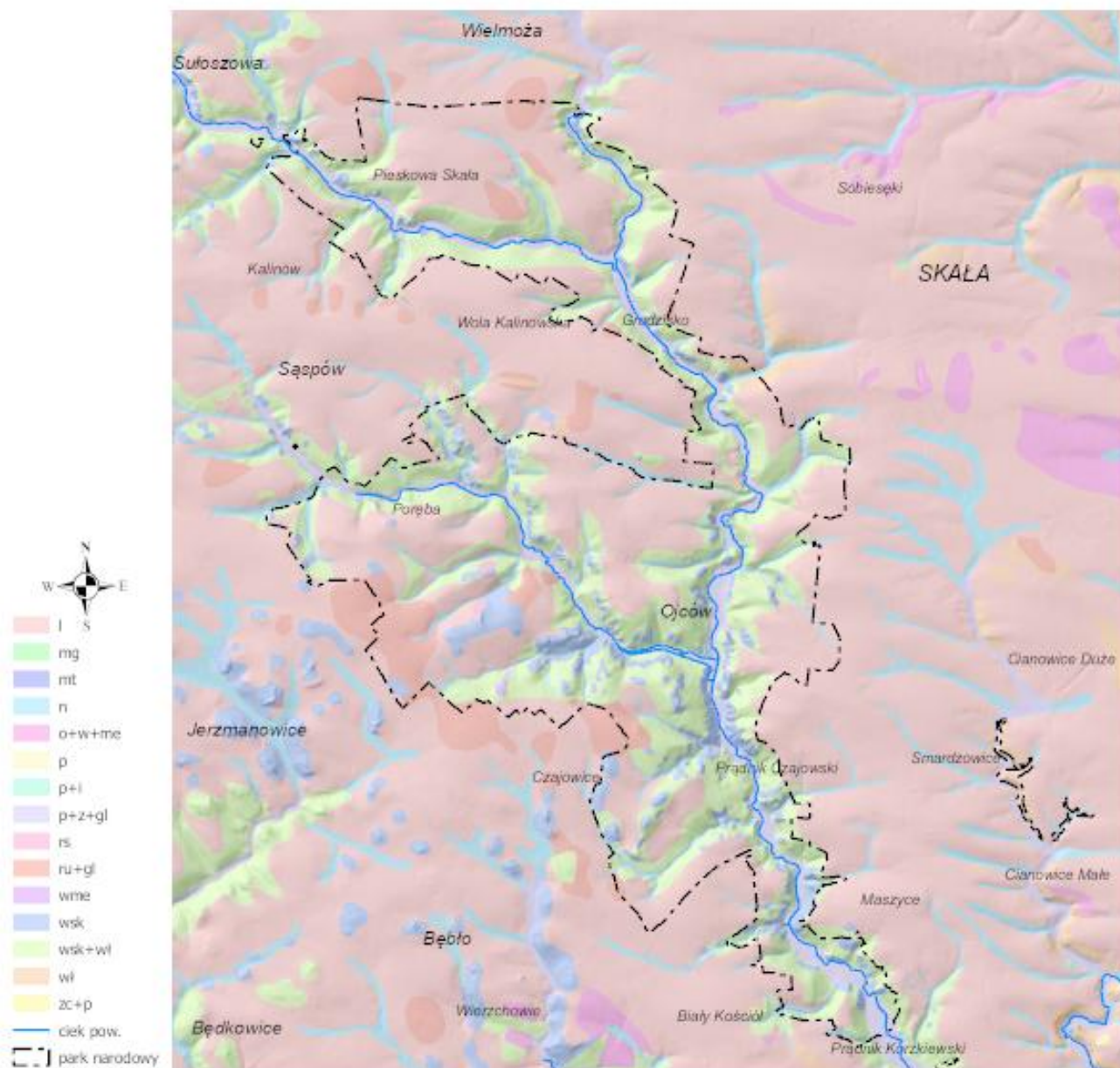


## Klasyfikacja danych

Klasyfikując dane możemy skorzystać z jednej z wielu standardowych metod klasyfikacji dostępnych w ArcGIS Pro lub można zdefiniować własne, niestandardowe zakresy klas.



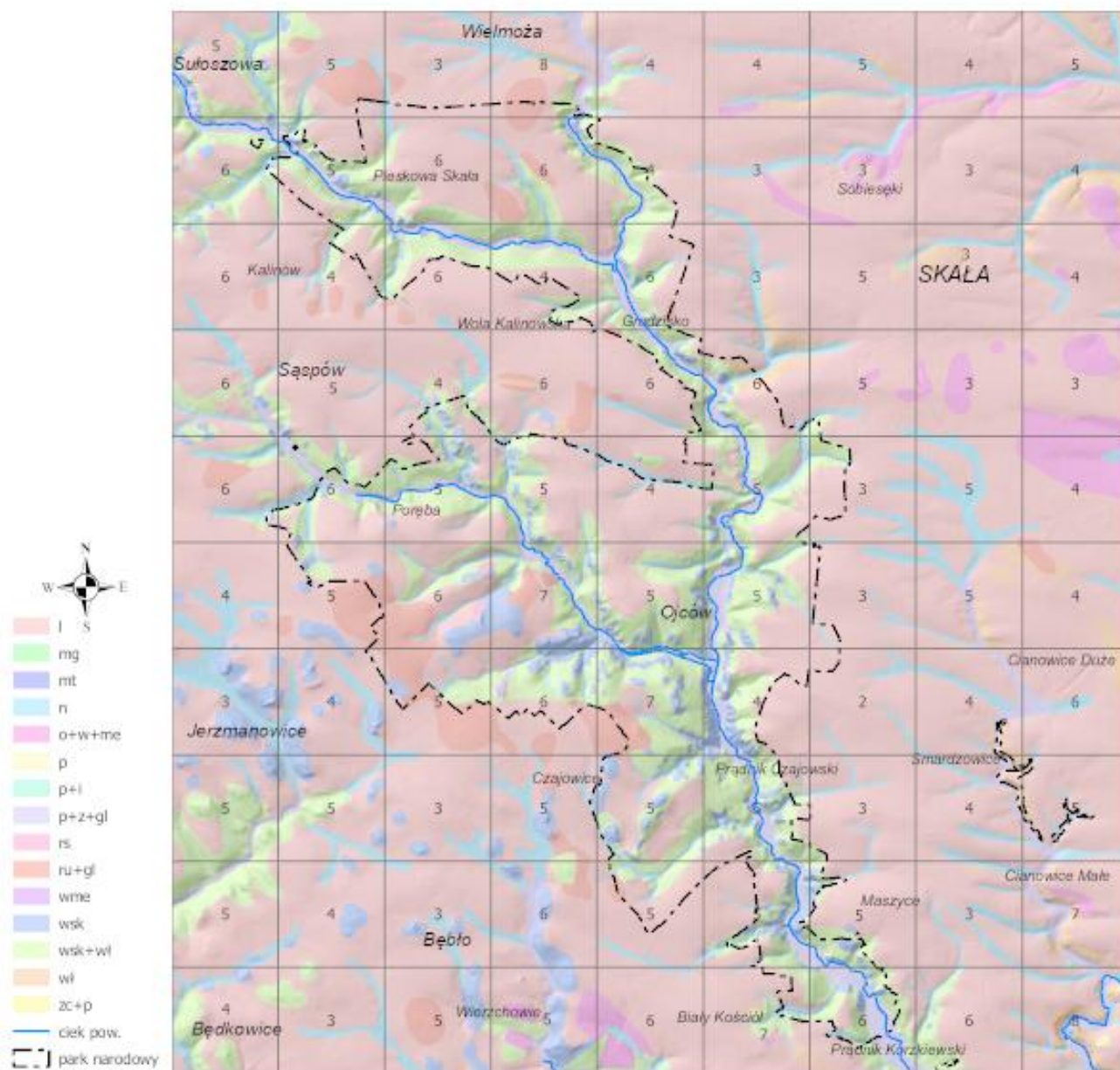
## Mapy tematyczne



Zmienność litofacjalna  
w rejonie Ojcowskiego Parku Narodowego



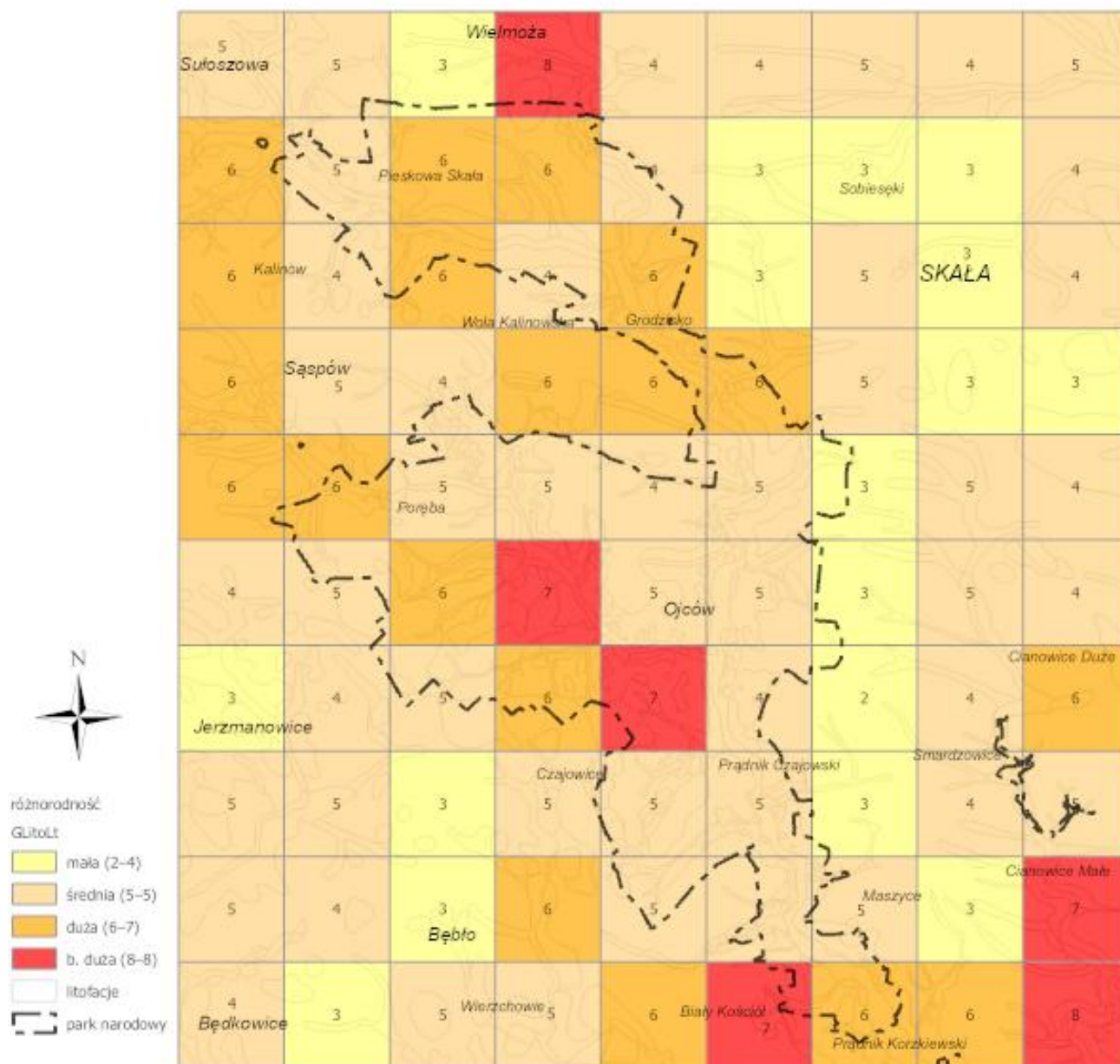
## Mapy tematyczne



Zmienność litofacjalna  
w rejonie Ojcowskiego Parku Narodowego

## Mapy tematyczne

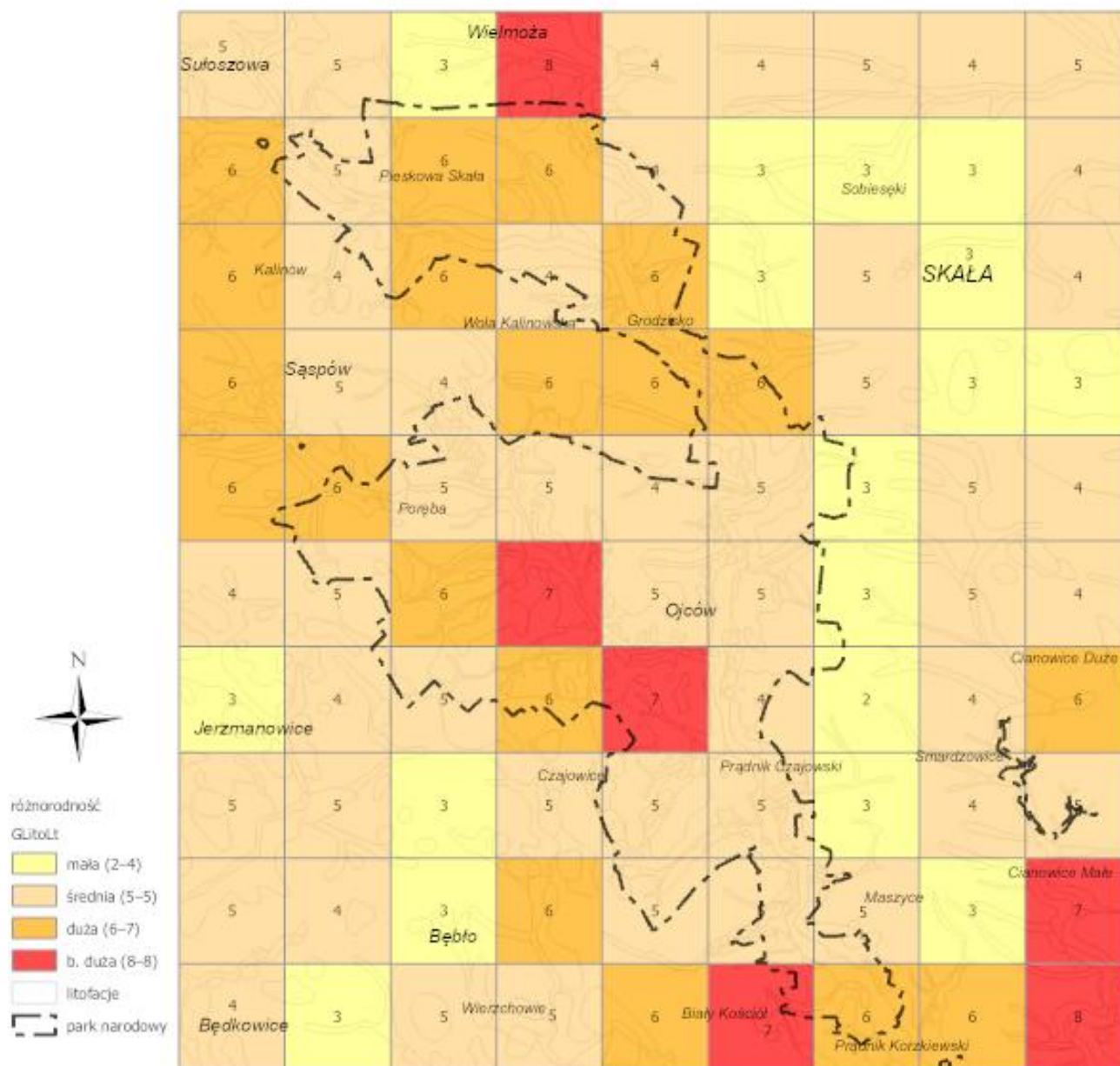
**Mapy tematyczne** są to mapy ilustrujące zmienność przestrzenną określonych zjawisk przyrodniczych i społeczno-gospodarczych.



Różnorodność ogniw litofacjalnych  
na podstawie liczby kategorii (*GLitoLt*)



## Decyzja o wyborze schematu klasyfikacji



- **Schemat klasyfikacji**  
danych obejmuje dwa zagadnienia:
  - wybór metody klasyfikacji,**
  - wybór liczby klas**  
służących do  
**pogrupowania danych.**

## Wybór metody klasyfikacji

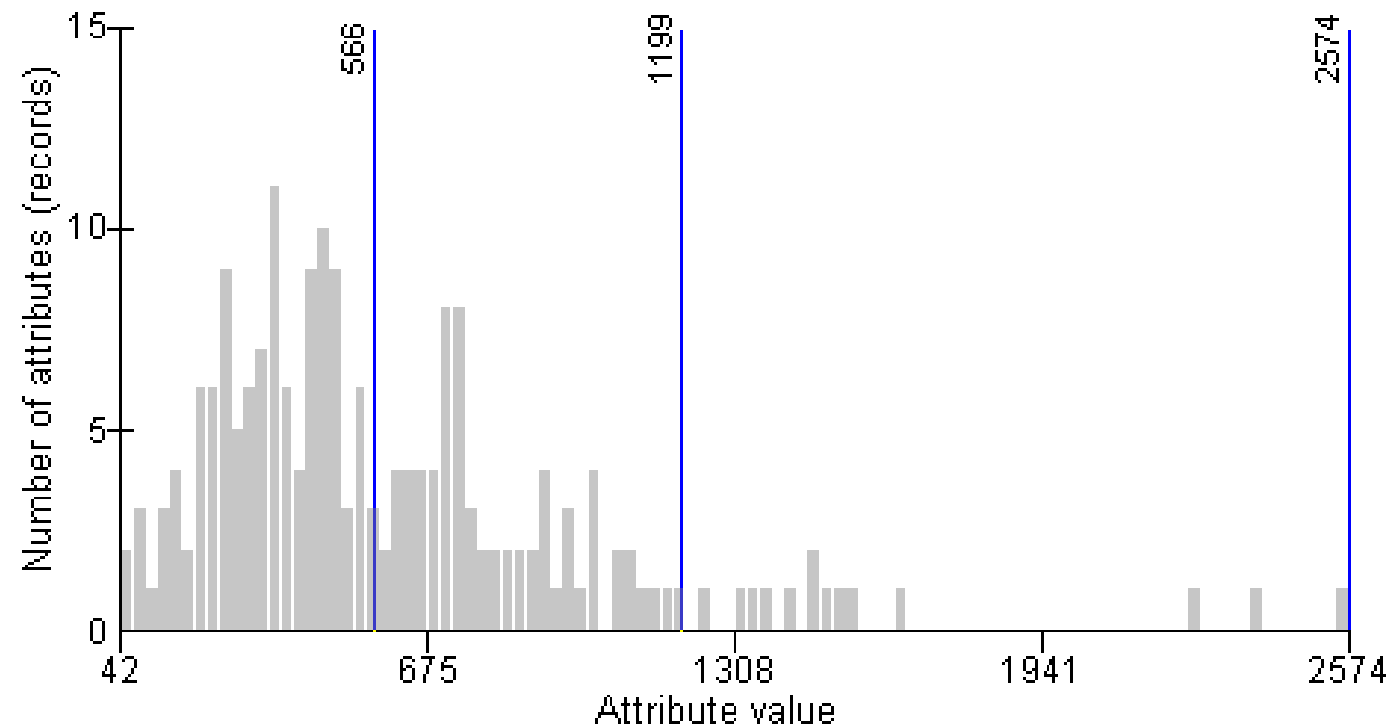
- Jednym rozsądnym rozwiązaniem jest próba oparcia tej decyzji o same dane. Można np. spróbować różnych metod klasyfikacji i wizualnie ocenić mapy wynikowe, a następnie wybrać metodę, która wydaje się najlepiej ilustrować przedstawiane zjawisko.
- Do oceny schematów klasyfikacji można użyć **histogramu**.



## Analiza rozkładu

**Rozkład** to sposób, w jaki wartości cechy (atrybutu) są rozłożone w całym zakresie jego zmienności.

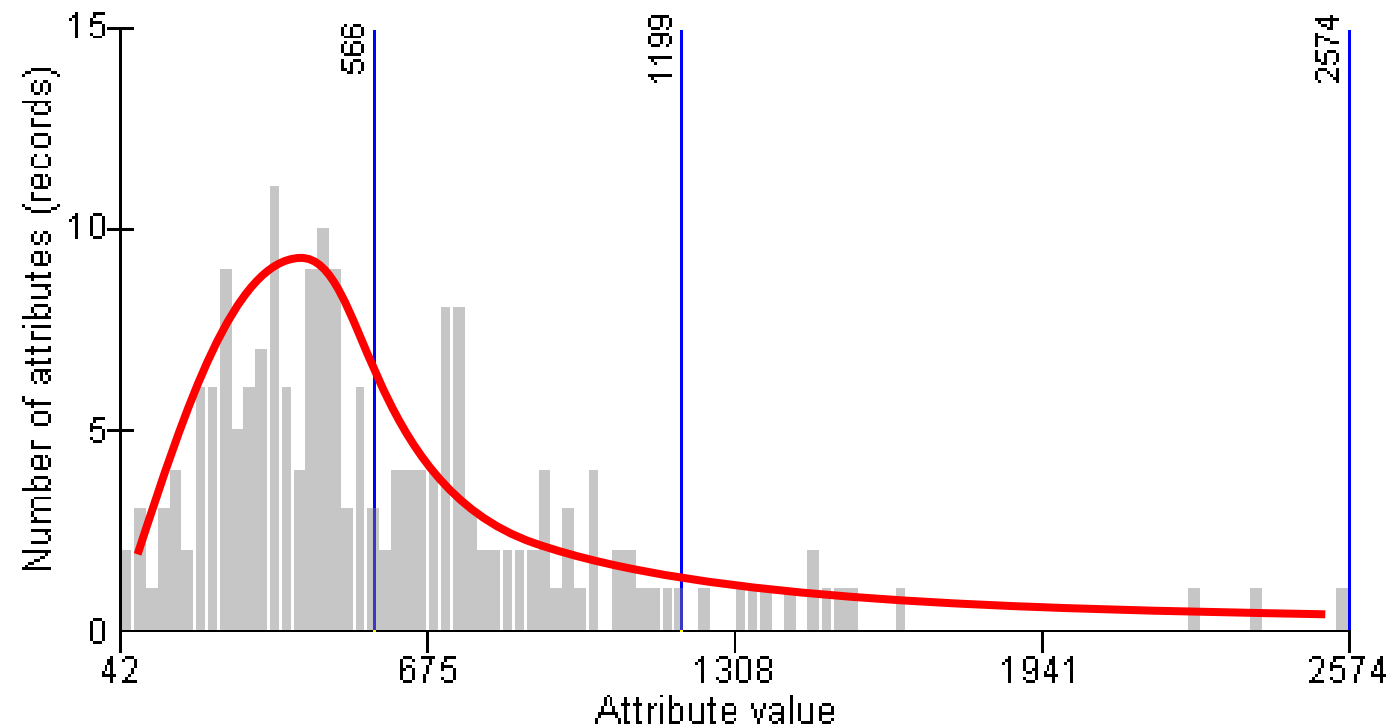
Jednym z podstawowych sposobów analizy rozkładu jest histogram.



## Analiza rozkładu

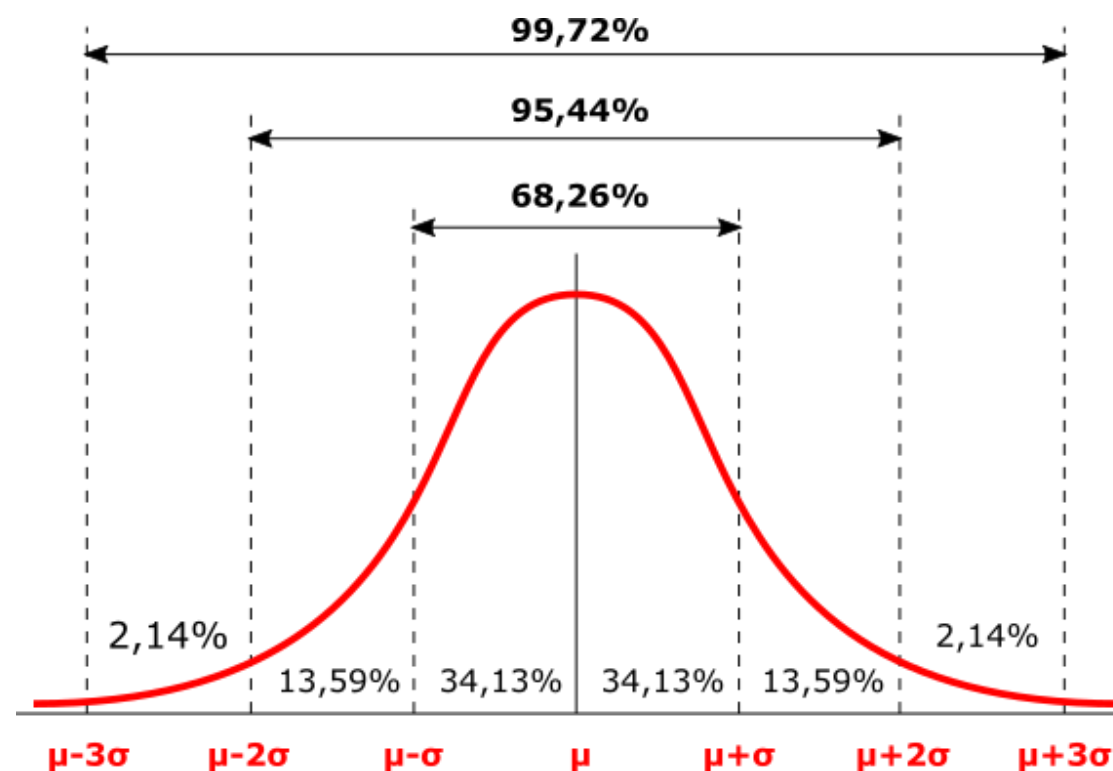
**Rozkład** to sposób, w jaki wartości cechy (atrybutu) są rozłożone w całym zakresie jego zmienności.

Jednym z podstawowych sposobów analizy rozkładu jest histogram.



## Rozkład normalny

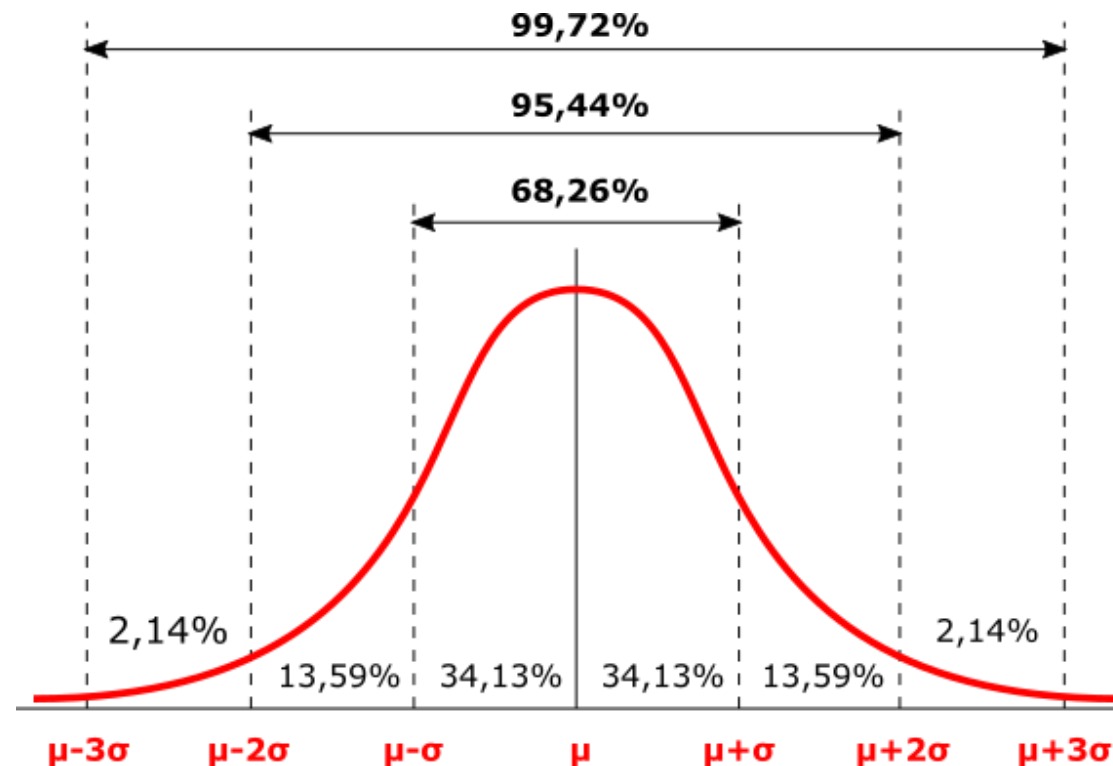
Istnieją dziesiątki zdefiniowanych rozkładów teoretycznych, które mają różnorodne zastosowania w statystyce, najbardziej znanym rozkładem cech przyrodniczych i społecznych jest **rozkład normalny**.





## Parametry rozkładów

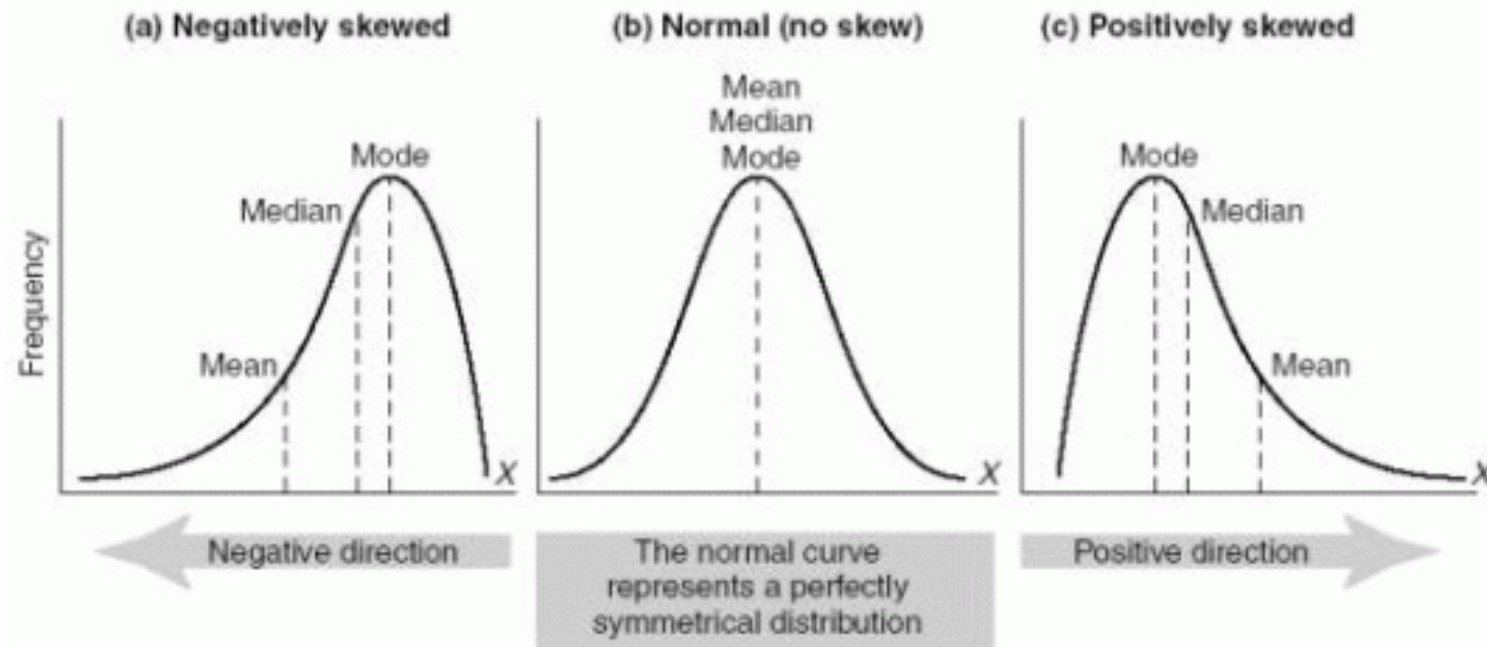
Rozkłady idealnie normalne występują tylko w abstrakcyjnym świecie matematyki. W rzeczywistości rozkłady empiryczne najczęściej różnią się od ideału matematycznego.



- odchyłka OX
- odchyłka OY
- liczba maksimumów

## Skośność

Jednym z powszechnych odstępstw od normalności jest **skośność** (asymetria), w której średnia wartość parametru jest wyższa lub niższa niż środek przedziału zmienności.



## Skośność

$X_i$ :

18, 6, 8, 2, 4, 3, 15, 10, 7, 9, 7, 6, 4, 6, 9

- Średnia arytmetyczna (*Mean*) ?
- Mediana (*Median*) ?
- Moda (*Mode*) ?



## Skośność

$X_i$ : 18, 6, 8, 2, 4, 3, 15, 10, 7, 9, 7, 6, 4, 6, 9

$n = 15$

Szereg rozdzielczy:

2, 3, 4, 4, 6, 6, 6, 7, 7, 8, 9, 9, 10, 15, 18

## Skośność

Szereg rozdzielczy:

2, 3, 4, 4, 6, 6, 6, 7, 7, 8, 9, 9, 10, 15, 18

$n = 15$

Średnia arytmetyczna =

Mediana (*Median*) =

Moda (*Mode*) =

$$\bar{X}_n = \frac{X_1 + X_2 + \dots + X_n}{n} = \frac{\sum_{i=1}^n X_i}{n}$$

Gdy  $n$  jest nieparzyste:  $Me = x_{\frac{n}{2}+1}$

Gdy  $n$  jest parzyste:  $Me = \frac{1}{2}(x_{n/2} + x_{n+1/2})$

## Skośność

Szereg rozdzielczy:

2, 3, 4, 4, 6, 6, 6, 7, 7, 8, 9, 9, 10, 15, 18

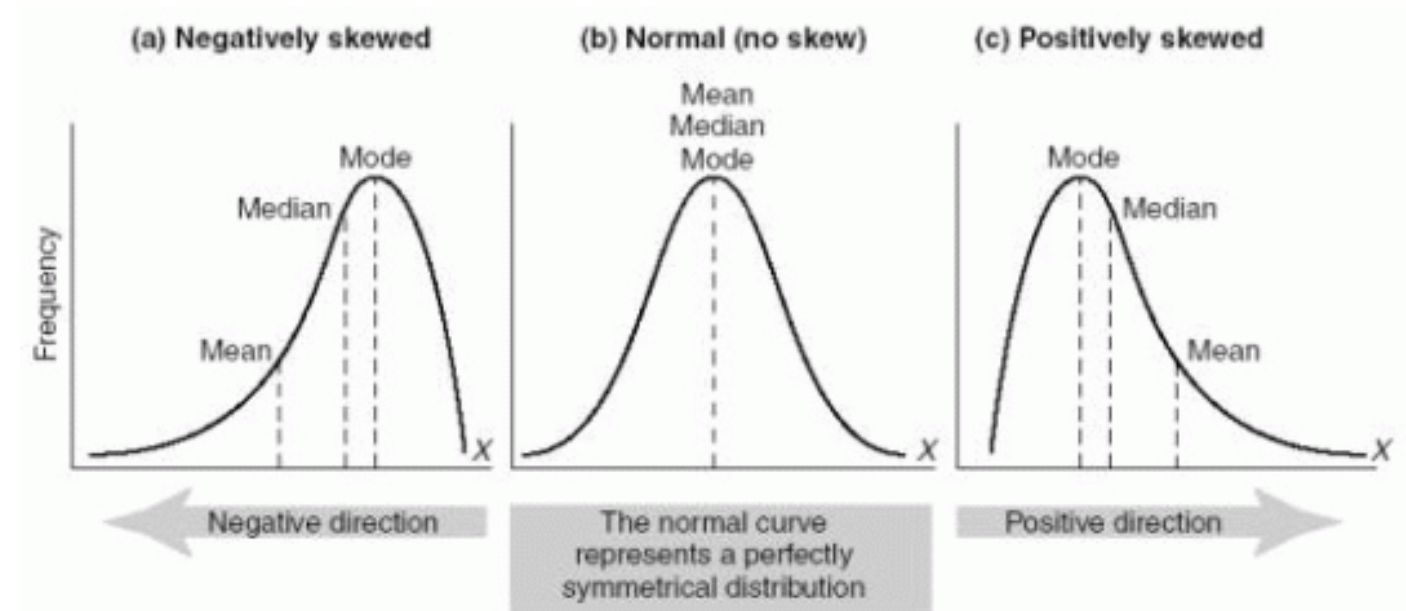
$n = 15$

Średnia arytmetyczna = **7,6**

Mediana = **7**

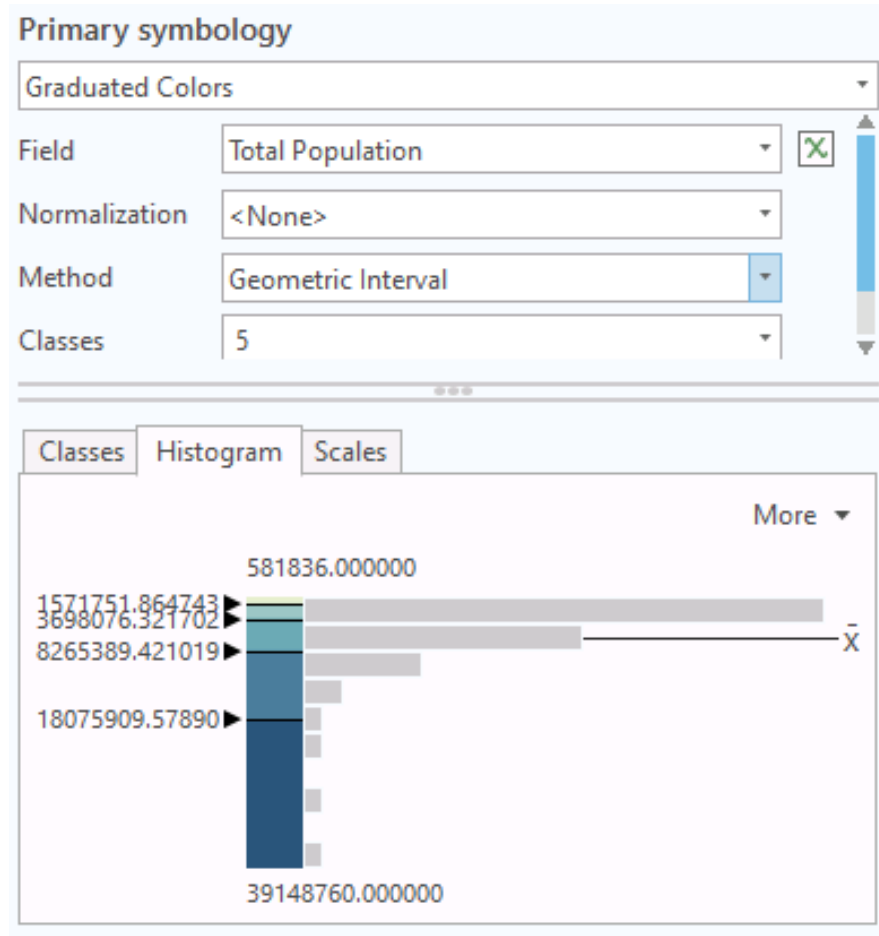
Moda = **6**

**Mo < Me < Śr**  $\Rightarrow$





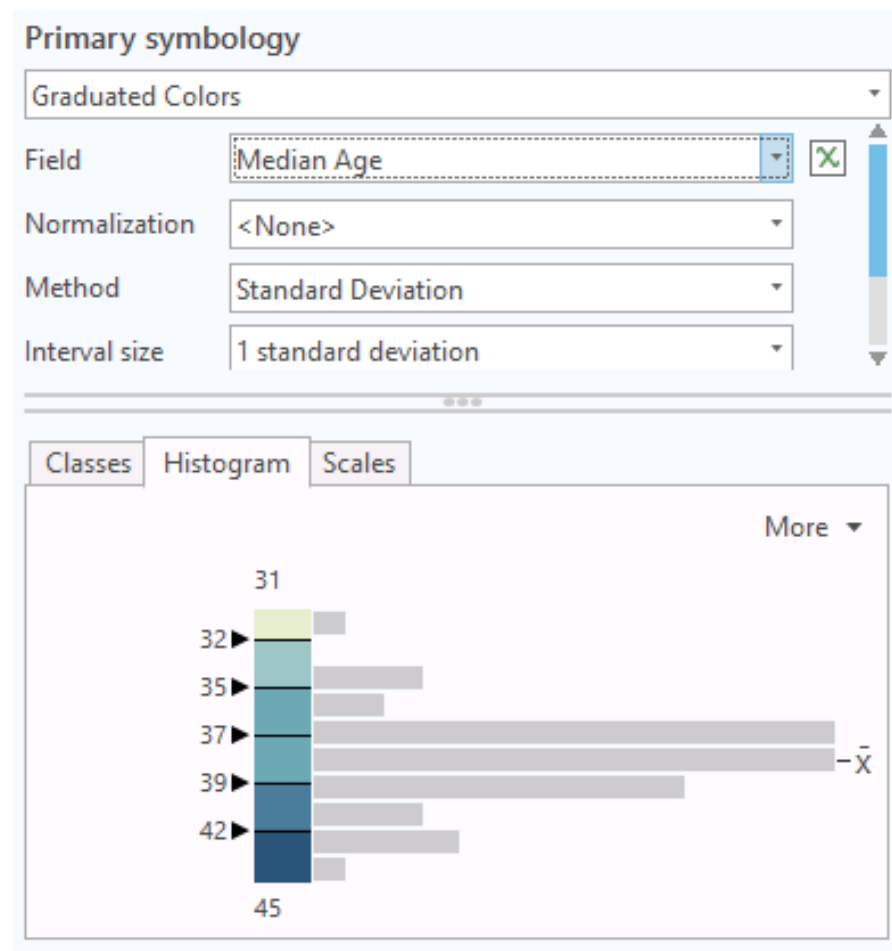
## Skośność geometryczna



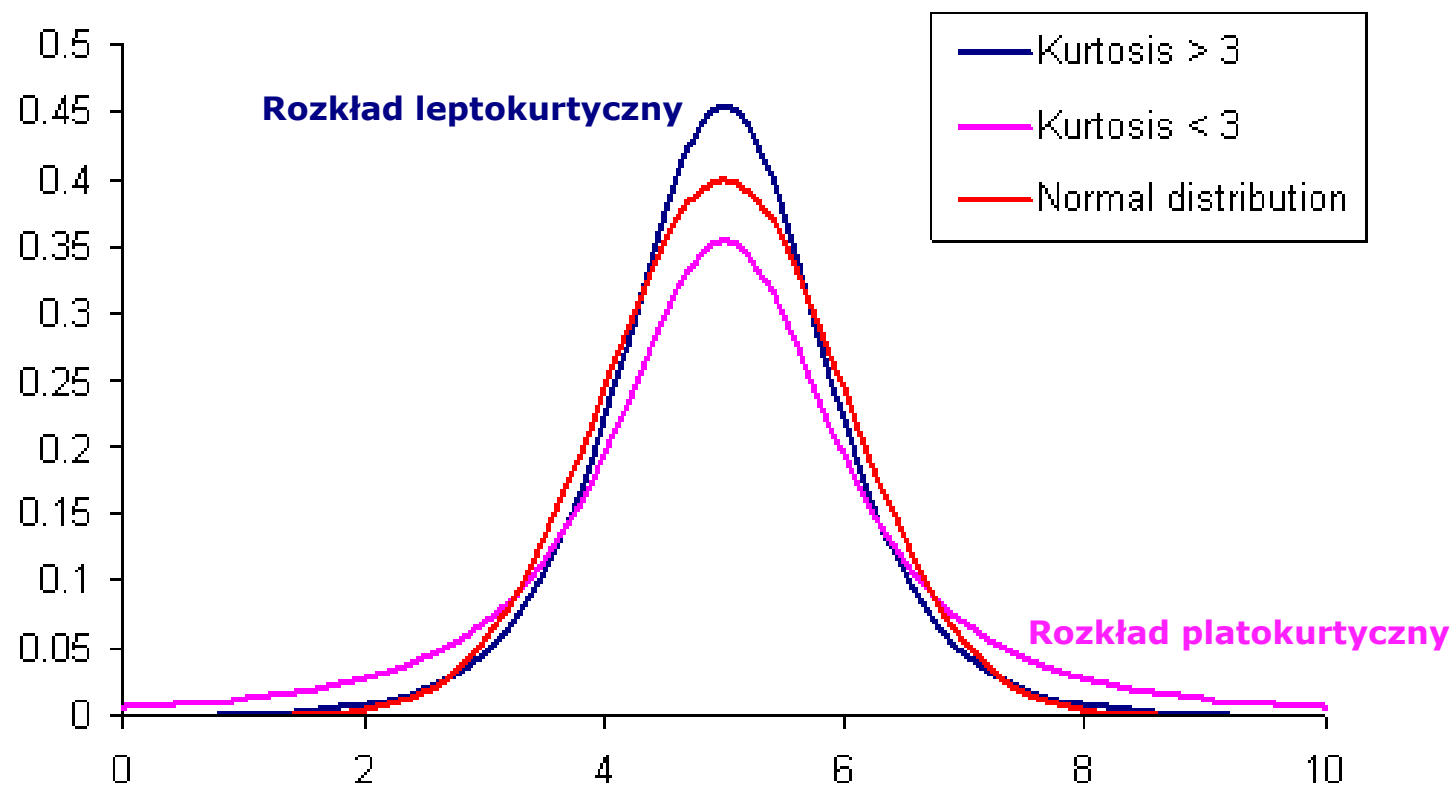
Jedną ze skrajnych form skośności jest **rozkład geometryczny**, w którym większość wartości jest skupiona w lewym końcu rozkładu.

Rozkład tego typu jest typowy w przypadku liczby ludności państw, gdzie kilka dużych krajów ma liczną populację obywateli ale większość krajów jest mała i ma niską populację. Często są to tzw. rozkłady **logarytmiczno-normalne**, ponieważ logarytmy wartości tworzą rozkład normalny.

# Kurtoza

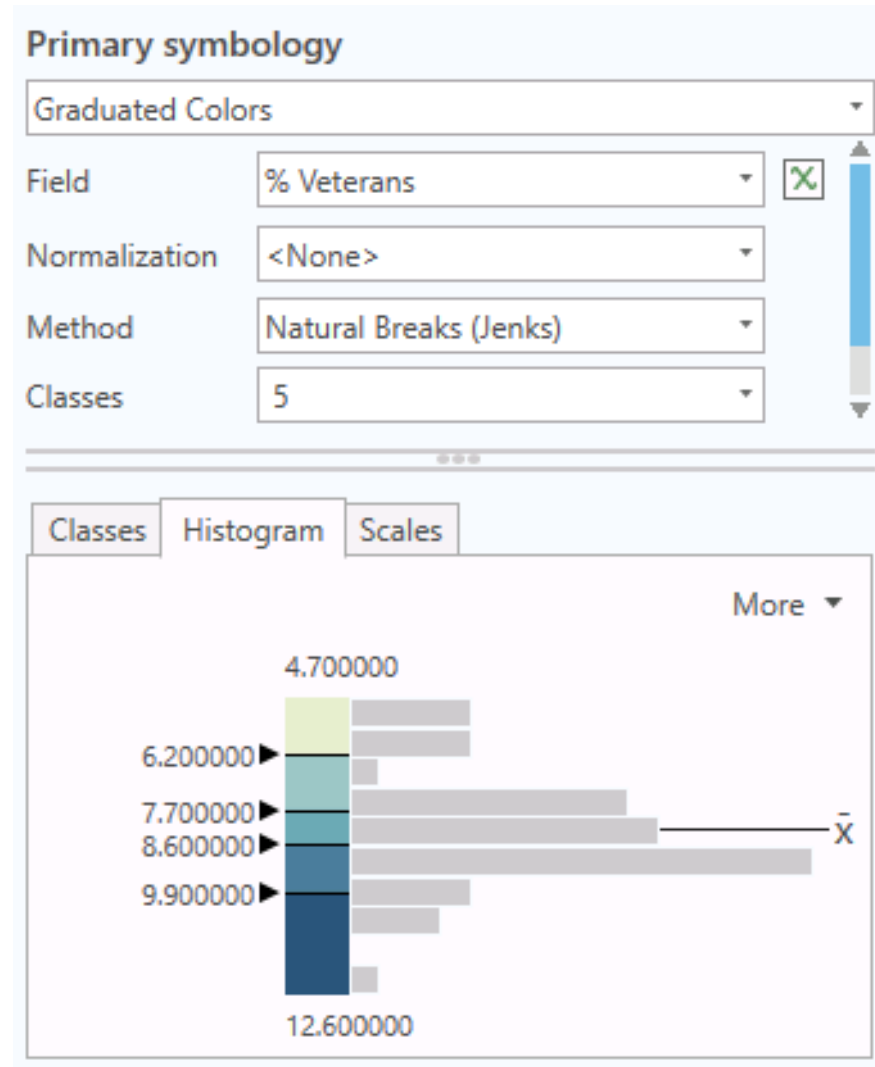


**Kurtoza** to ostrość maksimum rozkładu klasy centralnej.



Skrajnym przykładem niskiej kurtozy są rozkłady równomierne.

## Wielomodalność

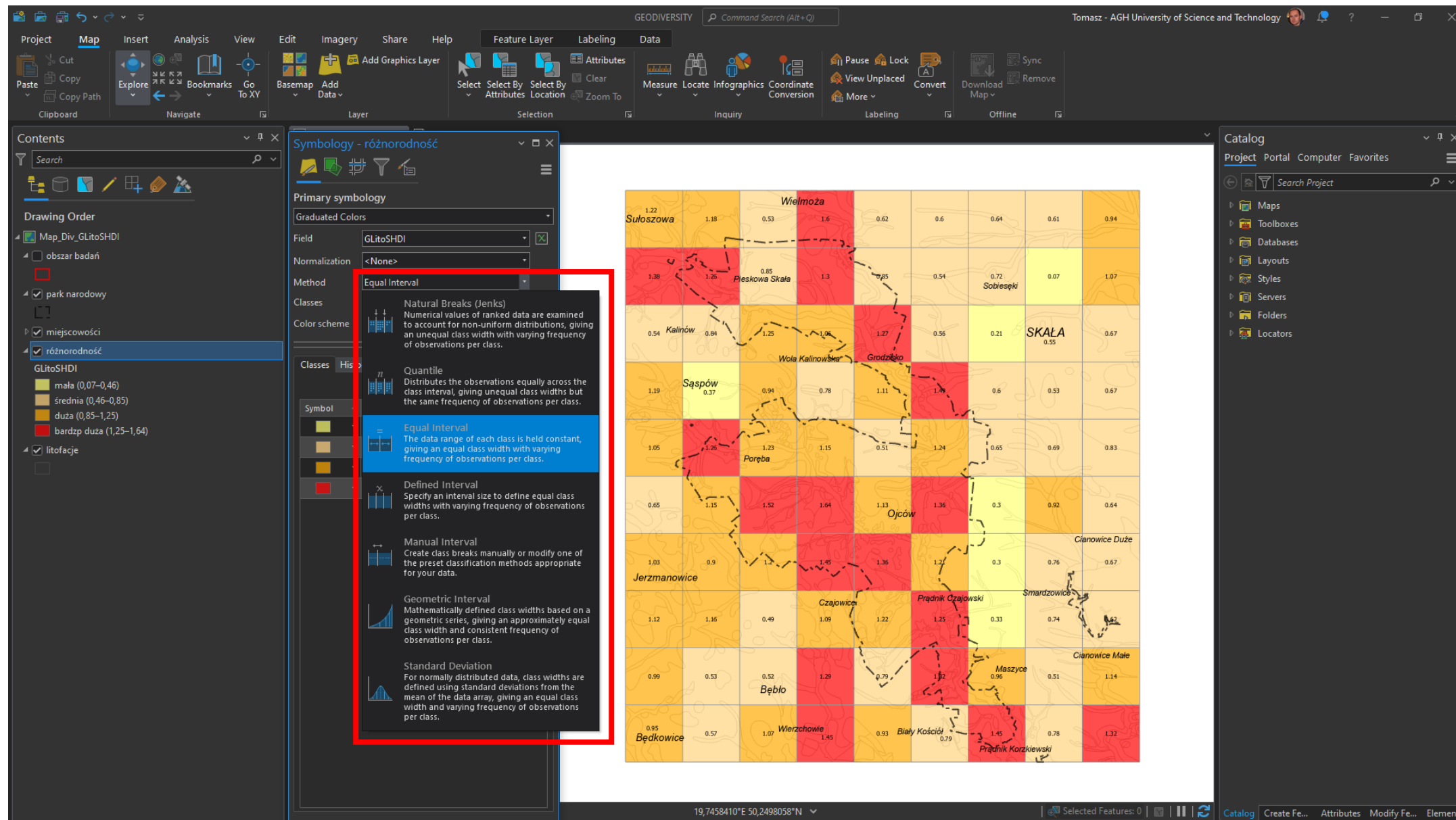


Rozkłady często zawierają wiele dominujących grup wartości. Nawet rozkłady z dominującą klasą centralną mogą mieć mniejsze maksima wtórne w lewym lub prawym ogonie. Takie rozkłady z wieloma skupiskami nazywane są **wielomodalnymi**.

## Wybór metody klasyfikacji

- Innym podejściem jest wybór schematu klasyfikacji na podstawie analizy (naukowej lub statystycznej) zmienności badanego atrybutu.
- Może wreszcie być tak, że możemy dysponować z góry określonymi standardami i kryteriami, określającymi metodę klasyfikacji lub liczbę klas.

# Metody klasyfikacji danych w ArcGIS Pro



The screenshot displays the ArcGIS Pro interface with the 'Symbology - różnorodność' (Symbology - diversity) pane open. The 'Primary symbology' is set to 'Graduated Colors'. The 'Field' is 'GLitoSHDI' and 'Normalization' is '<None>'. The 'Method' dropdown is set to 'Equal Interval', which is highlighted with a red box. The 'Color scheme' is set to 'Sequential'. The 'Classes' pane shows four classes: 'mała (0,07-0,46)', 'średnia (0,46-0,85)', 'duża (0,85-1,25)', and 'bardzo duża (1,25-1,64)'. The map view shows a grid of values representing GLitoSHDI, with a red dashed line indicating a specific area of interest.

**Equal Interval**  
The data range of each class is held constant, giving an equal class width with varying frequency of observations per class.

**Natural Breaks (Jenks)**  
Numerical values of ranked data are examined to account for non-uniform distributions, giving an unequal class width with varying frequency of observations per class.

**Quantile**  
Distributes the observations equally across the class interval, giving unequal class widths but the same frequency of observations per class.

**Defined Interval**  
Specify an interval size to define equal class widths with varying frequency of observations per class.

**Manual Interval**  
Create class breaks manually or modify one of the preset classification methods appropriate for your data.

**Geometric Interval**  
Mathematically defined class widths based on a geometric series, giving an approximately equal class width and consistent frequency of observations per class.

**Standard Deviation**  
For normally distributed data, class widths are defined using standard deviations from the mean of the data array, giving an equal class width and varying frequency of observations per class.

## Metoda naturalnych przerw



1886–1889

HEIGHT	
11	Class: 11 - 11
15	Break: 11
18	Class: 12 - 19
19	Break: 19
29	Class: 20 - 35
30	
35	Break: 35
44	Class: 36 - 44

- Metoda nazwana na cześć twórcy algorytmu George'a Jenksa (*Natural breaks (Jenks)*)
- algorytm wyszukuje skupione grupy wartości w celu utworzenia kategorii, które mogą odzwierciedlać zjawisko grupowania.
- Tworzony jest szereg rozdzielczy, a następnie grupy są identyfikowane na podstawie wyraźnych przerw występujących w utworzonym ciągu danych.
- Jest to domyślna metoda klasyfikacji danych w ArcGIS Pro.

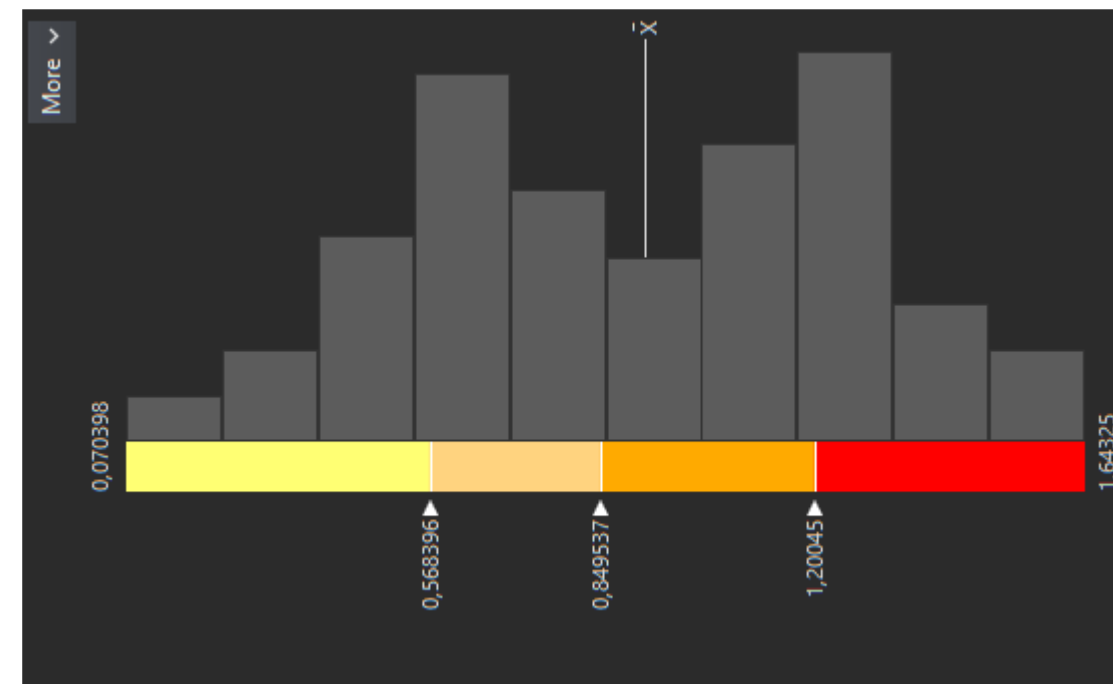
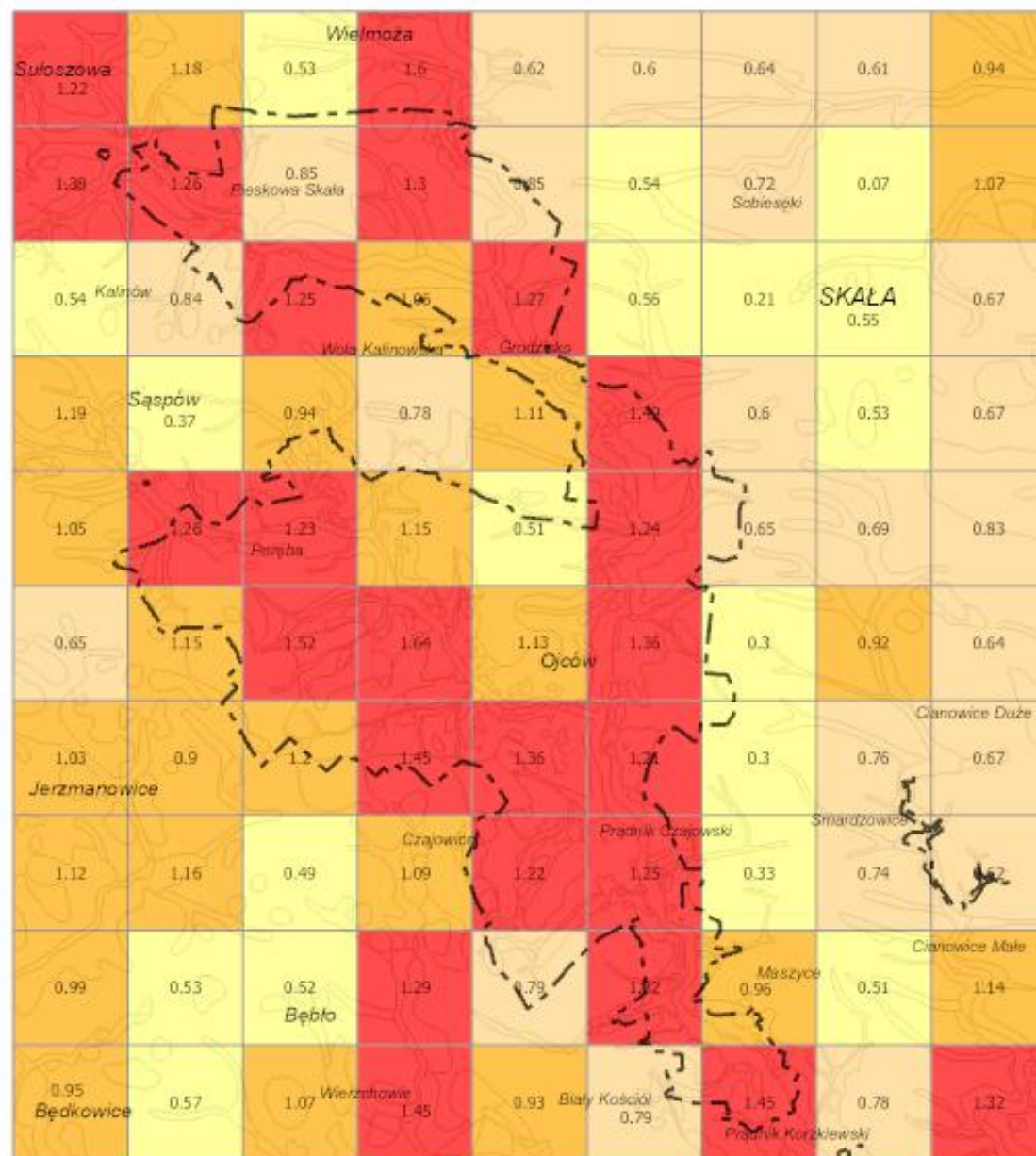


## Metoda naturalnych przerw

- Naturalne przerwy stanowią bezpieczny, ogólny wybór w przypadku większości rozkładów, stosujemy kiedy mapa musi jedynie dawać ogólny obraz rozkładu wartości w obszarach.
- Pojawia się problem w przypadku gdy dane zawierają skupiska wartości, które w rzeczywistości nie są znaczącymi grupami. Klasyfikacja naturalnych przerw połączy te skupiska razem i utworzy fałszywe wrażenie wizualne, które w rzeczywistości nie odzwierciedla wizualizowanego zjawiska.

HEIGHT	
11	Class: 11 - 11
15	Break: 11
18	Class: 12 - 19
19	Break: 19
29	Class: 20 - 35
30	
35	Break: 35
44	Class: 36 - 44

# Klasyfikacja metodą naturalnych przerw



## Klasyfikacja kwantylowa (*Quantile*)

HEIGHT	
11	Class: 10 - 15
15	Break: 15
18	Class: 16 - 19
19	Break: 19
29	Class: 20 - 30
30	Break: 30
35	
44	Class: 31 - 44

- Każda klasa zawiera jednakową liczbę elementów badanej cechy,
- np. gdy 15 województw chcemy pogrupować w 3 klasach, niezależnie od wartości atrybutów, w każdej klasie będzie się znajdowało po 5 województw.
- Jeśli używamy domyślnych 5 kategorii, dzielimy populację na pięć uporządkowanych kategorii według percentyla 20%.

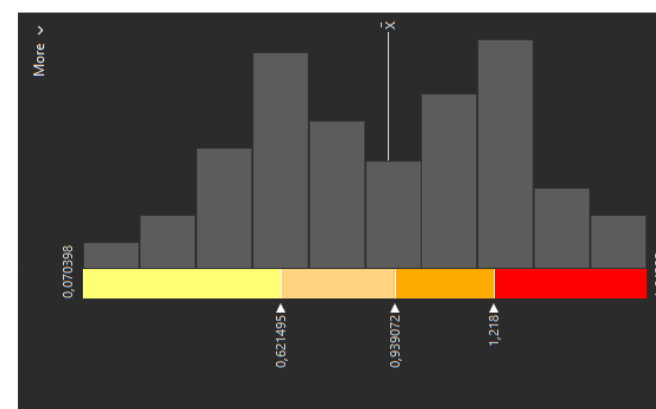
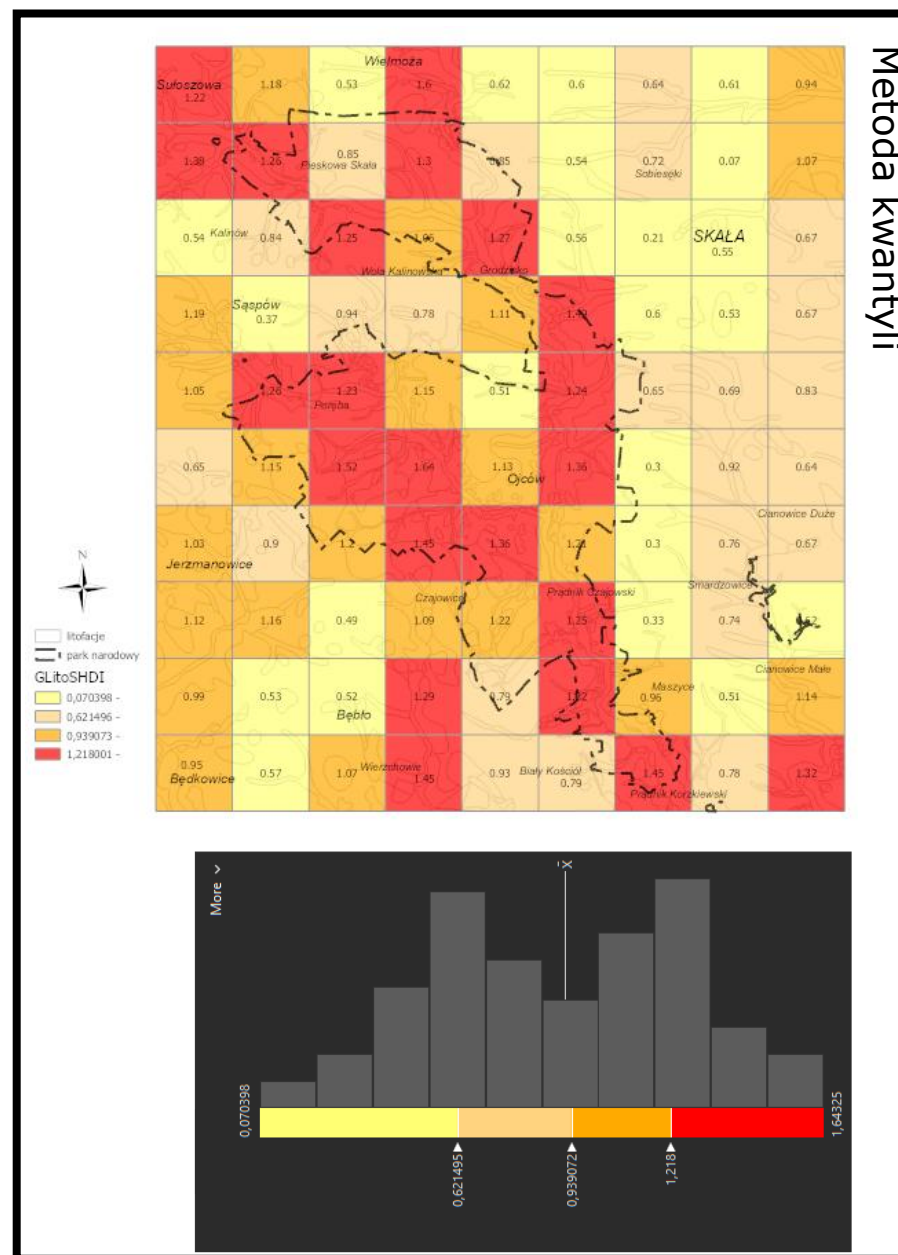
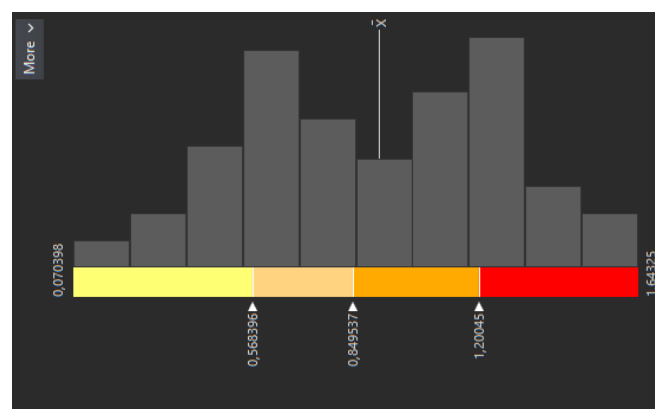
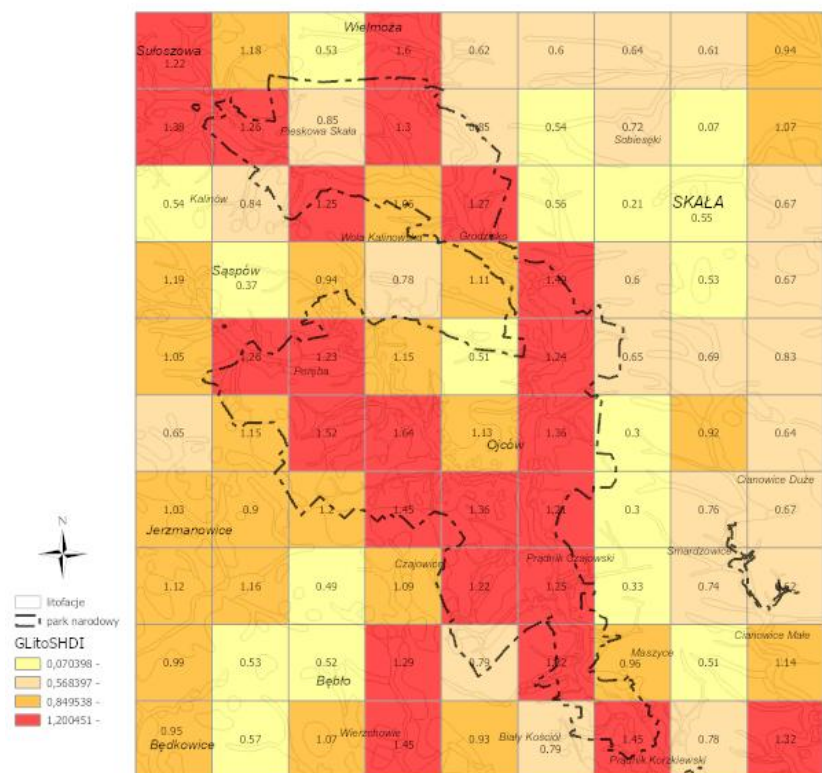
## Klasyfikacja kwantylowa

HEIGHT	
11	Class: 10 - 15
15	Break: 15
18	Class: 16 - 19
19	Break: 19
29	Class: 20 - 30
30	Break: 30
35	
44	Class: 31 - 44

### UWAGA!

Ponieważ obiekty są grupowane na podstawie równych liczb elementów w każdej klasie, wynikowa mapa często może wprowadzać w błąd. Podobne cechy można umieścić w sąsiednich klasach lub cechy o bardzo różnych wartościach można umieścić w tej samej klasie. Można zminimalizować to zniekształcenie, zwiększając liczbę klas.

# Klasyfikacja kwantylowa



## Metoda równych przedziałów (*Equal interval*)

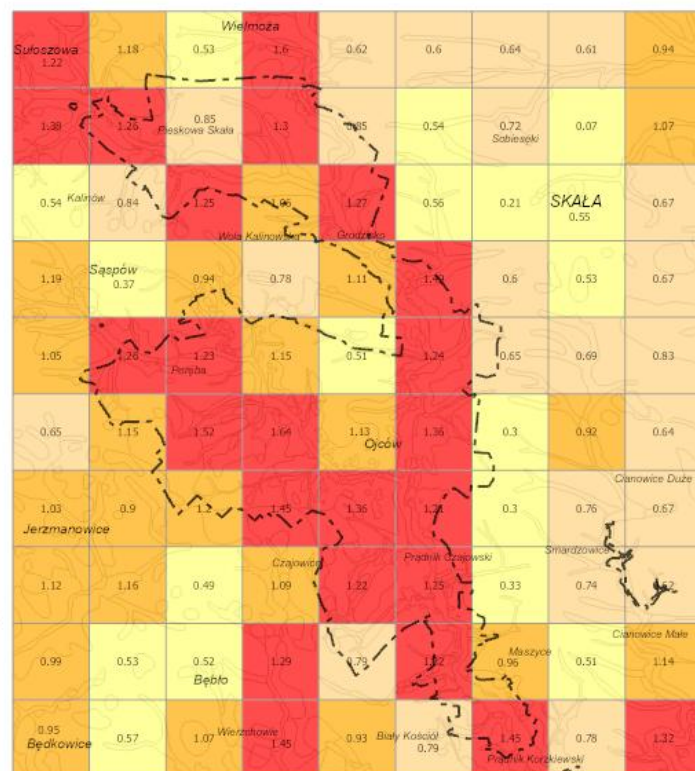
HEIGHT	
11	
15	Class: 11 - 22
18	
19	Break: 22
29	Class: 23 - 33
30	Break: 33
35	Class: 34 - 44
44	

- Definiujemy **liczbę klas**.
- Jest to metoda którą można porównać do linijki. Szerokości przedziałów klasowych są zawsze takie same, np. przedziały o szerokości 10% (1–10%, 11–20%, 21–30%, itd.).
- Metodę tę najlepiej zastosować do znanych zakresów danych, takich jak wartości procentowe i temperatura. Ta metoda podkreśla wielkość wartości atrybutu w stosunku do innych wartości.

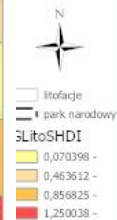
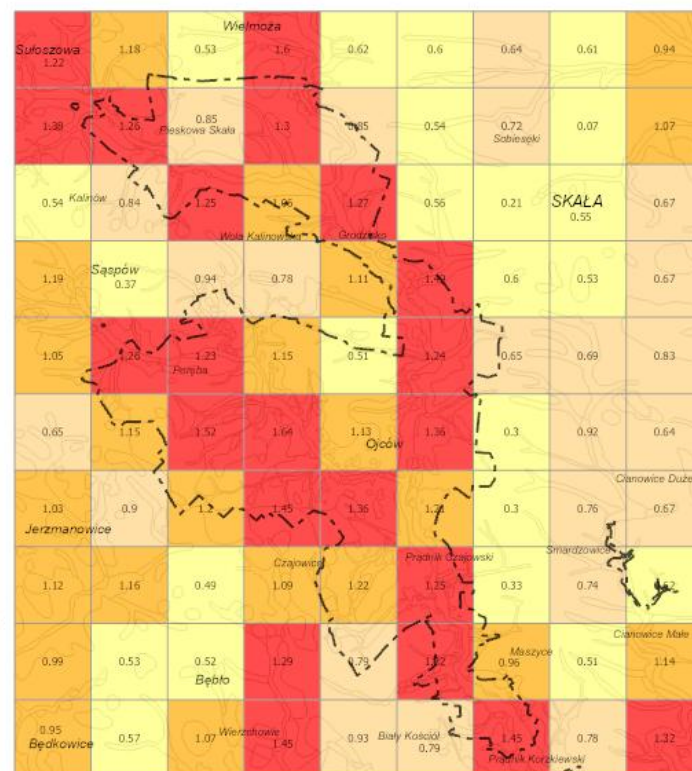


# Klasyfikacja równych przedziałów

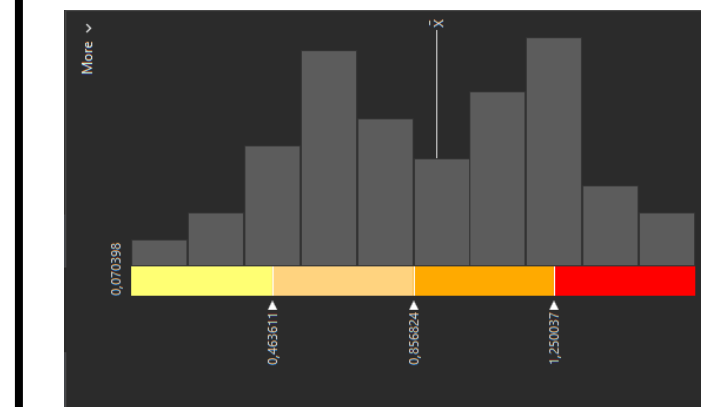
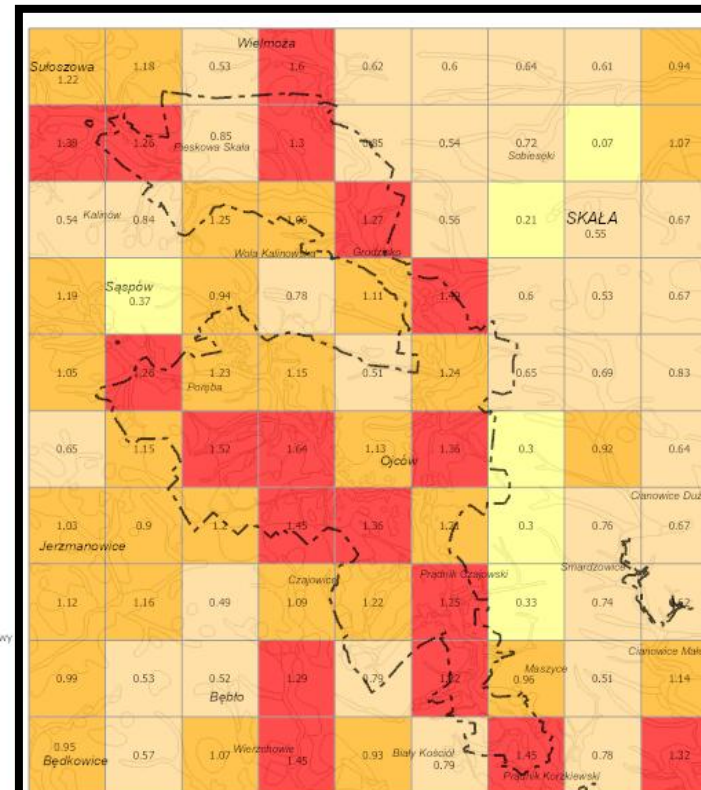
Metoda naturalnych przerw



Metoda kwantyli



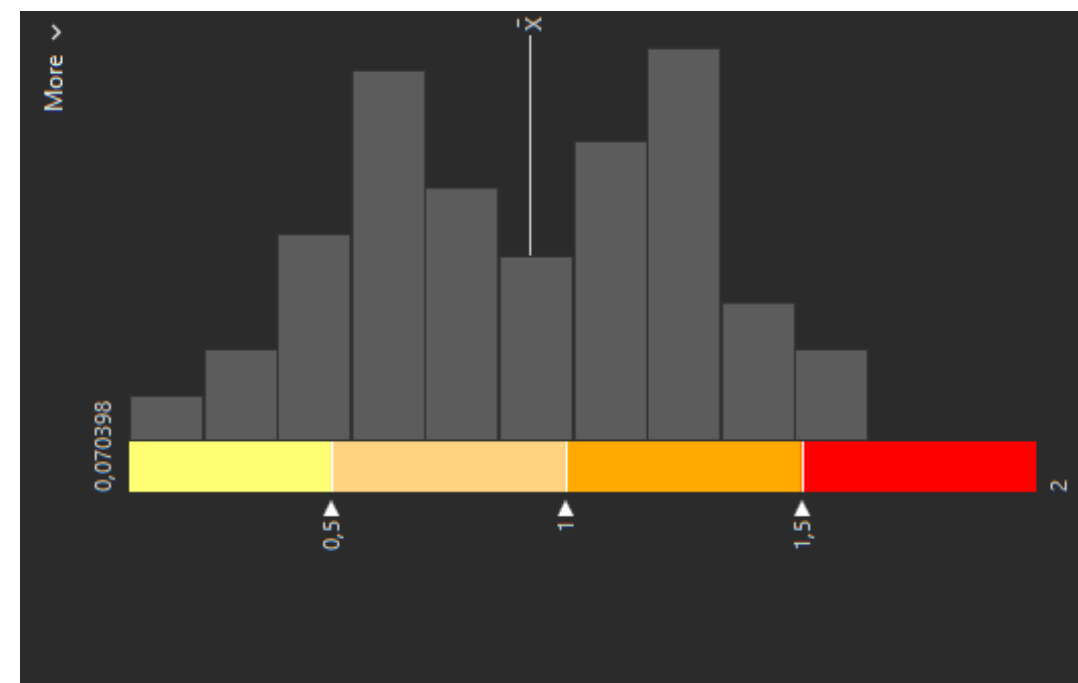
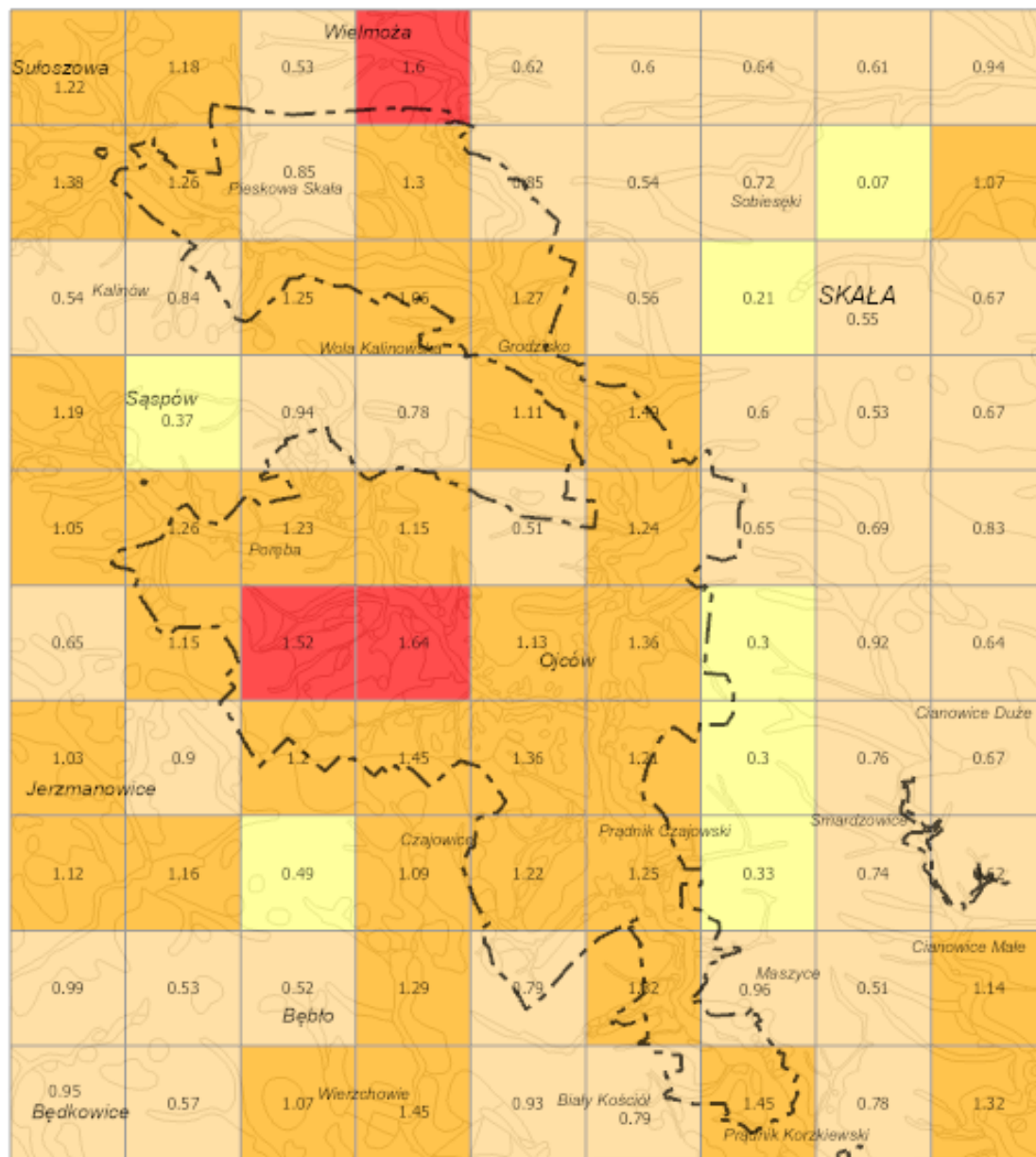
Metoda równych przedziałów



## Metoda definiowanego interwału

- Definiujemy liczbę klas **rozmiar interwału**.
- Np. jeśli rozmiar interwału wynosi 55, każda klasa będzie obejmować 55 jednostek.
- Liczba klas jest ustalana automatycznie na podstawie wielkości interwału i maksymalnej wartości z próby. Rozmiar interwału musi być wystarczająco mały, aby zmieścić się w minimalnej dozwolonej liczbie klas, czyli trzech.

# Metoda definiowanego interwału



## Metoda manualna

HEIGHT
11
15
18
19
29
30
35
44

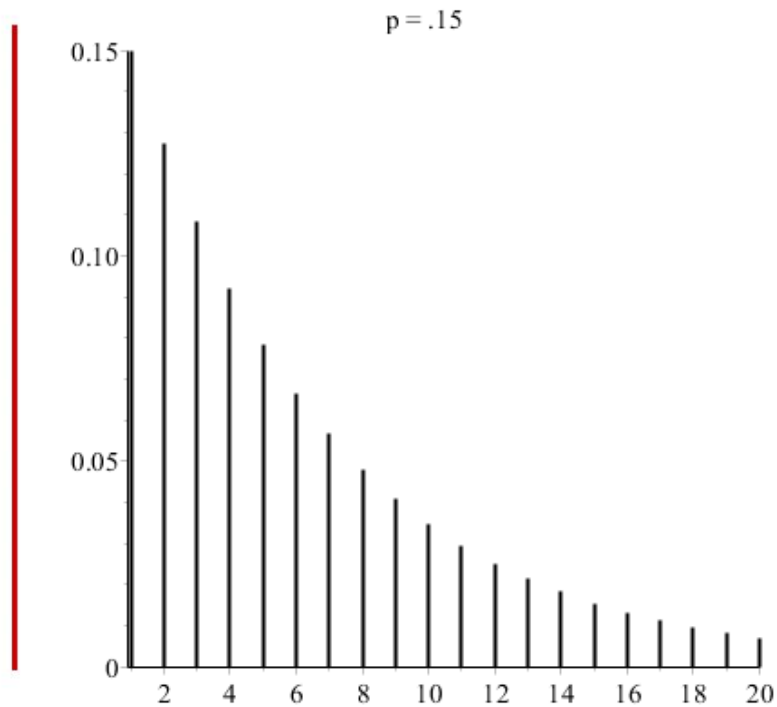
Class: Less than 30

Break: 29

Class: More than 30

- Zakres każdej klasy jest określany przez użytkownika. Metoda ta jest szczególnie przydatna, gdy mamy odzwierciedlić jakieś szczególne kryteria lub typy danych.
- Np. mamy dane dotyczące temperatury. Możemy zechcieć określić granicę przedziałów klasowych dokładne na 32° Fahrenheita (0°C).

## Interwał geometryczny (*geometrical interval*)



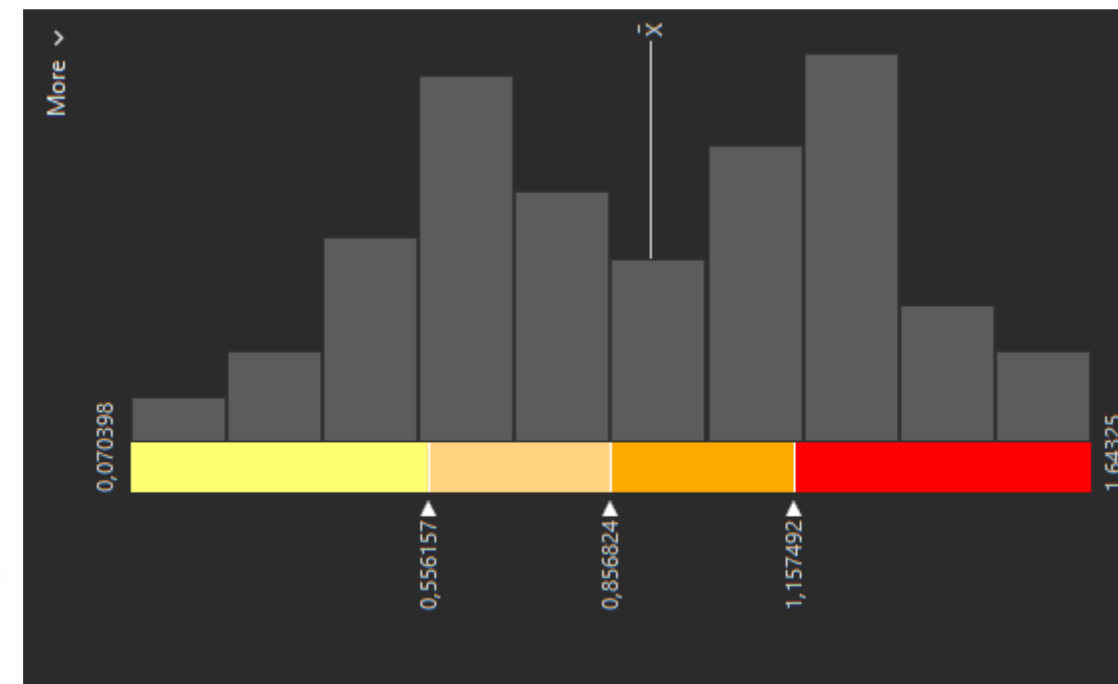
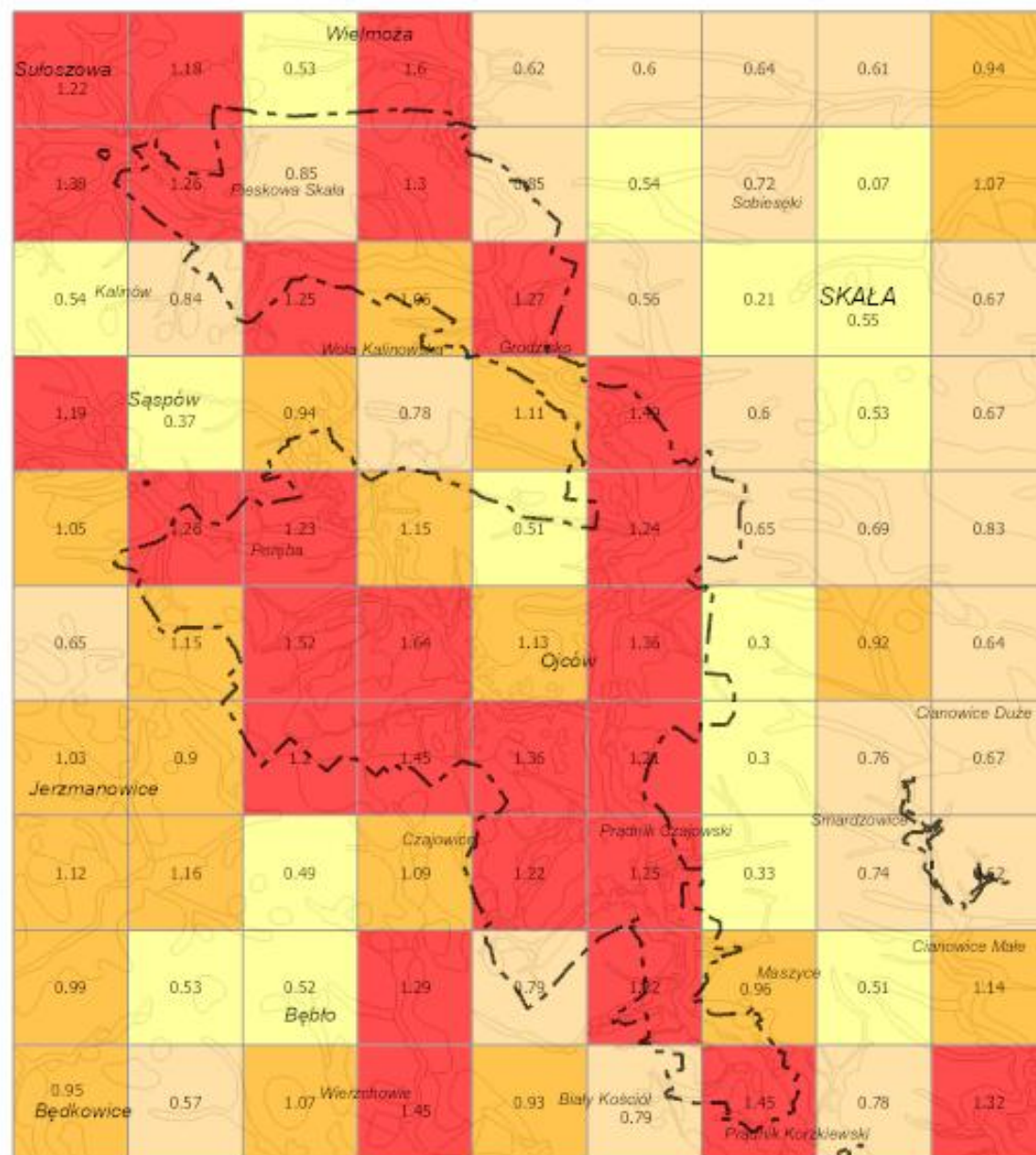
- Algorytm został zaprojektowany do obsługi zmiennych ciągłych o rozkładach geometrycznych.
- Tworzy równowagę pomiędzy podkreślaniem zmian wartości średnich i wartości skrajnych, tworząc w ten sposób efekt atrakcyjny wizualnie i kompleksowy pod względem kartograficznym.
- Klasy są definiowane matematycznie poprzez minimalizację sumy kwadratów liczby elementów w każdej klasie. Zapewnia to, że każdy zakres klas ma w przybliżeniu tę samą liczbę wartości w każdej klasie i że zmiana między przedziałami jest w miarę stała.

## Metoda interwału geometrycznego

- Przykładem zastosowania jest zbiór danych o opadach, w którym tylko 15 ze 100 stacji pogodowych (mniej niż 50%) zarejestrowało opady, a pozostałe nie zarejestrowały opadów, więc ich wartości atrybutów wynoszą zero.



# Metoda interwału geometrycznego

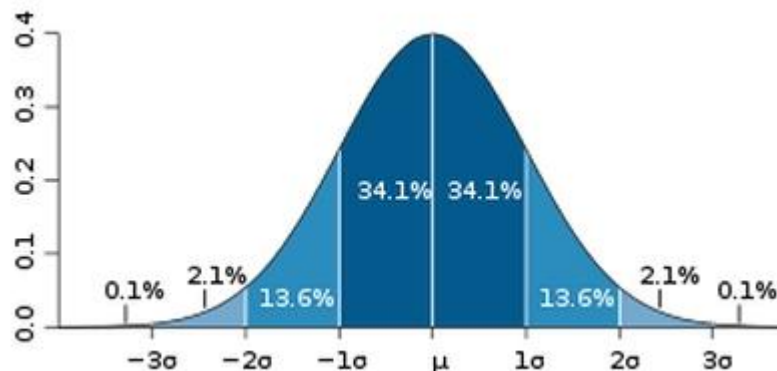


## Metoda odchylenia standardowego (*Standard deviation*)

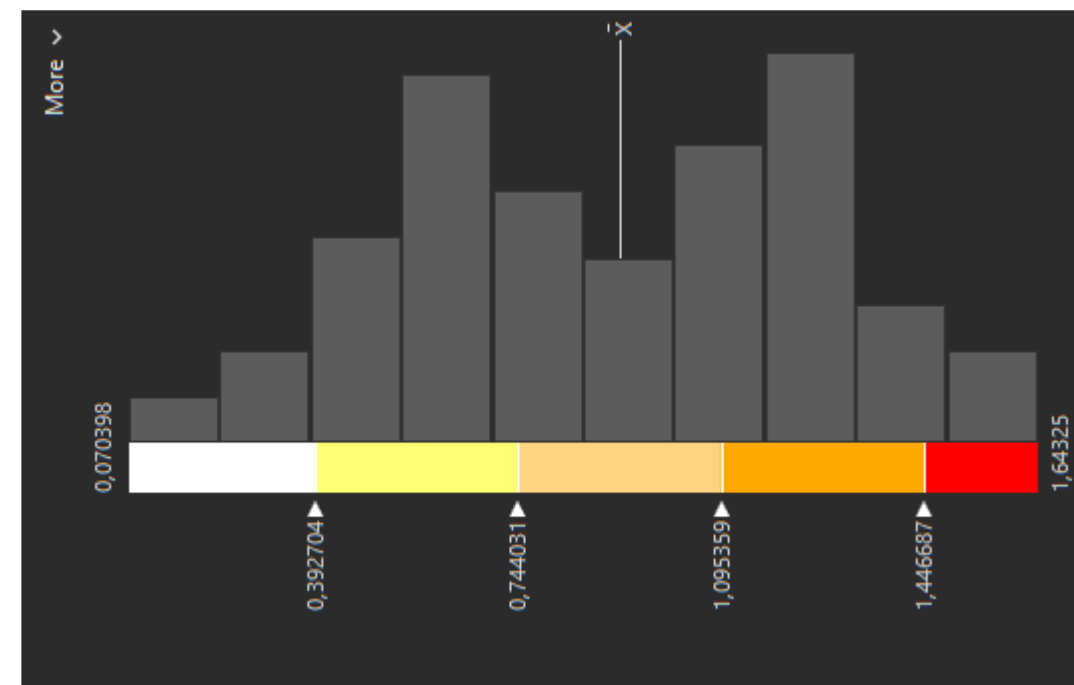
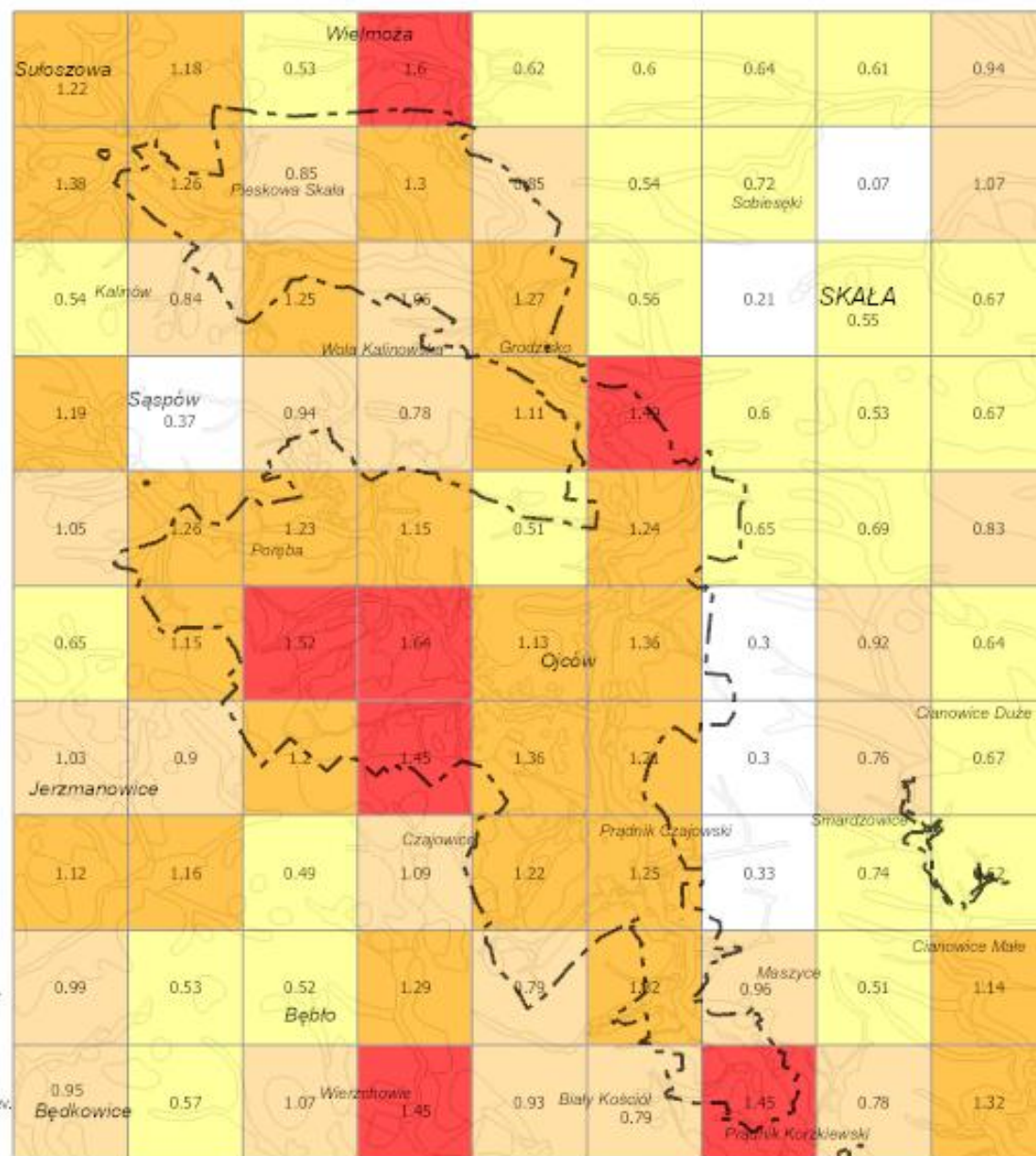
$$\bar{X}_n = \frac{X_1 + X_2 + \dots + X_n}{n} = \frac{\sum_{i=1}^n X_i}{n}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$

- Klasyfikacja tworzy klasy na podstawie średniej arytmetycznej i odchylenia standardowego.
- Pokazuje, jak bardzo wartość atrybutu obiektu różni się od średniej.
- Podziały klas są tworzone z jednakowymi zakresami wartości, które stanowią proporcję odchylenia standardowego – zwykle w odstępach jednej, połowy, jednej trzeciej lub jednej czwartej.



# Klasyfikacja metodą odchylenia standardowego



## Wybór metody klasyfikacji

- Która metoda klasyfikacji jest najlepsza?

Nie ma „poprawnej” odpowiedzi na tak postawione pytanie.

- Najlepszy schemat klasyfikacji dla danej klasy zależy od przeznaczenia analizy, rodzaju analizowanych danych, przyjętej dokładności i innych.

## Wybór metody klasyfikacji

Metoda	Kiedy używać	Ile klas
<b>Naturalnej przerwy</b>	Jeśli atrybuty są rozmieszczone nierównomiernie w całym zakresie zmienności analizowanego atrybutu (np. asymetria, wielomodalność rozkładów).	Algorytm sam wybierze liczbę, która najlepiej odzwierciedla naturalne grupy atrybutu, których zmienność chcemy zobrazować.
<b>Równych rozstępów</b>	Gdy chcemy, aby wszystkie klasy miały ten sam zakres.	Wybierz „okrągłą” liczbę, która będzie łatwa do zapamiętania i interpretacji np.: 2, 50, 1000, itd..,
<b>Kwantyli</b>	Jeśli atrybuty są rozmieszczone w sposób liniowy (równomierny rozkład w całym zakresie zmienności i niewielkie odchylenie liczebności elementów dla każdego przedziału klasowego).	Wybierz liczbę sensowną z punktu widzenia celów klasyfikacji.
<b>Manualna</b>	Gdy chcemy tworzyć klasy w oparciu o nasze doświadczenie i z uwzględnieniem specyfiki danych.	Wybierz liczbę adekwatną do zjawiska, które wizualizujemy. Np. potrzeba 2 klas, aby pokazać wartości powyżej i poniżej pewnej wartości progowej.



## Liczba klas

- Przy podejmowaniu decyzji o liczbie klas, powinno się raczej używać ograniczonej liczby przedziałów klasowych.
- Z reguły najlepszym wyborem jest użycie 3–7 klas.



## Materiały dodatkowe

- Bartuś T., Miary wartości przeciętnej, URL: <http://home.agh.edu.pl/~bartus/click.php?id=296>.
- Bartuś T., Szeregi rozdzielcze, URL: <http://home.agh.edu.pl/~bartus/click.php?id=306>.
- Bartuś T., Miary asymetrii, URL: <http://home.agh.edu.pl/~bartus/click.php?id=301>.
- Bartuś T., Miary skupienia, URL: <http://home.agh.edu.pl/~bartus/click.php?id=302>.
- Minn M., 2023. *Classification in ArcGIS Pro*. URL: <https://michaelminn.net/tutorials/arcgis-pro-classification/index.html>.