Korekta danych modelowych z wykorzystaniem metody mapowania kwantylowego (ang. *quantile-quantile mapping*)

Wprowadzenie:

Modele stosowane w naukach atmosferycznych (np. (AO)GCM, RCM, NWP, itd.) nie są wolne od błędów. Częstym zabiegiem mającym na celu ich wyeliminowanie są tzw. metody korekcji "BIAS"-u, spośród których za jedną z najbardziej uniwersalnych jest uznawana metoda **mapowania kwantylowego**.

Dane:

Dane do niniejsze zadania dostępne sa na stronie: https://www.dropbox.com/s/dveqyga2akf6hl2/qmap.rdata?dl=0 ; Pobierz dane na dysk w znanej lokalizacji.

Jest to standardowy zbiór danych języka programowania R, który można wczytać za pomocą funkcji `load()`. Pamiętaj o ustawieniu katalogu roboczego!

Wczytany zbiór danych powinien zawierać 2 obiekty:

1. `dane1` z kolumnami:

- `obserwacje` – zbiór danych zawierający średnie dobowe temperatury powietrza na wysokości 2 m w latach 1971-2000 (1 losowo wybrany punkt dla obszaru Polski)
- `rcm_hist` - wyniki symulacji temperatury powietrza na wysokości 2 m według (losowo wybranego) regionalnego modelu klimatu (RCM) dla analizowanej lokalizacji.

2. `dane2` z kolumną:

- `rcm_scen` - wyniki symulacji jednego z RCP scenariuszowych zmian klimatu. Zakres danych: 2021-2050

```
load("qmap.rdata")
head(dane1, 3)
```

```
##    rok miesiac dzien obserwacje  rcm_hist
## 1 1971       1     1     -10.23 -7.977941
## 2 1971       1     2     -11.69 -4.440670
## 3 1971       1     3     -14.15 -5.761170
```

```
head(dane2, 3)
```

```
##    rok miesiac dzien   rcm_scen
## 1 2021       1     1  0.3733483
## 2 2021       1     2  0.4474250
## 3 2021       1     3 -0.8529596
```

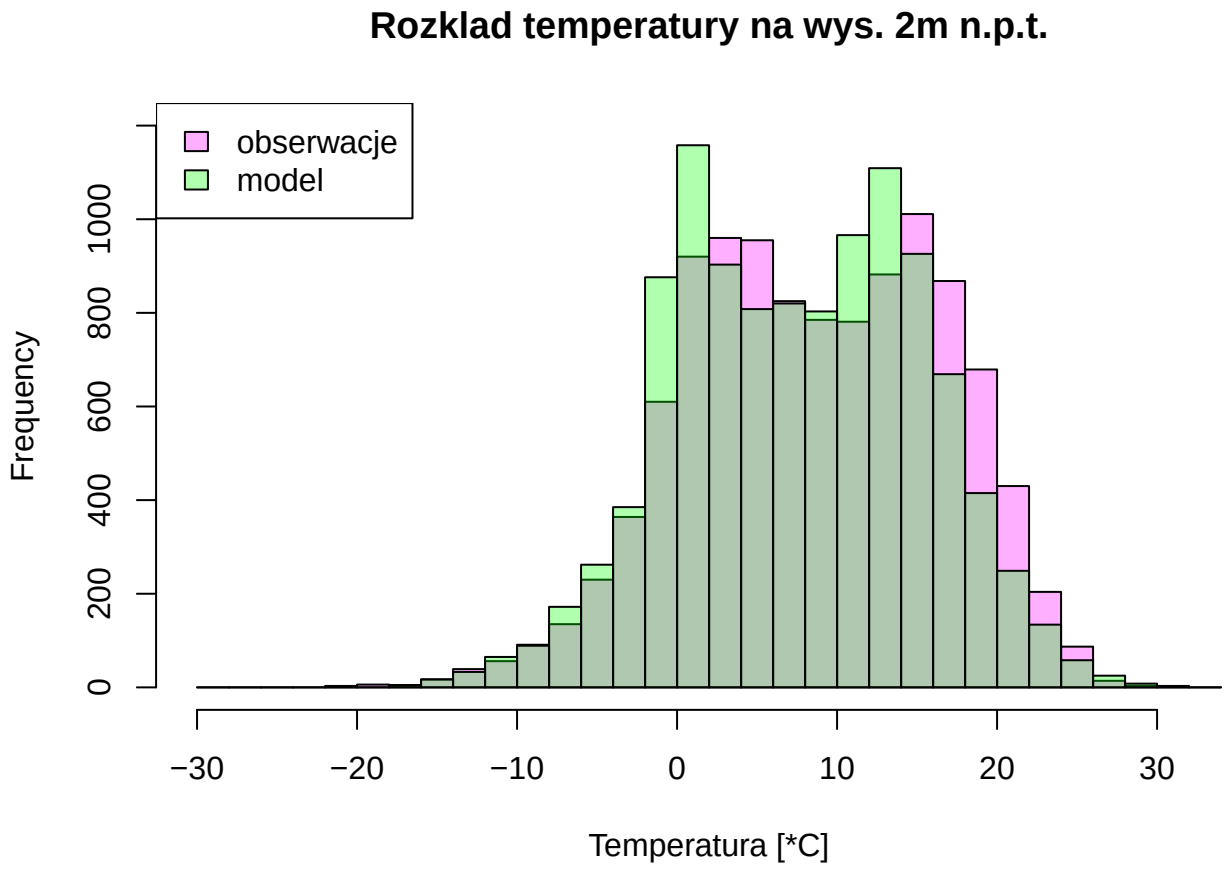
Rozkład danych

Modele GCM/RCM są tworzone z myślą o symulowaniu zmian klimatu a nie pogody (tj. symulowana wartość dla 03.06.2030 r. nie oznacza, że taka pogoda będzie w istocie tego dnia). Z tego względu konieczne jest spojrzenie na całą informację zawartą w symulowanym polu np. przez pryzmat statystycznego rozkładu danych opisującego klimat a nie pogodę.

Sprawdźmy zatem jak radził sobie z tym nasz model w okresie historycznym:

```
par(mfrow=c(1,2))
hist(dane1$obserwacje, main="OBSERWACJE")
hist(dane1$rcm_hist, main="MODEL")
```

... lub w nieco bardziej przystępny sposób te same dane można przedstawić jako:



Pytania kontrolne:

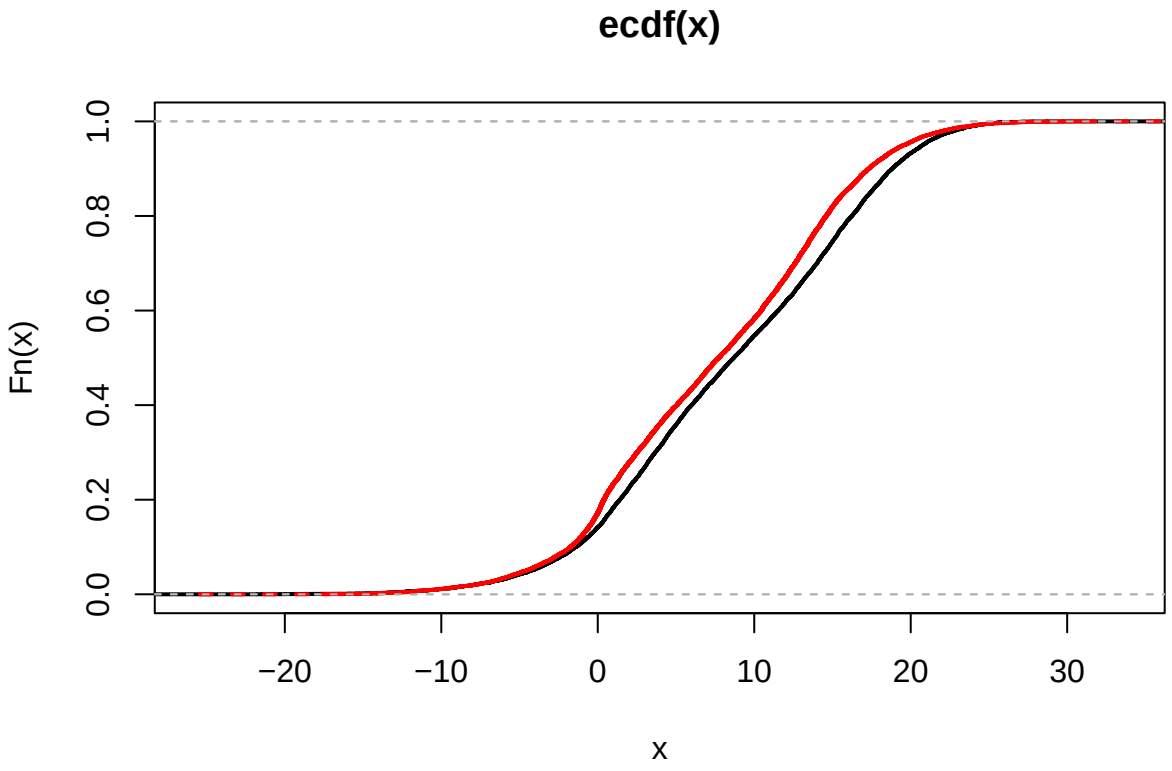
1. Która z serii danych w okresie historycznym jest cieplejsza (model czy obserwacje). Oblicz wartości średnie i ekstremalne dla obu serii
2. Dlaczego temperatury powietrza w okolicach 0*C występują częściej niż wynikało by z wartości średniej i (teoretycznego) rozkładu normalnego
3. Czy przesunięcie całego rozkładu (w lewo lub w prawo), które można łatwo uzyskać dzięki metodzie *delta-change* pozwoliło by na dobre dopasowanie obu serii danych na siebie? Jakie potencjalne problemy mogą wystąpić w przypadku elementów meteorologicznych ograniczonych jedno- lub dwustronnie (np. opad atmosferyczny, prędkość wiatru)?

Histogram (PDF) a dystrybuanta (CDF)

Informacja o rozkładzie danych wizualizowana w postaci histogramu to tzw. funkcja gęstości prawdopodobieństwa (ang. *probability density function*). Ten sam zbiór danych można także przedstawić za pomocą dystrybuanty (CDF, ang. *cumulative density function*).

W R łatwo wyświetlić dystrybuantę dzięki funkcji `plot.ecdf()`. Dorysowanie kolejnej dystrybuanty do istniejącego wykresu jest możliwe dzięki zastosowaniu parametru `add=T`. Pozostałe parametry sa analogiczne jak dla funkcji `plot()`:

```
plot.ecdf(dane1$obserwacje, col="black", lwd=2)
plot.ecdf(dane1$rcm_hist, col="red", lty=2, lwd=2, add=T)
```



Konstrukcja i interpretacja dystrybuanty

Każda wartość temperatury w analizowanej serii danych pojawia się z pewnym (określonym) prawdopodobieństwem i jest ona opisana na osi Y wykresu. Najczęściej są to percentyle (tj. kwantyle rzędu k/100, gdzie k=1, ... , 99) lub bardziej intuicyjnie - wartości procentowe, które odpowiadają prawdopodobieństwu wystąpienia danej temperatury poniżej zadanego odsetka w analizowanej serii danych.

Teoretycznie wartości kwantyli serii danych można obliczyć samodzielnie korzystając z funkcji `quantile()`.

Obliczmy wartości percentyli analizowanych serii danych historycznych i scenariuszowych z dokładnością do 1 percentyla.

```
quantile(x = dane1$obserwacje, probs = 0.01) # temperatura o prawdodpobieństwie wystąpienia 1%
```

```
##       1%
## -10.3143
```

```
quantile(x = dane1$rcm_hist, probs = 0.01) # temperatura o prawdodpobieństwie wystąpienia 1%
```

```
##        1%
## -10.35185
```

Do tego celu najlepiej stworzyć wektor składający się z 1000 elementów o wartościach od 0.01 do 1.00.

```
percentyle <- seq(from=0.01,to=1,by=0.01)
q_obs <- (quantile(dane1$obserwacje,percentyle))
q_rcm_hist <- (quantile(dane1$rcm_hist,percentyle))
q_rcm_scen <- (quantile(dane2$rcm_scen,percentyle))
```

Stworzone serie percentyli dla analizowanych serii danych należy złączyć po kolumnach i zapisać w postaci nowego obiektu.

```
tabelka <- data.frame(obs = q_obs, rcm_hist = q_rcm_hist, rcm_scen = q_rcm_scen)
```

Obliczenie różnic percentyli i wprowadzenie wartości poprawkowej

Polecenie:

1. W następnym kroku oblicz różnice pomiędzy serią danych obserwacyjnych i modelowych w analogicznym okresie (1971-2000) dla wyznaczonych wartości percentyli.
2. Wykreśl dystrybuantę dla obu serii danych w celu sprawdzenia poprawności wyniku
3. Uzyskane różnice uwzględnij w symulacji scenariuszowych zmian temperatury powietrza na wysokości 2m w latach 2021-2050.

```
tabelka$roznica <- tabelka$obs-tabelka$rcm_hist # Ad.1. różnica w odległości pomiędzy dystrybuantami
tabelka$rcm_hist_kor <- tabelka$roznica+tabelka$rcm_hist
tabelka$rcm_scen_kor <- tabelka$roznica+tabelka$rcm_scen
head(tabelka)
```

```
##         obs   rcm_hist  rcm_scen     roznica rcm_hist_kor rcm_scen_kor
## 1% -10.3143 -10.351848 -6.400917  0.03754795     -10.3143   -6.3633694
## 2%  -7.9286  -7.879132 -4.637112 -0.04946755      -7.9286   -4.6865794
## 3%  -6.2800  -6.426876 -3.519992  0.14687605      -6.2800   -3.3731160
## 4%  -5.2144  -5.519694 -2.397561  0.30529456      -5.2144   -2.0922667
## 5%  -4.2715  -4.590418 -1.786032  0.31891817      -4.2715   -1.4671139
## 6%  -3.5500  -3.893649 -1.317809  0.34364912      -3.5500   -0.9741594
```

Kontrolnie:

Ponowne wykreślenie analogicznych serii danych w formie dystrybuanty z porównaniem:

1. Serii danych obserwacyjnych (np. na czerwono)
2. Serii danych modelowych w okresie historycznym po uwzględnieniu wartości korekty. Jeśli dane skorygowane się pokrywają z serią danych obserwacyjnych oznacza to, że mapowanie kwantylowe zostało wykonane poprawnie
3. Możemy teraz dodać pogrubioną linią, w innym kolorze skorygowaną dystrybuantę dla scenariuszowych zmian klimatu

//Zadanie dla chętnych:

W jaki sposób można zrekonstruować serię danych modelowych reprezentującą “rzeczywiste”, skorygowane wartości temperatury (czyli np. -20*C 1. stycznia, itd.)