Analiza danych pogodowych

Stacja: Swieradów Zdrój

Adrian Frankowski

Wydział Matematyki, Fizyki i Informatyki
Uniwersytet Gdański

1 Wstęp

Głównym celem tej pracy jest wyestymowanie 20-letniego oraz 50-letniego poziomu zwrotu (który oznaczamy odpowiednio przez x_{20} i x_{20}) dla sezonu letniego oraz dla pozostałych pór roku.

Powiemy wówczas, że w okolicach danej stacji średnio raz na 20 lat (na 50 lat), w miesiącach letnich możemy spodziewać się temperatur co najmniej wielkości x_{20} (x_{50}).

2 Informacje o stacji

Nazwa stacji: Swieradów Zdrój
Kod stacji: X250150090
Długość i szerokość geograficzna: 15.35861, 50.89833

Rysunek 1: położenie stacji

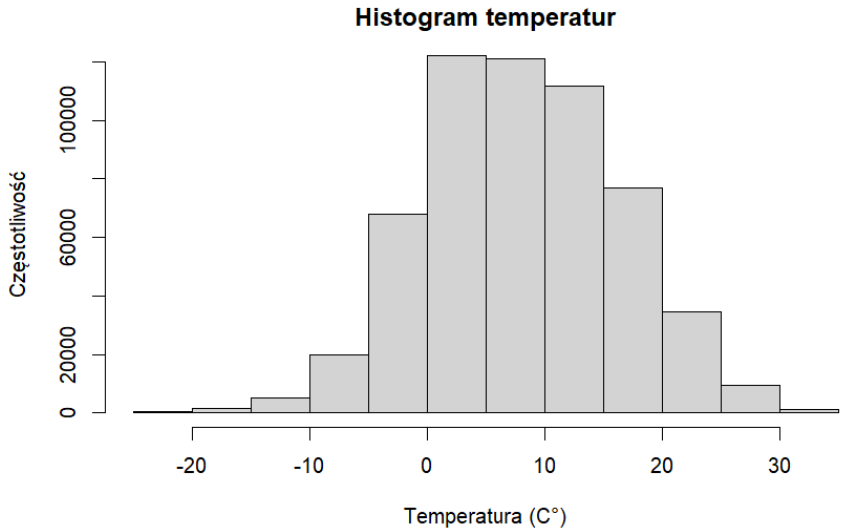
3 Dane

Będziemy pracować na danych 10-minutowych ze stacji Swieradów Zdrój z

lat 2008-2018. Każda z próbek została wyrażona w stopniach Stopień Celsjusza ($C°$) dlatego w pracy domyślnie będziemy korzystali ze skali Celsjusza. Szczegółowa charakterystyka danych prezentuje się następująco:

ilość próbek: 578592
ilość brakujących danych: 6847
temperatura minimalna: $-24.34°C$
temperatura maksymalna: $33.58°C$
średnia: $8.07°C$
mediana: $7.82°C$
odchylenie standardowe: $8.26°C$

W danych znajduje się wiele pustych i nierealistycznych wartości (bliskich $-50°C$). Po ich usunięciu możemy pokazać jak wygląda rozmieszczenie poszczególnych wartości w danych przedziałach liczbowych wykorzystując histogram (wykres poniżej).



Rysunek 2: Histogram temperatur

4 Dopasowanie najlepsze rozkładu

Korzystając z biblioteki gamlls, do maksimów 10-minutowych możemy dopasować najlepsz rozkład tam zaimplementowanych. Wyboru dokonamy korzystając z kryterium Akaike (AIC). Wybieramy z kilku modeli ten, który ma najmniejszą wartość

$$AIC = -2\sum_{i=1}^{n} \ln f_{\hat{\theta}}(X_i) + 2p,$$

gdzie $\hat{\theta}$ jest estymatorem MLE parametru θ, a p liczbą parametrów modelu.

W przypadku danych letnich najlepszym rozkładem okazał się:

Sinh-Arcsinh $SHASH(\mu, \sigma, \nu, \tau)$

Funkcja 4-parametrowego rozkładu Sinh-Arcsinh została opracowana w 2005 roku przez matematyka Chris'a Jones'a. Funkcje gęstości prawdopodobieństwa tego rozkładu oznaczamy przez

$$f_Y(y|\mu, \sigma, \nu, \tau) = \frac{c}{\sqrt{2\pi}\sigma(1+z^2)^{\frac{1}{2}}} exp(-\frac{1}{2}r^2).$$

Dopasowanie rozkładów dla każdej z poszczególnych pór roku prezentuje się następująco:

Pora roku	lato	jesień	zima	wiosna
Rozkład	SHASH	SEP2	SEP4	SEP1

5 Parametry rozkładu SHASH

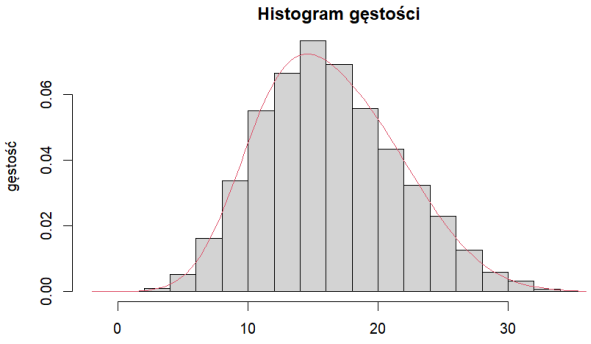
Dla danych letnich udało się wyestymować następujące parametry rozkładu:

$$\mu = 15.94, \sigma = 6.97, \nu = 1.35, \tau = 1.10.$$

6 Analiza dobroci dopasowania rozkładu SHASH

Korzystając z obliczonych parametrów możemy dokonać analizy dobroci dopasowania naszego rozkładu do naszych danych. Poniżej znajdują się 3 wykresy.

Histogram gęstości przestawia dopasowanie funkcji gęstości rozkładu do histogramu temperatur.

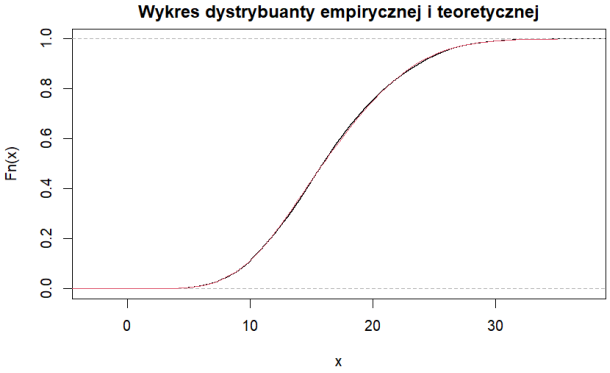


Rysunek 3: histogram gęstości

Wykres kwantyl-kwantyl (Q-Q plot) pokazuje dopasowanie kwantyli empirycznych i teoretycznych.

Rysunek 4: wykres kwantyl-kwantyl

Na wykresie CDF możemy zobaczyć dopasowanie dystrybuanty empirycznej i teoretycznej.



Rysunek 5: Wykres dystrybuanty empirycznej (kolor czarny) i teoretycznej (kolor czerwony)

Jak widzimy na powyższych wykresach rozkład SHASH znakomicie dopasował się do analizowanych danych.

7 Poziomy zwrotu

W tej części zajmiemy się wyestymowaniem 20-letniego i 50-letniego poziomu zwrotu. Poziomy zwrotu pokazują nam jakich maksymalnych temperatur możemy się spodziewać w danym przedziale czasu. Aby obliczyć 20-letni poziom zwrotu (odpowiednio 50-letni) posłużymy się równaniem $x_{20} = q(1 - \frac{1}{k})$, gdzie $k = 20 * 92 * 24 * 6$ (odpowiednio $x_{50} = q(1 - \frac{1}{k})$, gdzie $k = 50 * 92 * 24 * 6$).

W przypadku danych letnich otrzymaliśmy, że $x_{20} = 40.94$ oraz $x_{50} = 41.92$ co oznacza, że w przeciągu 20 lat możemy oczekiwać temperatury rzędu $40.94°C$

(odp. w przeciągu 50 lat możemy oczekiwać temperatury rzędu $41.92°C$). Poziomy zwrotu (x_{20} i x_{50}) dla innych pór roku wyglądają następująco:

	lato	jesień	zima	wiosna
x_{20}	40.94	36.08	16.85	33.24
x_{50}	41.92	37.12	17.48	33.65

8 Rozkład GEV

W tej części pracy zajmiemy się wyestymowaniem parametrów GEV w oparciu o maksima roczne. W rachunku prawdopodobieństwa i statystycę uogólnionym rozkładem wartości ekstremalnych (ang. generalized extreme value (GEV) distribution) nazywamy rodzinę ciągłych rozkładów prawdopodobieństwa stworzoną w ramach teorii wartości ekstremalnych aby połączyć rodziny rozkładów Gumbel'a, Fréchet'a i Weibull'a. Rozkład GEV jest jedynym możliwym rozkładem granicznym właściwie znormalizowanych maksimów ciągu niezależnych i identycznie rozłożonych zmiennych losowych. Należy zauważyć, że musi istnieć rozkład graniczny, który wymaga warunków regularności na krańcach rozkładu. Mimo to rozkład GEV jest często używany jako przybliżenie do modelowania maksimów długich (skończonych) sekwencji zmiennych losowych.

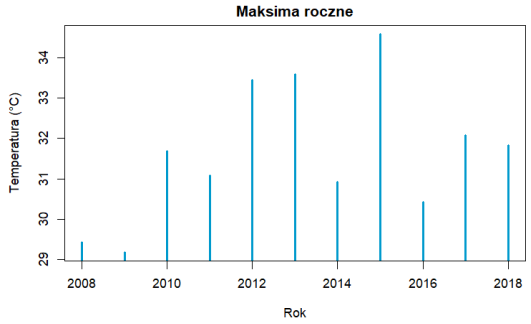
Na tę chwilę ograniczymy się do danych z sezonu letniego. Za pomocą biblioteki evir możemy wyestymować parametry rozkładu GEV(μ, σ, ξ), których wartości zostały przedstawione poniżej.

$$\mu = 31.06,\ \sigma = 1.55,\ \xi = -0.25,$$

gdzie:
μ - parametr położenia,
σ - parametr skali,
ξ - parametr kształtu.

9 Analiza dobroci dopasowania rozkładu GEV

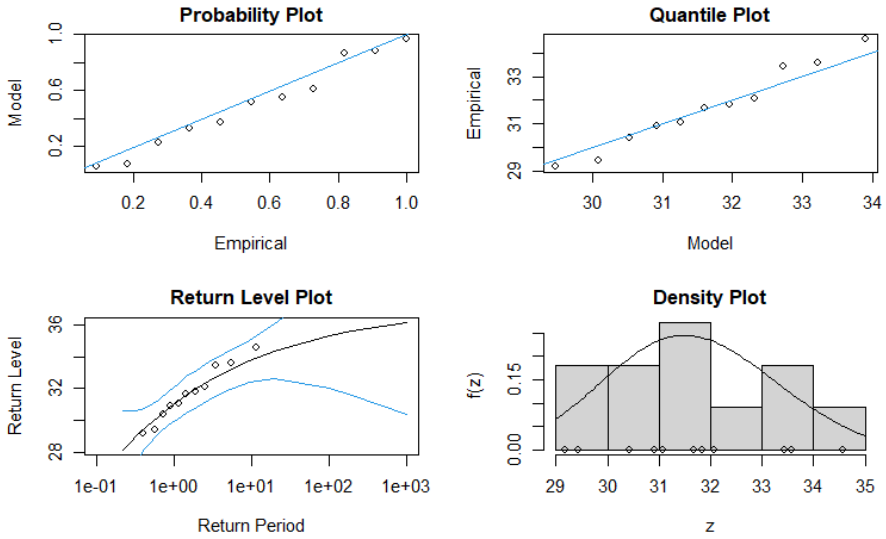
Analizę dobroci dopasowania rozkładu GEV rozpoczniemy od obliczenia maksimów rocznych, znajdujących się na wykresie poniżej.



Rysunek 6: maksima roczne

Jak widzimy najniższe maksimum zostało zarejestrowane w 2009 roku i wyniosło 29.18 $^{\circ}C$. Najwyższa z kolei temperatura wyniosła 34.57 $^{\circ}C$ i została zarejestrowana w roku 2015.

Aby dokonać bardziej zaawansowanej analizy dopasowania rozkładu skorzystamy z biblioteki ismev. Biblioteka ta zawiera zestaw wielu przydatnych narzędzi w tym pozwala narysować dodatkowe wykresy diagnostyczne opisane poniżej.



Rysunek 7: wykresy diagnostyczne z biblioteki ismev

- **Probability Plot** - Wykres prawdopodobieństwa służy do wizualnego porównania danych pochodzących z dwóch różnych zbiorów danych (rozkładów).
- **Quantile Plot** - W statystyce wykres Q-Q jest wykresem prawdopodobieństwa, który jest graficzną metodą porównania dwóch rozkładów prawdopodobieństwa poprzez wykreślenie ich kwantyli względem siebie.
- **Return level Plot** - Wykres poziomu zwrotu jest wykresem poziomu, który ma zostać przekroczony przez proces średnio raz na T-lat (poziom zwrotu y_T) w stosunku do (logarytmicznego) okresu zwrotu T.
- **Density Plot** - Wykres gęstości jest reprezentacją rozkładu zmiennej numerycznej. Pokazuje on funkcję gęstości prawdopodobieństwa danej zmiennej. Można powiedzieć, że jest to wygładzona wersja histogramu i jest używana w tym samy celu.

Jak możemy zauważyć powyższe wykresy potwierdzają poprawne dopasowanie rozkładu GEV do analizowanych danych.

Na koniec obliczymy poziomy zwrotu ($x_{20} = q(0.95)$ i $x_{50} = q(0.98)$) dla wszystkich pór roku. W tym celu ponownie wykorzystamy bibliotekę evir. Otrzymujemy, że:

	lato	jesień	zima	wiosna
x_{20}	34.31	30.53	14.49	28.90
x_{50}	34.92	31.65	15.35	29.18

10 Metoda przekroczeń progu (POT)

Do analizy w tej części pracy wykorzystamy maksima 10-minutowe. Próg dla lata został odgórnie ustalony na $u = 27°C$. Następnie wyestymujemy parametry rozkładu GPD dla nadwyżek nad ten próg oraz przeprowadzimy analizę oceniającą dobroć dopasowania. Na koniec wyznaczymy poziomy zwrotu x_{20} i x_{50} oraz przeprowadzimy uproszczoną analizę dla pozostałych pór roku.

Metoda POT, to sposób estymacji k-letniego poziomu zwrotu x_k, w której nadwyżki ponad wybrany próg modelowane są rozkładem GPD. Niech $X_1, \ldots, X_n$ będą niezależnymi zmiennymi losowymi o jednakowym rozkładzie o nieznanej dystrybuancie F, natomiast $Y1, \ldots, Y_{N_u}$ to nadwyżki nad wybrany „duży" próg u.

W statystycę uogólnionym rozkładem Pareto (GPD) nazywamy rodzinę ciągłych rozkładów prawdopodobieństwa. Używany jest on często do modelowania "ogonów" innych rozkładów prawdopodobieństwa. Rozkład GPD jest rozkładem o parametrach ξ, β.
Dystrybuanta uogólnionego rozkładu Pareto dana jest wzorem:

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

gdzie $\beta > 0$ oraz:

$$\begin{cases} x \geqslant 0 & \text{dla} \quad \xi \geqslant 0, \\ 0 \leqslant x \leqslant -\frac{\beta}{\xi} & \text{dla} \quad \xi < 0 \end{cases}$$

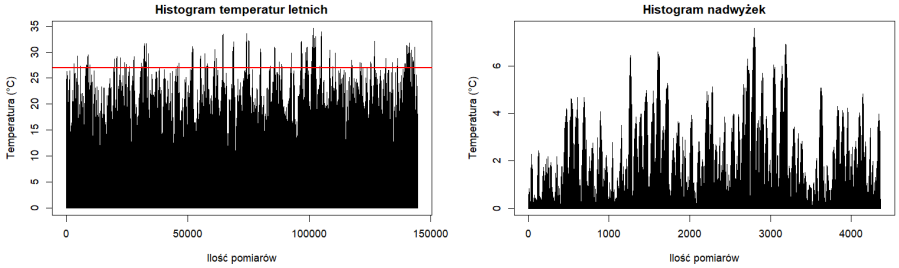
Za pomocą biblioteki ismev możemy westymować następujące parametry rozkładu GPD:
$\xi = -0.33$,
$\beta = 2.70$.

11 Analiza dobroci dopasowania rozkładu GPD

Wyzwaniem w metodzie GPD jest wybór progu u, ponieważ:

- Za niski próg może dawać za dużo danych pochodzących nie z ogona rozkładu, co zwiększy obciążenia estymatorów.
- Wybór zbyt dużego progu spowoduje, że pozostanie nam za mało danych, co wpłynie na wariancję.
- Wielu autorów sugeruje taki wybór progu, aby liczba pozostałych danych była nie większa niż 10-15%. Często wybierany jest próg tak aby liczba ta była rzędu 5-10%.

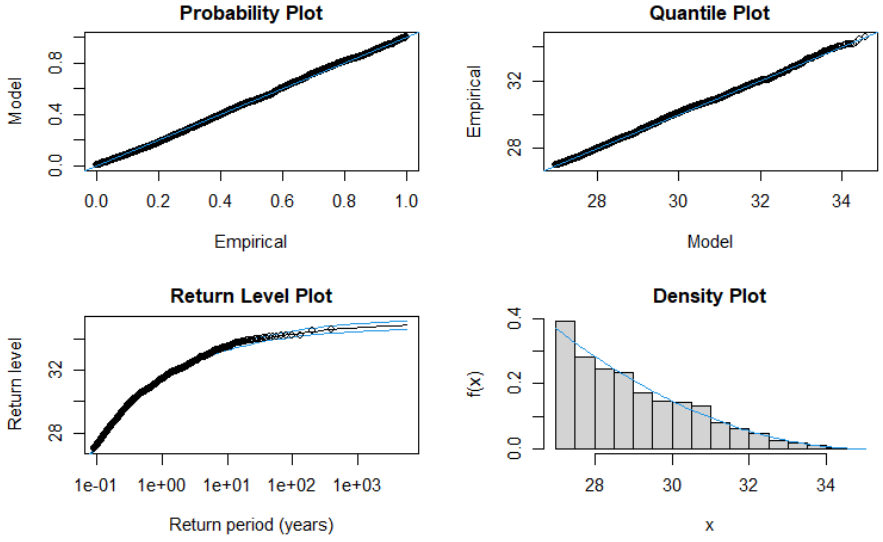
Analizę dopasowania dobroci rozkładu GPD rozpoczniemy od narysowania histogramów przedstawiających rozkład temperatur przed i po ustaleniu progu $u = 27°C$ dla danych z okresu letniego. W tym przypadku ilość pozostałych danych wynosi 3.01%.



(a) kolorem czerwonym zaznaczono próg $27°C$ (b) nadwyżki nad próg $27°C$

Rysunek 8: Histogramy temperatur

W celu narysowania bardziej szczegółowych wykresów ponownie posłużymy się funkcją biblioteką ismev. Biblioteka ta jest prosta w użyciu oraz dostarcza komplet wykresów diagnostycznych znajdujących się poniżej. Jak widzimy rozkład GPD prawidłowo dopasował się do naszych danych. W szczególności warto spojrzeć na wykres dopasowania gęstości **(Density Plot)**, który został dokładniej opisany w analizie rozkładu GEV.



Rysunek 9: wykresy diagnostyczne z biblioteki ismev

Na koniec obliczymy poziomy zwrotu dla pozostałych pór roku. W tym celu możemy posłużymyć się następującymi wzorami (lub skorzystać z gotowych bibliotek):

$x_k = u + \frac{\beta}{\xi}((k \cdot p)^{\xi} - 1)$ dla $\xi \neq 0$,

$x_k = u + \beta \ln(k \cdot p)$ dla $\xi = 0$.

Progi u zostały dobrane tak aby ilość pozostałych danych mieściła się w przedziale od 5 do 15%. Aby wyestymawać poziomy zwrotu (x_{20}, x_{50}) posłużymy się biblioteką evir. Tabela wyników łącznie z progami prezentuje się następująco:

	lato	jesień	zima	wiosna
próg ($^\circ C$)	27	18	5	18
x_{20}	25.50	18.61	7.01	19.21
x_{50}	28.03	21.42	8.57	21.66

Podsumowanie

Celem niniejszej pracy było wyestymowanie 20-letniego oraz 50-letniego poziomu zwrotu dla sezonu letniego i pozostałych pór roku. W tym celu posłużyliśmy się interpretowanym języki programowania R, który doskonale sprawdza się w rozwiązywaniu problemów obliczeniowych. R dostarcza szeroką gamę technik statystycznych, które można rozszerzyć za pomocą dodatkowych pakietów. Szczególnie przydatny okazał się pakiet ismev zawierający gotowe narzędzia do rysowania wykresów diagnostycznych.

W celu obliczenia poziomów zwrotu posłużyliśmy się kilkoma różnymi metodami estymacji takimi jak: metoda maksimów blokowych (BMM) oraz metoda przekroczeń progu (POT). Do analizy danych wykorzystaliśmy szereg różnych rozkładów prawdopodobieństwa m.in. SHASH, GEV oraz GPD. Warto zwrócić uwagę, że każda z metod dostarczyła nam różnych wyników. Oznacza to, że nie ma jednego, najlepszego sposobu rozwiązania problemu postawionego w tej pracy.

Biorąc pod uwagę powyższe rozważania możemy stwierdzić, że każda z wymienionych metod dostarcza nam wielu ciekawych wniosków. Każda z tych metod ma swoje wady i zalety ale z pewnością warto przyjrzeć się im z osobna.