# Analiza 1

Jakub Szarafin

# Wstęp

Podczas tego projektu przeprowadzę analizę zarejestrowanych temperatur w latach 2008-2018 przez stację meteorologiczną w Borusowie-wsi położonej w południowej Polsce w województwie małopolskim, w powiecie dąbrowskim, w gminie Gręboszów.


Projekt ten będzie składał się z trzech części. W pierwszej sekcji posługując się biblioteką "gamlls"dopasuję do danych najlepszy z zaimplementowanych w niej rozkład, następnie obliczę 20-letnie oraz 50-letnie poziomu zwrotu. Druga sekcja skupia się na metodzie maksimów blokowych, w oparciu o maksima roczne, wyestymuję parametry rozkładu GEV oraz ocenię dobroć dopasowania za pomocą wykresów diagnostycznych, a następnie obliczę poziomy zwrotu. W trzeciej sekcji zajmę się metodą przekroczeń progu, posługując się przy tym bibliotekami evir i ismev, zamieszczę wykresy diagnostyczne i obliczę poziomy zwrotu dla każdej pory roku, dla indywidualnie dobranego progu.

## O danych

Dane, którymi się posługuję to maksima 10-minutowe. Podstawowe statystyki (już po oczyszczeniu danych) zamieszczam w poniższej tabeli:

| | **Lato** | **Jesień** | **Zima** | **Wiosna** |
|---|---|---|---|---|
| **Średnia** | 19.466 | 9.773 | -0.261 | 9.721 |
| **Odchylenie standardowe** | 5.268 | 6.321 | 5.551 | 7.022 |
| **Ilość obserwacji** | 145532 | 143740 | 137664 | 144687 |
| **Minimum** | 4.62 | -21.11 | -26.47 | -17.24 |
| **Maksimum** | 37.8 | 35.8 | 16.99 | 30.92 |
| **Ilość wartości odstających i NA** | 196 | 404 | 5328 | 1041 |

Miesiące przypisane do poszczególnych pór roku:
Wiosna:

- Marzec
- Kwiecień
- Maj

Lato:

- Czerwiec
- Lipiec
- Sierpień

Jesień:

- Wrzesień
- Październik
- Listopad

Zima:

- Grudzień
- Styczeń
- Luty

# Sekcja 1

W dobraniu odpowiedniego rozkładu pomoże nam kryterium Akaike information criterion (AIC), które jest estymatorem błędu przewidywań. Pozwala nam znaleźć rozkład, najmniej różniący się od naszego zbioru danych.

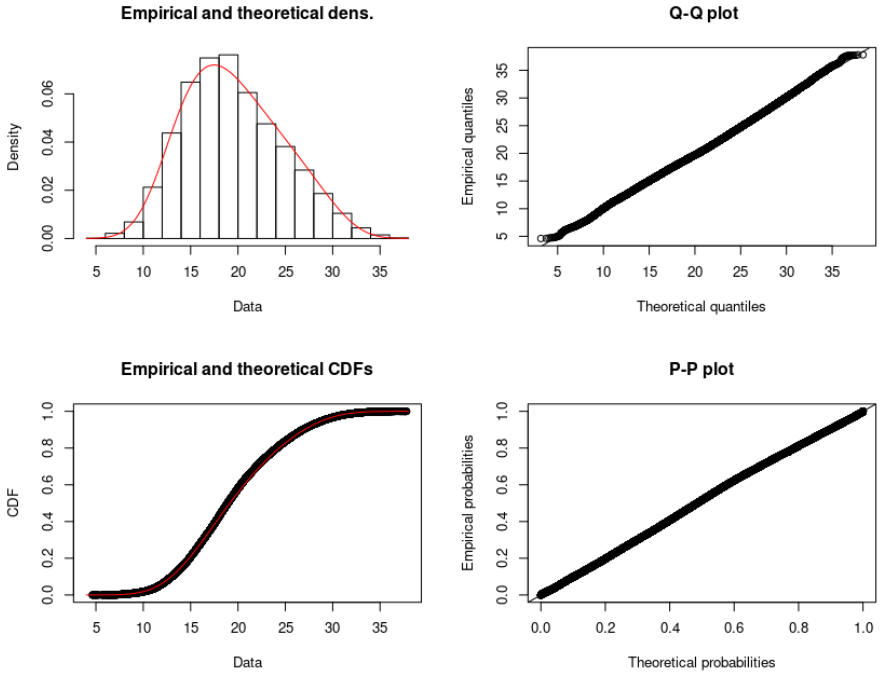
## Lato

W pierwszej kolejności, po wyizolowaniu danych letnich ze zbioru przechodzę do ich czyszczenia. W moim zbiorze znalazło się 196 wierszy z NA, w związku z czym zostały one usunięte. Po próbie dopasowania rozkładu dla tych danych otrzymałem następujące wyniki:
SHASH: 890805.8 SHASHo: 890814.8 SHASHo2: 890856.9 SEP1: 891043.1
Zgodnie z kryterium AIC wybrałem ten, który uzyskał najmniejszą wartość, czyli "SHASH"- Sinh-Arcsinh, wyraża się on następującym wzorem.

$$X = \mu + \delta \cdot \sinh[\frac{\sinh^{-1}(Z) + v}{\tau}].$$

$\mu$ kontroluje położenie rozkładu (gdzie jest on "wyśrodkowany"),
$\sigma$ kontroluje skalę (im jest większa, tym bardziej rozkład jest rozłożony),
$\nu$ kontroluje asymetrię rozkładu (może mieć dowolną wartość rzeczywistą, większa wartość dodatnia oznacza większą prawoskośność, większa wartość ujemna - większą lewoskośność),
$\tau$ kontroluje wagę ogona.
Na poniższych wykresach przedstawione zostają kolejno:

- Histogram prezentujący dane z naniesionym i dopasowanym rozkładem.
- Q-Q plot, który jest graficzną metodą porównywania dwóch rozkładów prawdopodobieństwa przez wykreślanie ich kwantyli względem siebie.
- Porównanie empirycznej i teoretycznej dystrybuanty.
- P-P plot to wykres prawdopodobieństwa służący do oceny zgodności dwóch zestawów danych, który przedstawia dwie funkcje rozkładu skumulowanego względem siebie.




Możemy wizualnie ocenić, że zgodność dopasowanego rozkładu do danych empirycznych jest wysoka.

## Poziomy zwrotu

W okolicach stacji, średnio raz na 20 lat (na 50 lat), w miesiącach letnich możemy spodziewać się temperatur co najmniej wielkości 43.937 (44.906).

# Wszystkie pory roku

| | **Wiosna** | **Lato** | **Jesień** | **Zima** |
|---|---|---|---|---|
| **Rozkład** | SEP1 | SHASH | SEP2 | SEP4 |
| **Poziom zwrotu** $x_{20}$ | 37.498 | 43.937 | 42.329 | 20.7909 |
| **Poziom zwrotu** $x_{50}$ | 38.296 | 44.906 | 44.174 | 21.7234 |

Najwyższe przewidywane poziomy zwrotu obserwujemy dla lata, są one bardzo wysokie i wydają się być w rzeczywistości mało prawdopodobne.

# Sekcja 2

## BMM-Maksima blokowe

Wiadome jest to, że maksima $M_n$ po odpowiedniej normalizacji, zbiegają według rozkładu do jednego z trzech rozkładów: Weibulla, Gumbela lub Frecheta. Wyestymujemy paramety uogólnionego rozkładu wartości ekstremalych, w którego skład wchodzą trzy wyżej wymienione rozkłady. GEV wyraża się następującymi wzorami:

$$H_\xi(x) = \exp\left(-\left(1+\xi\cdot\frac{x-\mu}{\sigma}\right)_+^{-\frac{1}{\xi}}\right) \quad \text{dla } \xi \neq 0$$

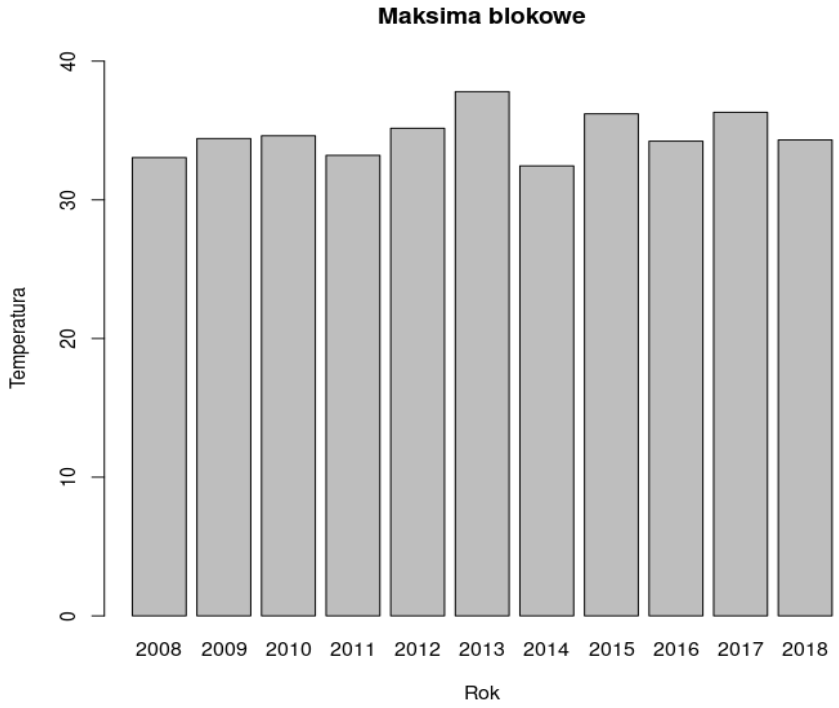
$$H_0(x) = \exp\left(-\exp\left(-\frac{x-\mu}{\sigma}\right)\right) \quad dla \quad \xi = 0$$

W tej metodzie skupiamy się na estymacji wartości ekstremalnych, a nie na dopasowywaniu rozkładu do całości danych. Poprawne użycia metody wygląda następująco:

1. Dzielimy zbiór danych na rozłączne bloki o równej długości.
2. Z każdego bloku wybieramy wartość największą.
3. Na podstawie wybranych danych estymujemy parametry rozkładu GEV.

## Lato

Dysponujemy danymi z 11 lat, więc taką też przejmuję ilość bloków. Poniżej znajduje się histogram i tabela z najwyższymi wartościami w danym bloku.



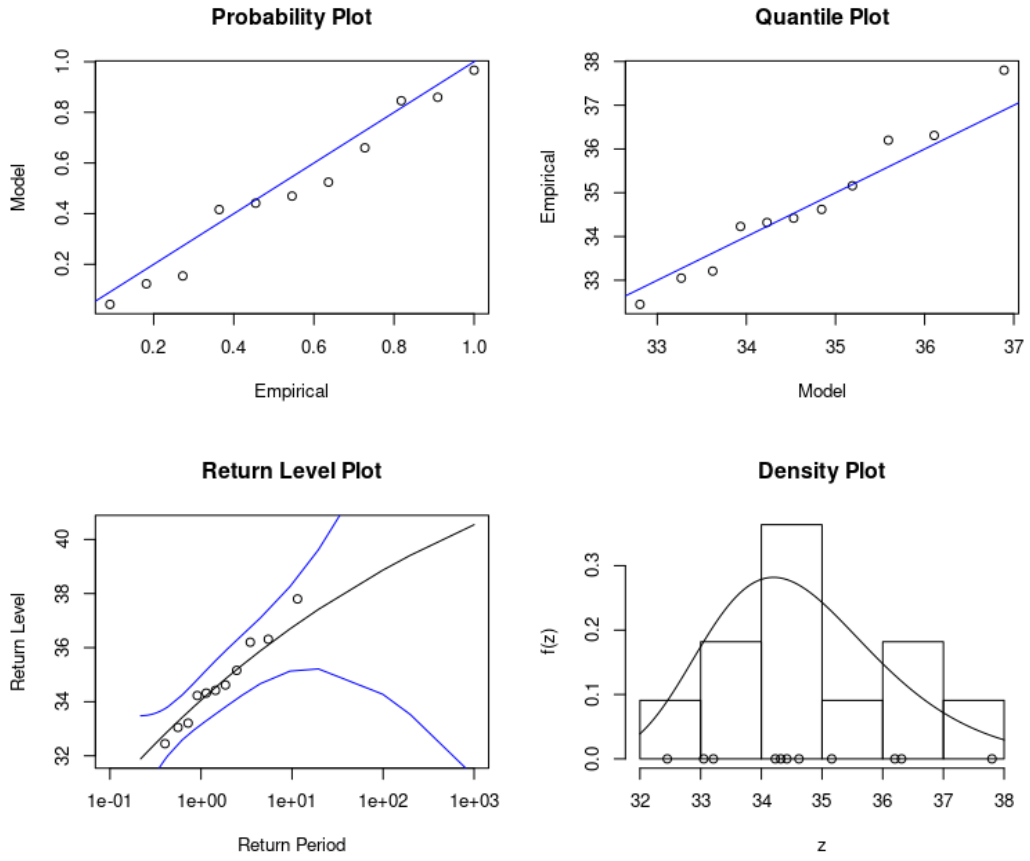

Maksimum z każdego bloku:

| **2008** | **2009** | **2010** | **2011** | **2012** | **2013** | **2014** | **2015** | **2016** | **2017** | **2018** |
|---|---|---|---|---|---|---|---|---|---|---|
| 33.05 | 34.42 | 34.62 | 33.21 | 35.16 | 37.80 | 32.45 | 36.20 | 34.23 | 36.31 | 34.32 |

Wyestymowane parametry rozkładu GEV:

| $\xi$ | $\sigma$ | $\mu$ |
|---|---|---|
| $-0.1029706$ | 1.3121663 | 34.0584911 |

Poniżej znajdują się wykresy diagnostyczne.




Wykresy informują nas o w całkiem dobrym, ale z drobnymi odstępami dopasowaniu do danych. W przypadku wykresu kwantyli wysokie wartości nieco odstają od przewidywanych, w przypadku gęstości również możemy zaobserwować nieco niedokładne dopasowanie.

## Poziomy zwrotu

Według rozkładu GEV w okolicach stacji, średnio raz na 20 lat (na 50 lat), w miesiącach letnich możemy spodziewać się temperatur co najmniej wielkości 37.41627 (38.27487).

# Wszystkie pory roku

| | **Wiosna** | **Lato** | **Jesień** | **Zima** |
|---|---|---|---|---|
| **Poziom zwrotu** $x_{20}$ | 30.74004 | 37.41627 | 34.32480 | 16.52759 |
| **Poziom zwrotu** $x_{50}$ | 31.23025 | 38.27487 | 35.76445 | 16.97801 |

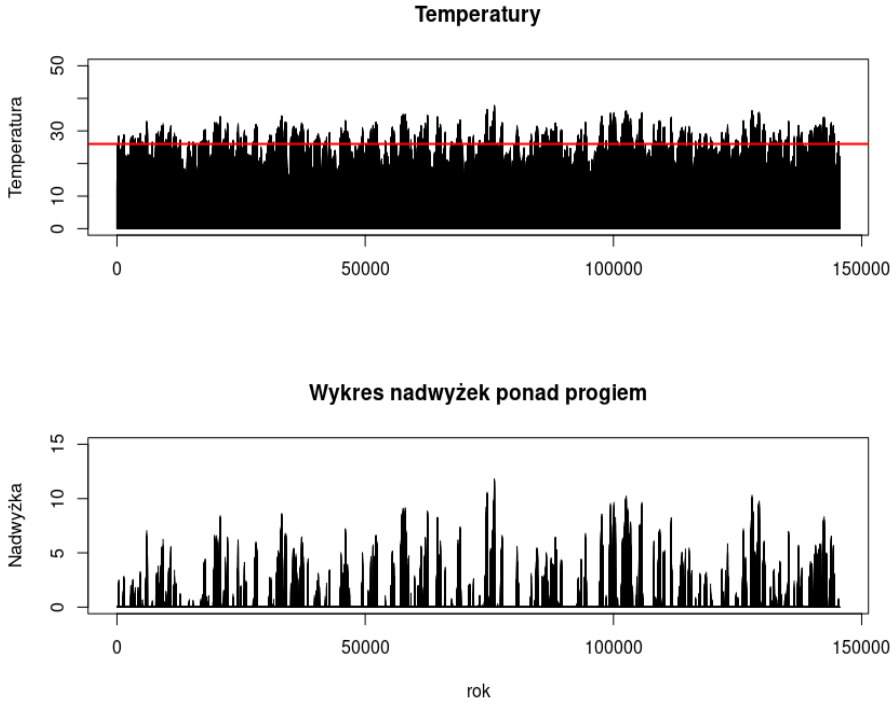
Najwyższe przewidywane poziomy zwrotu przewidywane są dla miesięcy letnich. Największa pomiędzy poziomami zwrotu wynosi

# Sekcja 3
# Metoda przekroczeń progu (POT)

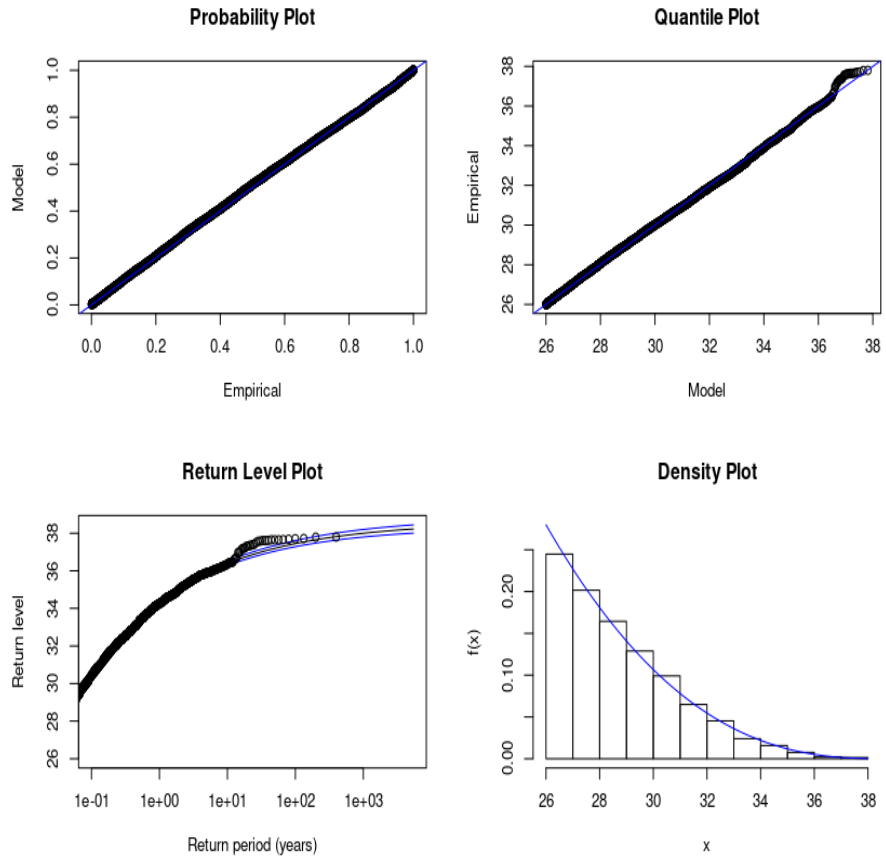
Metoda POT, to sposób estymacji $k-$letniego poziomu zwrotu $x_k$, w której nadwyżki ponad wybrany próg modelowane są rozkładem GPD. Próg ten ustalamy z góry tak, by około 10% naszych danych znalazło się ponad nim. Dystrybuanta tego rozkładu wyraża się następującymi wzorami:

$$G_{\xi,\beta} = \begin{cases} -\left(1+\frac{\xi \cdot x}{\beta}\right)^{-\frac{1}{\xi}} & \text{dla } \xi \neq 0 \\ 1-\exp\left(-\frac{x}{\beta}\right) & \text{dla } \xi = 0 \end{cases}$$

Poniżej przedstawiam wykres danych wraz z ustalonym progiem $u = 26$.




Na poniższych wykresach przedstawiona została dobroć dopasowania do danych letnich.




Na powyższych wykresach możemy zaobserwować bardzo dobre dopasowanie na całym zbiorze danych.

### Poziomy zwrotu

Przy pomocy tej metody okazuje się, że w okolicach stacji, średnio raz na 20 lat (na 50 lat), w miesiącach letnich możemy spodziewać się temperatur co najmniej wielkości 37.94144 (38.09129).

## Wszystkie pory roku

| | **Wiosna** | **Lato** | **Jesień** | **Zima** |
|---|---|---|---|---|
| **Próg** | 19 | 26 | 17 | 6 |
| **Poziom zwrotu** $x_{20}$ | 31.45199 | 37.94144 | 36.05007 | 16.98212 |
| **Poziom zwrotu** $x_{50}$ | 31.55782 | 38.09129 | 36.45265 | 17.2558 |

Jak możemy odczytać z tabeli różnica między poszczególnymi 20 i 50-letnimi poziomami zwrotu (w każdej porze roku) różni się niewiele. Największe poziomy temperatury przewidujemy dla lata.

## Podsumowanie

W przypadku użycia każdej z powyższych metod otrzymaliśmy największe przewidywane poziomy zwrotu dla miesięcy letnich. Metoda przekroczeń progu i metoda maksimów blokowych okazały się przewidywać zbliżone do siebie wyniki, natomiast wykorzystanie kryterium Akaike, do dopasowania rozkładu, prezentuje poziomy zwrotu znacznie wyższe niż pozostałe metody.