# Wprowadzenie do modelowania wartości ekstremalnych (*Extreme Value Theorem*)

Prognozowanie w hydrologii i klimatologii

22 marca 2015 r.

## 1 Wprowadzenie

Określenie powtarzalności występowania ekstremalnych zdarzeń naturalnych jest zagadnieniem problematycznym, głównie ze względu na brak wystarczająco długich serii pomiarowych. Z punktu widzenia praktyki (zwłaszcza inżynierskiej) często przed klimatologami i hydrologami stawiane są pytania natury praktycznej dotyczące np. sfery planowania inwestycyjnego, potencjalnego zagrożenia zdrowia mieszkańców, określenia zagrożenia powodziowego, itp.

Spośród stawianych pytań najczęściej pojawiają się dwie zasadnicze podgrupy, które w języku matematyki wyrażane są jako **okres powtarzalności** (ang. *return period*), np.:

- ile wynosi przepływ w danym profilu rzeki występujący raz na 20, 100, 200 lub X lat?

oraz **stopa zwrotu** (ang. *return value*):

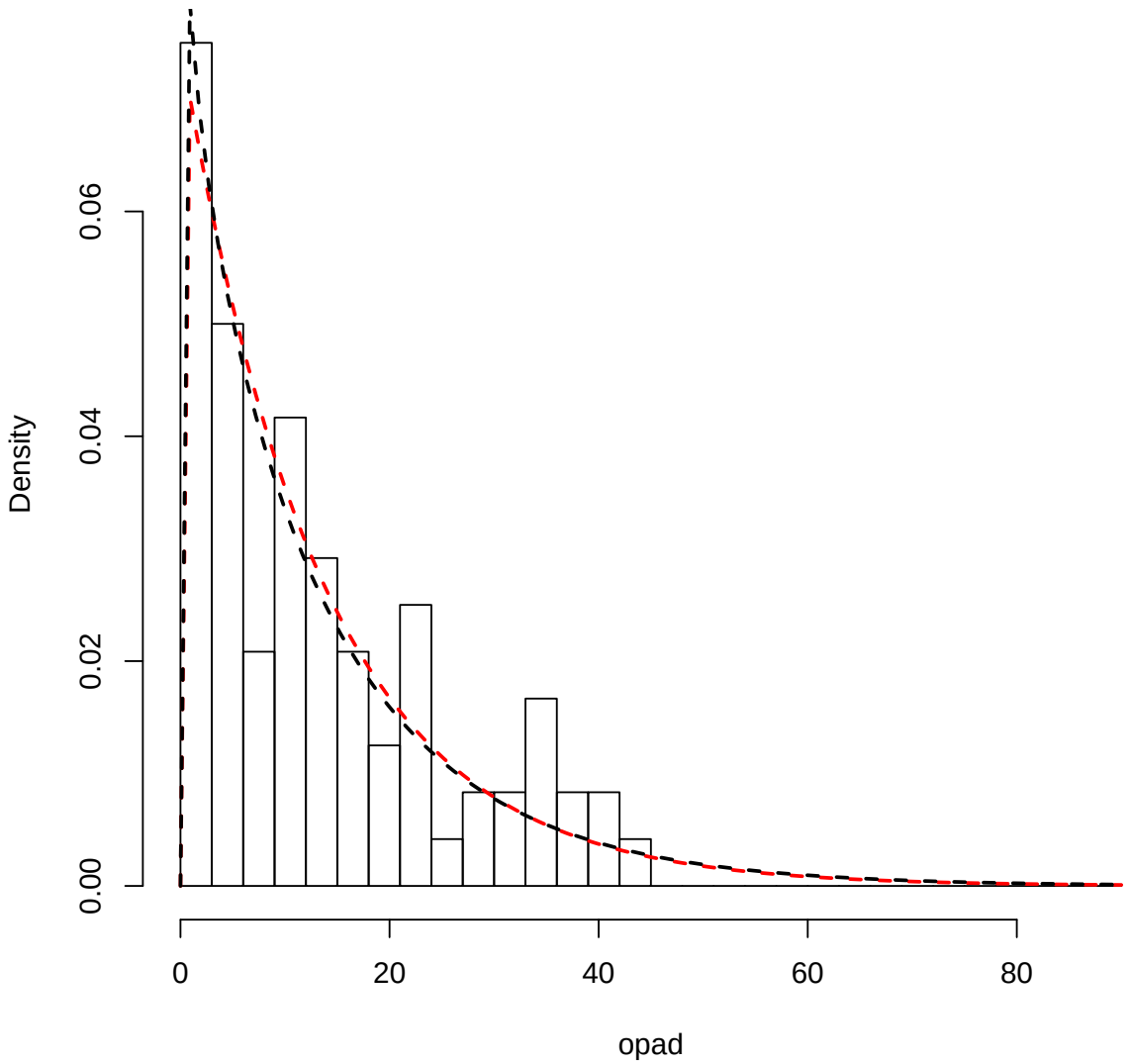
- jak często (raz na ile lat) występuje przepływ w rzece przekraczający np. $> 1000\ m^3/s$ ?

Brak długich serii pomiarowych w znacznym stopniu ogranicza możliwości stosowania podejścia probabilistycznego w klasycznym ujęciu. Dodatkowo kwestię problemową stanowi niezbyt dokładne dopasowanie najczęściej stosowanych rozkładów (np. rozkładu normalnego dla temperatury) w "ogonach" histogramu. Pewien ogólny zarys problemu dla rozkładów ciągłych można zauważyć np. za pomocą poniższego kodu w R:

```
library(fitdistrplus) # pakiet pozwalajacy na dopasowanie rozkladu
opad <- rexp(80,rate = 0.09) # generujemy losowa probe
wykladniczy=fitdist(opad,"exp") # testujemy rozklad wykladniczy
r_gamma=fitdist(opad,"gamma") # testujemy rozklad gamma (np. opady)

## Warning: wyprodukowano wartości NaN

hist(opad, prob = TRUE, breaks=0:30*3) # rysujemy histogram
curve(dexp(x, rate = wykladniczy$estimate), col = 2, lty = 2,
      lwd = 2, add = TRUE) # dodajemy rozklad wykladniczy
curve(dgamma(x, shape = r_gamma$estimate[1], rate=r_gamma$estimate[2]),
      col = 1, lty = 2, lwd = 2, add = TRUE) # i r. gamma
```




Dlatego też kluczową kwestię przy analizie zjawisk ekstremalnych stanowi dobór odpowiedniego rozkładu statystycznego, pozwalającego na ujęcie analizowanego zagadnienia w ujęciu probabilistycznym (Friedrichs i Thorarinsdottir 2012). Wśród rozkładów statystycznych pozwalających na określenie prawdopodobieństwa takiego zdarzania najczęściej stosuje sie dwa rodzaje rozkładów teoretycznych należacych do rodziny rozkładów Pareto (GPD, Generazlied Pareto Distribution) oraz rodziny rozkładów Fishera-Tippeta (GEV, Generalized Extreme Value).

W języku programowania R oba podejścia można zrealizować za pomocą pakietu `extRemes` specjalnie dedykowanego do tego celu. Szczegółowy opis pakietu dostępny jest na stronie:
www.ral.ucar.edu/ ericg/extRemes/

## 1.1 Teoria rozkładów asymptotycznych

- **Centralne twierdzenie graniczne** - uzasadnia powszechne występowanie w przyrodzie rozkładów zbliżonych do rozkładu normalnego. Średnia z dużej populacji losowych zmiennych ma rozkład normalny bez względu na rozkład "bazowy" z którego te średnie są liczone
- **Twierdzenie odnoszące się do rozkładów ekstremalnych** - maksimum dużej populacji zmiennych losowych ma rozkład: Gumbela, Frecheta lub Weibulla, niezależnie od rozkładu z którego te dane pochodzą.

Ogólna postać rozkładu ekstremalnego Fishera-Tippeta (GEV) może zostać wyrażona za pomocą funkcji gęstości prawdopodobieństwa przez:

$$t(x) = (1 + (\frac{x-\mu}{\sigma})\xi\big)^{-1/\xi} \qquad \text{dla } \xi \neq 0 \tag{1}$$

$$t(x) = e^{-(x-\mu)/\sigma} \qquad \text{dla } \xi = 0 \tag{2}$$

natomiast w rozróżnieniu na rozkłady ekstremalne I, II i III typu przybierają one postać: (poniższe wzory dla dystrybuant rozkładów)

- Rozkład Gumbela (bez ograniczeń):

$$G(x) = exp(-exp(-x)) \tag{3}$$

- Rozkład Frecheta (dla x>0 i $\alpha$ >0) - (ograniczenie lewostronne, dla niskich wartości)

$$G(x) = exp(-x^{-a}) \tag{4}$$

- Rozkład Weibulla (dla x<0 i $\alpha$ >0) - (ograniczenie prawostronne, dla wysokich wartości)

$$G(x) = exp(-(-x)^{a}) \tag{5}$$

## 2 Polecenia wprowadzające do zadania

1. W pierwszej kolejności należy zainstalować pakiet `extRemes` z którego będziemy wykorzystywać zarówno przykładowe dane, jak i zawarte w tym pakiecie funkcje:

   - polecenie: `install.packages("extRemes")` instaluje pakiet z repozytorium internetowego. Po jednorazowym zainstalowaniu na komputerze nie ma potrzeby ponownej instalacji.
   - poleceniem `library(extRemes)` aktywujemy poprzednio zainstalowany pakiet (nie jest on aktywny w momencie uruchamiania programu R lub RStudio, więc przy każdym ponownym uruchomieniu programu pakiet ten musi być ponownie wczytany do pamięci komputera)

   ```
   library(extRemes)
   ```

2. **Dane:** Maksymalne roczne przepływy rzeki Potomac w latach hydrologicznych (Paź-Wrz) 1895-2000 w miejscowości Point Rocks, Maryland.
   Jednostki = cfs (1 cfs = 0.028317 $[m^3/s]$ )
   **Podpowiedź:** W przypadku modelowania zjawisk ekstremalnych metodami EVD korzysta się z podejścia "block maxima", w którym wykorzystuje się jedynie wartości maksymalne w zdefiniowanych okresach czasu (czyli bierzemy pod uwagę tylko maksima roczne jeśli naszym celem jest chęć uzyskania prawdopodobieństwa wystąpienia zjawiska o okresie powtarzalności "X" lat).

```
data(Potomac) # po poprawnym uruchomieniu dane powinny byc widoczne w
# zakładce Global Environment
```

   - W pierwszej kolejności sprawdźmy strukturę naszych wczytanych danych za pomocą polecenia: `head` wyświetlającego zawartość pierwszych kilku wierszy. Ogólne statystyki możemy sprawdzić np. za pomocą funkcji `summary`. Następnie wyrysujmy przebieg zmienności międzyrocznej przepływów rzeki Potomac (do stworzenia wykresu użyjemy funkcji `plot`). Po wstępnym przeanalizowaniu danych widzimy, że wartości przepływów są dość duże, zatem wykres stwórzmy w przeliczeniu na $m^3$ zamiast $cfs$ (1 cfs = 0.028317 cubic meters per second). Domyślnie w `R` rysowany jest wykres punktowy. Jeśli chcemy zmienić rodzaj wykresu na liniowy, musimy uwzględnić dodatkowo opcję rysowania: `type="l"`.

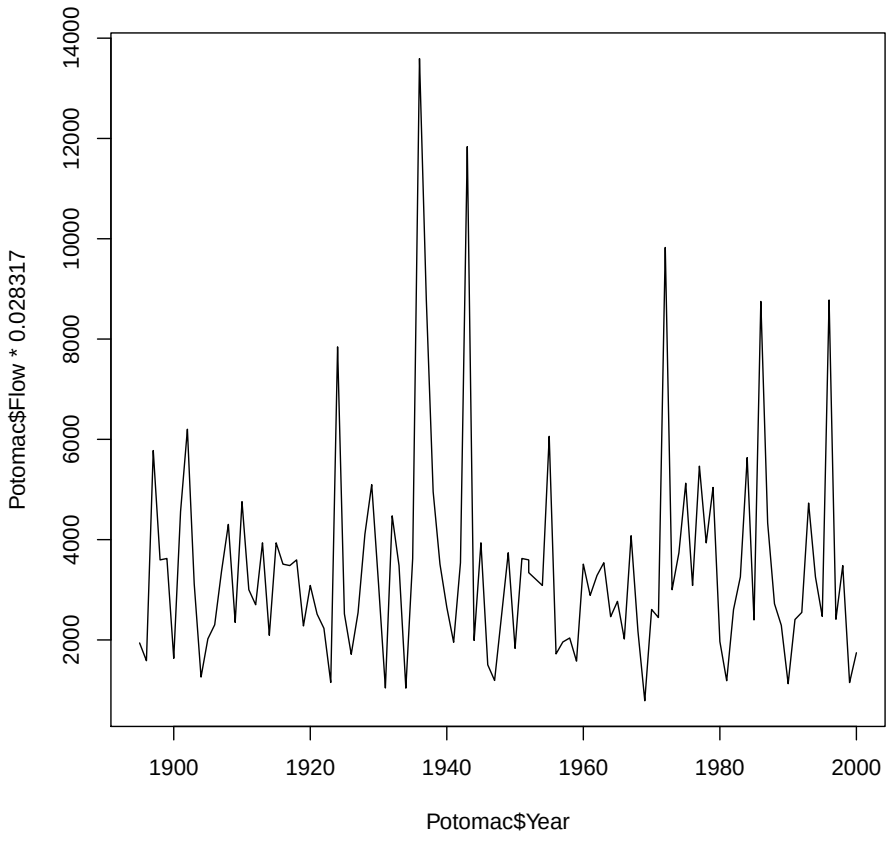
```
head(Potomac)

##   Year   Flow
## 1 1895  68500
## 2 1896  56000
## 3 1897 204000
## 4 1898 127000
## 5 1899 128000
## 6 1900  57700

summary(Potomac)
```

```
##       Year           Flow
##  Min.   :1895   Min.   : 27800
##  1st Qu.:1921   1st Qu.: 77300
##  Median :1948   Median :109000
##  Mean   :1948   Mean   :121949
##  3rd Qu.:1974   3rd Qu.:139000
##  Max.   :2000   Max.   :480000

plot(Potomac$Year, Potomac$Flow*0.028317, type="l")
```




3. Stworzone podsumowania wskazują, że w kilku latach maksymalny chwilowy przepływ przewyższył 10000 $m^3/s$. W praktyce, często interesuje nas ile było takich przypadków (powyżej zadanego progu), i w których latach takie przypadki miały miejsce?
   Można do tego celu wykorzystać wcześniej poznaną funkcję `which` aby znaleźć interesujące nas wartości:

```
which(Potomac$Flow*0.028317>10000) # otrzymujemy numery wierszy: 42 i 49 (2 przypadki)

## [1] 42 49
```

```
Potomac$Year[c(42,49)] # powinnismy otrzymac lata w ktorych przeplyw przekroczyl 1e4
```

```
## [1] 1936 1943
```

4. Aby określić prawdopodobieństwo dla wartości występujących stosunkowo rzadko ważne jest w pierwszej kolejności określenie rodzaju rozkładu, który w najlepszym stopniu będzie oddawał rozkład danych. Do tego celu zastosujemy polecenie `fevd`. Jako jeden z argumentów funkcji `fevd`, który należy uwzględnić dotyczy typu rozkładu do którego dane mają być dopasowane. Pakiet `extRemes` pozwala na dopasowanie danych do rozkładów: `GEV, GP, PP, Gumbel, Exponential`. Jeśli wyświetlimy histogram dla naszych danych (funkcja `hist`) powinniśmy dojść do wniosku, że warte przetestowania są przede wszystkim opcje numer 1 (`GEV`) i 4 `Exponential`.

5. Stwórzmy zatem pierwszy model dla typu ogólnego rodziny rozkładów ekstremalnych (Generalized Extreme Value Distribution) w przeliczeniu na $m^3$:

```
model1=fevd(Potomac[,2]*0.028317,type="GEV")
```

6. Zapiszmy wartość parametru AIC którą można uzyskać po wywołaniu stworzonego obiektu: `model1` . Powinny nam się wyświetlić wszystkie parametry naszej funkcji, błędy oszacowania oraz parametry AIC oraz BIC

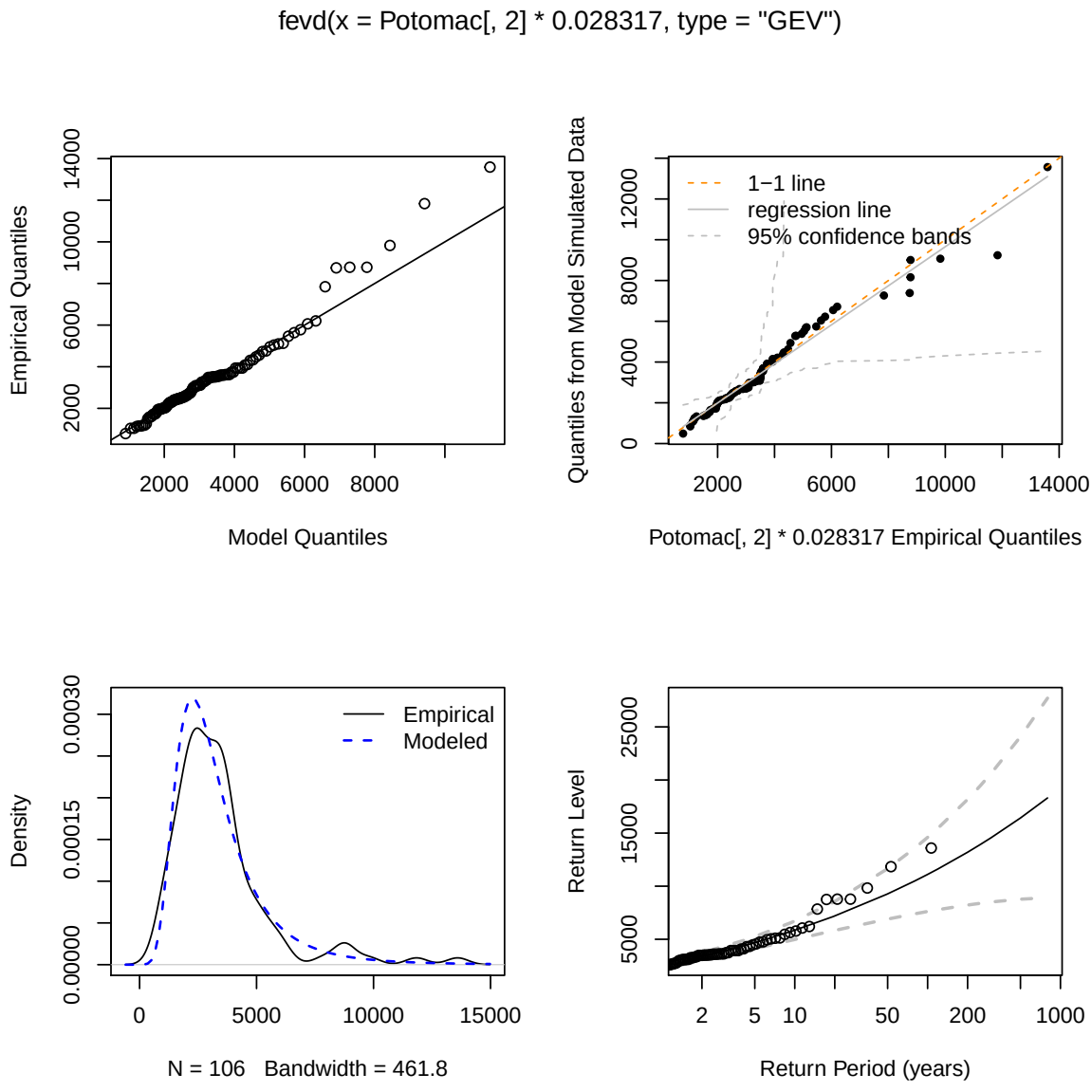
```
print(model1)
```

```
##
## fevd(x = Potomac[, 2] * 0.028317, type = "GEV")
##
## [1] "Estimation Method used: MLE"
##
##
##  Negative Log-Likelihood Value:  930.7
##
##
##  Estimated parameters:
##  location     scale     shape
## 2461.6825 1171.6934    0.1909
##
##  Standard Error Estimates:
##  location     scale     shape
## 129.71186  98.16734   0.07377
##
##  Estimated parameter covariance matrix.
##           location     scale     shape
## location 16825.168 7094.6102 -3.115367
## scale     7094.610 9636.8267 -0.477662
## shape       -3.115   -0.4777  0.005442
##
```

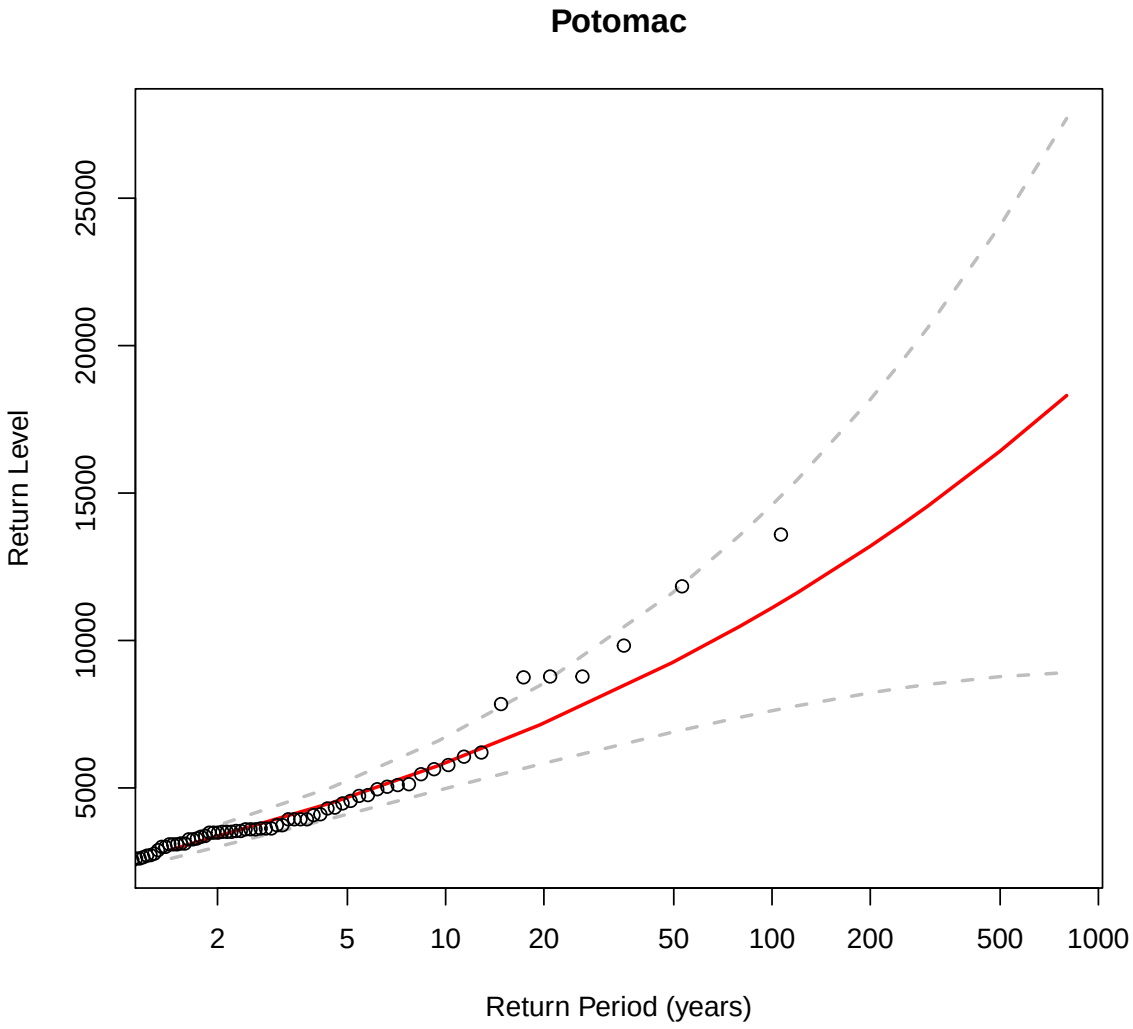
```
##  AIC = 1867
##
##  BIC = 1875
```

7. Bardzo ważnym etapem jest wykonanie i sprawdzenie wykresów diagnostycznych dla uzyskanego modelu. Dzięki nim możemy wizualnie określić jakość dopasowania danych do modelu, a także określić wiarygodność np. okresu zwrotu. Zapoznajmy się z poniższymi poleceniami:

```
plot(model1)
```




```
# poniewaz wczesniejsza funkcja dzieli ekran na 4 czesci
# musimy zresetowac ustawienia okna wyswietlajacego grafike
par(mfrow=c(1,1))
plot(model1, "rl", main="Potomac", col="red", lwd=2)
```




8. Obliczmy wartości dla często stosowanych w hydrologii okresów zwrotu wraz z podaniem przedziałów ufności na poziomie 1 - $\alpha = 0.05$

```
return.level(model1,do.ci=T) # opcji do.ci=T oznacza załącz współczynniki ufności

## fevd(x = Potomac[, 2] * 0.028317, type = "GEV")
##
## [1] "Normal Approx."
##
##                       95% lower CI Estimate 95% upper CI
## 2-year return level           2609     2906         3204
## 20-year return level          5806     7144         8482
## 100-year return level         7613    11093        14573
```

9. Utwórz analogiczny model dla rozkładu Gumbela. Następnie odpowiedz na pytanie który z testowanych modeli (Gumbel czy GEV) jest Twoim zdaniem bardziej wiarygodny? Odpowiedź uzasadnij.

# 3 Zadanie

1. Odpowiedz na ostatnie pytanie z poprzedniego rozdziału oraz podaj wartość przepływu występującego w rzece Potomac występującego wg rozkładu Gumbela ze 100-letnim okresem powtarzalności.

2. Na podstawie zbioru danych maksymalnych rocznych porywów wiatru w Wielkopolsce `evdporyw` umieszczonego na stronie Zakładu Klimatologii:
www.klimat.amu.edu.pl → `Dydaktyka` → `Materiały do zajęć dydaktycznych`
wykonaj następujące zadania:

   (a) Wczytaj i przeanalizuj dostarczony zbiór danych

   (b) Posługując się funkcją `fevd` przetestuj wszystkie możliwe rodzaje modeli z rodziny funkcji ekstremalnych dostępnych w pakiecie `extRemes`

   (c) Oceń i opisz jakość dopasowania każdego z tych modeli. Który z nich jest Twoim zdaniem najbardziej wiarygodny dla okresów powtarzalności powyżej 50 lat?

   (d) Posługując się najlepszym z uzyskanych modeli oblicz wartość porywu wiatru w Wielkopolsce występującą z okresem powtarzalności co 2, 10, 20, 50 i 100 lat? Użyj funkcji `return.level` wraz ze zdefiniowanymi przedziałami ufności (`do.ci=T`) oraz zadeklarowanymi okresami powtarzalności (opcja `return.period = c(2,10,20,50,100)`). Oprócz wartości najbardziej prawdopodobnej (estimate) zapisz w tabeli wartości dla dolnego i górnego ograniczenia (`lower CI, upper CI`)

   (e) Zadanie opcjonalne: Oblicz co ile lat może wystąpić w Wielkopolsce poryw wiatru przekraczający 50 m/s (180 km/h)?