Modelowanie zdarzeń ekstremalnych w R

Analiza 3
Projekt egzaminacyjny

Julia Romatowska, 266542
Maria Małasiewicz, 260252

Wstęp

Głównym celem pracy jest wyestymowanie 20-letnich oraz 50-letnich poziomów zwrotu temperatur oraz wiatrów, dla sezonu wiosennego (marzec, kwiecień, maj). W tym celu wyznaczymy wspólne rozkłady dla temperatur oraz wiatrów w stacjach w Polsce oraz na podstawie wybranych kopuł wyznaczymy poziomy zwrotów i przeprowadzimy analizę otrzymanych wyników dla wszystkich stacji na terenie Polski, dla których dane z wybranego sezonu nie zawierały wielu braków.

W oryginalnej bazie danych mamy 182 stacje meteorologiczne, które zawierają informacje na temat temperatur oraz 165 stacji, które zawierają informacje o prędkościach wiatrów w miesiącach wiosennych lat 2008-2018. Bazy zawierały jednak wartości odstające, więc te temperatury, które nie należały do przedziału $[-45°C, 45°C]$ oraz te wiatry które nie należały do przedziału $[0\frac{m}{s}, 50\frac{m}{s}]$ zostały zamienione na *NA*. Następnie, te stacje które zawierały więcej niż 1200 braków danych zostały usunięte. Wspólnych stacji, które zawierały kompletne - według naszych założeń - dane zarówno o temperaturach, jak i wiatrach zostało 69, które widzimy na rysunku 1. Część nazw na mapie częściowo się pokrywała, dlatego niektóre ze stacji nie zostały podpisane.

Rysunek 1: Mapa Polski przedstawiające stacje pozostawione do analizy.

Do dalszej analizy wykorzystane zostały maksima dzienne. Rysunki 2 oraz 3 przedstawiają fragmenty baz danych wykorzystanych do dalszej analizy.

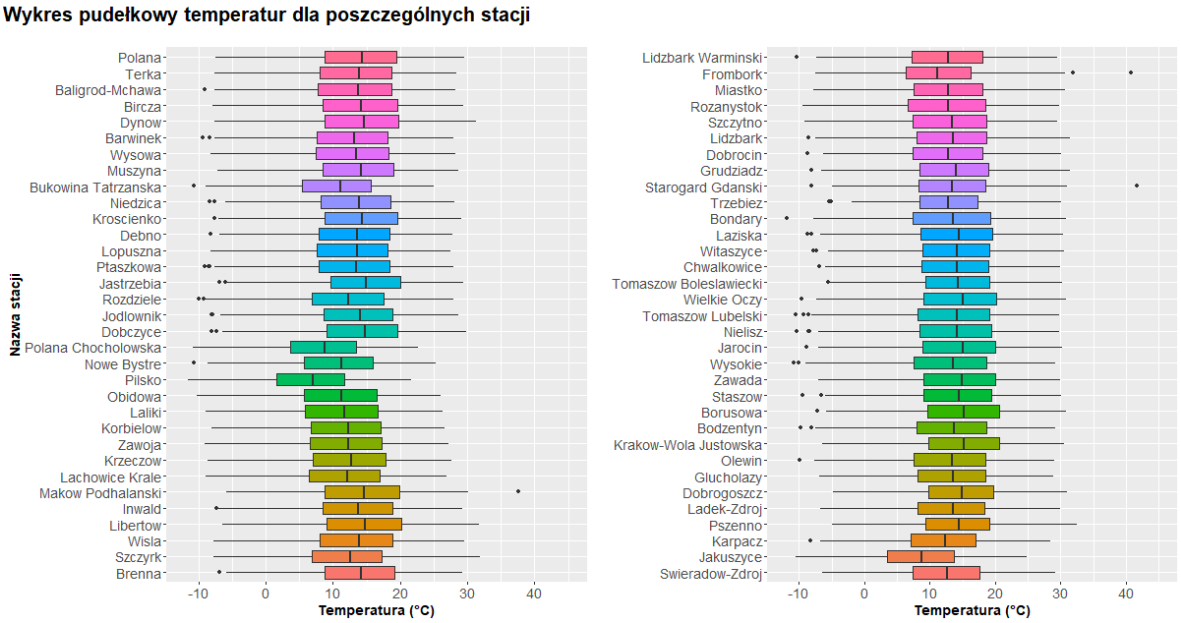
year	mth	day	X249180160	X249180210	X249180230
2008	3	1	11.21	6.73	8.80
2008	3	2	7.58	5.23	6.97
2008	3	3	9.51	8.02	9.08
2008	3	4	4.87	2.34	4.13
2008	3	5	1.98	-0.61	0.87
2008	3	6	6.44	4.89	6.19
2008	3	7	10.79	9.48	11.66
2008	3	8	8.30	7.42	9.21
2008	3	9	11.72	10.20	10.09
2008	3	10	12.01	10.82	10.72

Rysunek 2: Baza maksymalnych dziennych temperatur.

year	mth	day	X249180160	X249180210	X249180230
2008	3	1	15.3	14.2	16.8
2008	3	2	17.6	12.3	16.0
2008	3	3	10.4	9.2	8.8
2008	3	4	9.8	6.1	9.7
2008	3	5	8.0	5.3	7.1
2008	3	6	7.0	9.1	6.2
2008	3	7	9.3	7.5	5.7
2008	3	8	8.1	8.7	8.9
2008	3	9	6.9	6.5	6.1
2008	3	10	14.3	16.4	10.0

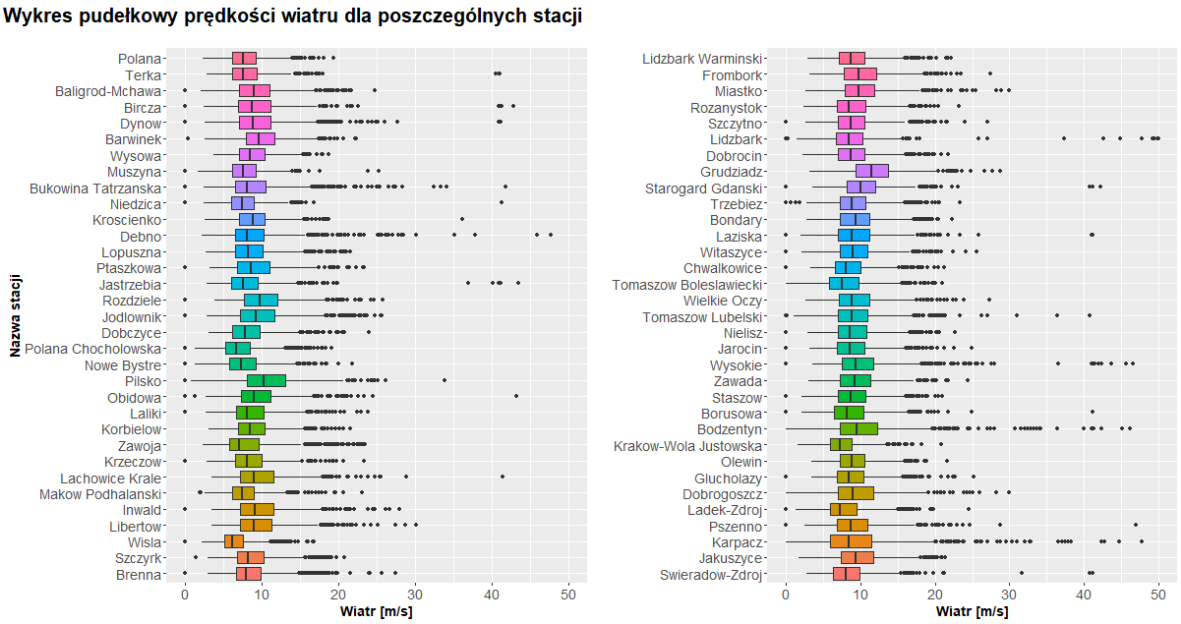
Rysunek 3: Baza maksymalnych dziennych prędkości wiatrów.

Na rysunku 4 widzimy wykresy pudełkowe maksymalnych dziennych temperatur dla każdej z analizowanych stacji. Dla niektórych stacji obserwujemy wartości odstające, jednakże ciężko jest stwierdzić czy takie obserwacje mogły czy nie mogły wystąpić. Nie odrzucamy tych wartości, jednakże warto jest je mieć na uwadze podczas formułowania wniosków.



Rysunek 4: Wykres pudełkowy temperatur z sezonu wiosennego dla poszczególnych analizowanych stacji.

Rysunek 5 przedstawia wykresy pudełkowe maksymalnych dziennych prędkości wiatrów odnotowanych w poszczególnych stacjach. Widzimy, że w porównaniu do wykresu temperatur 4, dla wiatrów istnieje więcej wartości, które w teorii uznawane są za odstające. Intuicyjne jednak wydaje się, że bardziej prawdopodobny jest duży skok prędkości wiatru niż skok temperatury. To spostrzeżenie może tłumaczyć ilość odstających obserwacji.



Rysunek 5: Wykres pudełkowy prędkości wiatru z sezonu wiosennego dla poszczególnych analizowanych stacji.

1 Analiza zależności między maksymalnymi dobowymi wartościami temperatur i wiatrów z wykorzystaniem kopuł

Twierdzenie Sklara mówi:

Twierdzenie 1.1
Dla dowolnej dystrybuanty d-wymiarowej F, z rozkładami brzegowymi $F_1, \ldots, F_d$ funkcja zdefiniowana

$$C(u_1, \ldots, u_d) = F(F_1^{-1}(u_1), \ldots, F_d^{-1}(u_d)),$$

gdzie $u_1, \ldots, u_d \in [0, 1]$, jest kopułą rozkładu F.

Kopuła C jest dystrybuantą więc możemy wykorzystać klasyczne metody estymacji. Nie posiadamy jednak próbki z rozkładu C a realizację $(x_1, y_1), \ldots, (x_n, y_n)$ z rozkładu $F(x, y) = C(F_1(x), F_2(y))$. Często jest tak, że nie znamy dystrybuant brzegowych F_1, F_2. W takich sytuacjach stosujemy podejście, w którym generujemy pseudo-obserwacje $(u_1, v_1), \ldots, (u_n, v_n)$, które możemy wyznaczyć za pomocą dwóch metod:

- *metoda parametryczna* - w której dopasowujemy najlepsze rozkłady F_1 i F_2 do danych brzegowych (np. według kryterium Akaike), następnie estymujemy parametry oraz tworzymy pseudo-obserwacje generując pary z wyznaczonych rozkładów $(u_i, v_i) = (F_1(x_i), F_2(y_i))$.
- *metoda nieparametryczna* - zakładamy, że

$$u_i = \frac{r_i}{n+1}, \; v_i = \frac{s_i}{n+1}, \tag{1}$$

  gdzie r_i oznacza rangę x_i wśród wartości $x_1, \ldots, x_n$, a s_i rangę y_i wśród wartości $y_1, \ldots, y_n$. Jest to inaczej złożenie z dystrybuantą empiryczną, w definicji której mamy dzielenie przez $n + 1$, zamiast n.

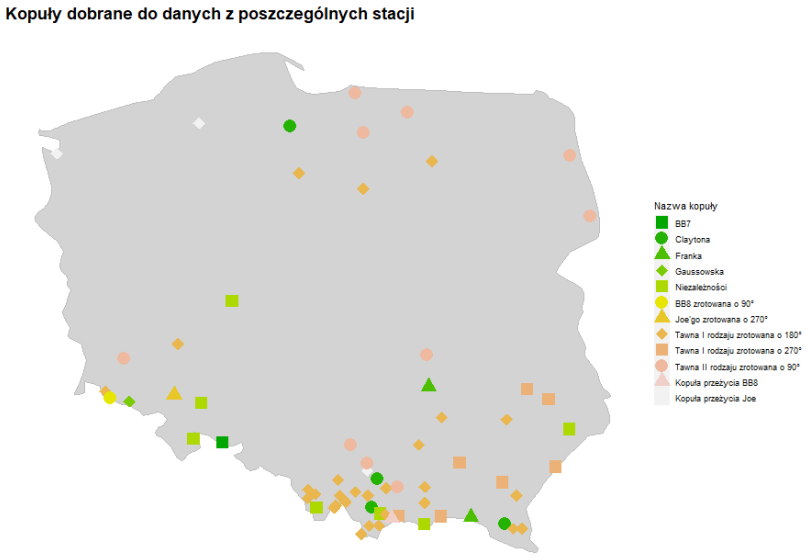
Mając kopułę i dopasowane rozkłady brzegowe lub w przypadku nieparametrycznym, dystrybuanty empiryczne, możemy generować duże próbki z rozkładu $F = C(F_1, F_2)$, wektora (X, Y).

Naszym zadaniem jest aby dla każdej ze stacji $s_1, s_2, \ldots, s_{59}$, wyznaczyć kopułę $C_{i,\ i=1,2,\ldots,59}$, która najlepiej opisuje zależności między maksymalnymi dobowymi temperaturami i wiatrami w tych stacjach. Do generowania pseudo-obserwacji wykorzystujemy podejście nieparametryczne, gdyż jest ono mniej wymagające czasowo. W tabeli 1 widzimy częstości występowania kopuł dopasowanych do naszych danych.

Tabela 1: Częstość występowania kopuł.

Rodzaj kopuły	Częstość występowania
Tawna I rodzaju zrotowana o 180°	28
Tawna II rodzaju zrotowana o 90°	10
Tawna I rodzaju zrotowana o 270°	7
Niezależności	7
Claytona	4
Joe'go zrotowana o 180°	3
Franka	2
Joe zrotowana o 270°	1
Gaussowska	1
BB7	1
BB8 zrotowana o 180°	1
BB8 zrotowana o 90°	1

Dodatkowo na rysunku 6 widzimy te same dane, które zostały zebrane w tabeli 1, ale zaznaczone na mapie Polski.



Rysunek 6: Mapa Polski przedstawiająca kopuły dopasowane dla poszczególnych stacji.

Widać, że najczęściej występującą kopułą jest kopuła ekstremalna - Tawna I rodzaju, zrotowana o 180°. Pochodzi ona z rodziny kopuł Tawna, które są asymetrycznym rozwinięciem kopuły Gumbela.

Jak możemy znaleźć w pracy Z. Gródek-Szostak i G. Malik [1], prezentowane są one za pomocą wzoru:

$$C_\theta(u_1, u_2; \theta) = u_1 u_2 \exp(\theta \frac{(\ln u_1)(\ln u_2)}{\ln u_1 + \ln u_2}), \tag{2}$$

gdzie $0 \leqslant \theta \leqslant 1$, a ich funkcje niezależności są dane wzorem

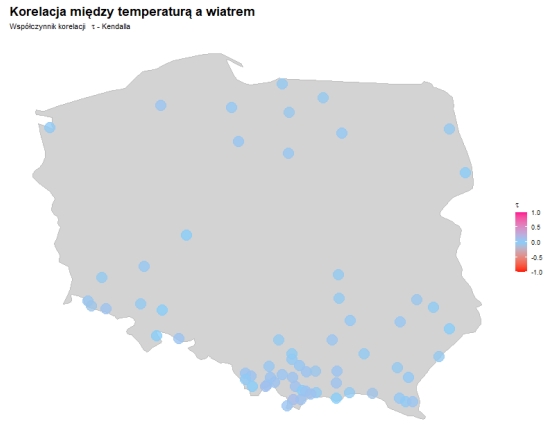
$$A(t) = \theta t^2 - \theta t + 1. \tag{3}$$

Najczęściej występująca w tej analizie kopuła jest kopułą Tawna I rodzaju, która posiada dwa parametry $u_1 \in (1, \infty)$, $u_2 \in (0, 1)$. Jest także kopułą zrotowaną o 180°, które zwane są również kopułami przeżycia. Opisane są one wzorem:

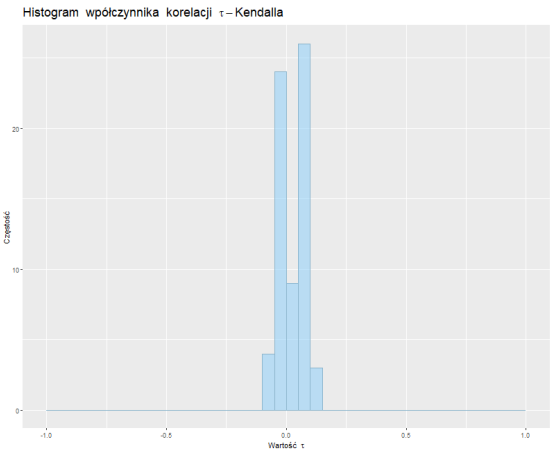
$$\bar{C}(u_1, u_2) = u_1 + u_2 - 1 + C(1 - u_1, 1 - u_2). \tag{4}$$

Oznacza to, że jeśli kopuła C jest kopułą wektora (u_1, u_2) to kopuła przeżycia $\bar{C}$ jest kopułą wektora $(1 - u_1, 1 - u_2)$.

Rysunek 7 przedstawia współczynniki korelacji τ, które określają zależności między temperaturą a wiatrem na terenie Polski. Widzimy, że te zależności są bardzo małe. Z histogramu 8 odczytujemy, że wartości współczynnika korelacji Kendalla zawierają się w przedziale $[-0.15, 0.15]$. Największa występująca zależność wynosi $\tau = 0.14$ i jest to wartość opisująca słabą korelację.



Rysunek 7: Korelacje temperatur z wiatrem na przestrzeni Polski.



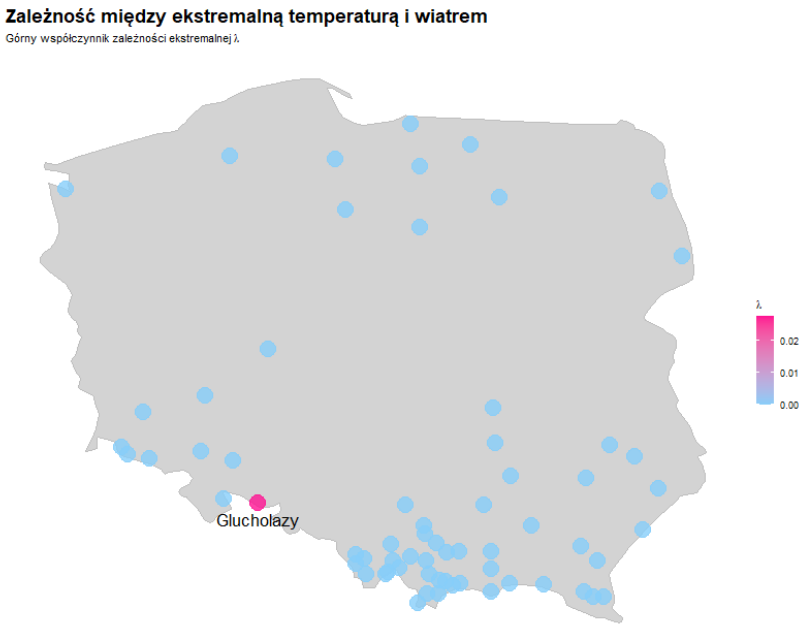
Rysunek 8: Histogram współczynników korelacji τ.

Dane współczynniki korelacji Kendalla nie mówią nam jednak nic o zależnościach między zdarzeniami ekstremalnymi. Aby sprawdzić jaki wpływ mają na siebie temperatura i wiatr, w takich sytuacjach, wyznaczamy górny współczynnik zależności ekstremalnej λ (gdyż badamy zależności w wysokich kwantylach rozkładów). Współczynnik jest definiowany wzorem

$$\begin{aligned} \lambda_u &= \lim_{u \to 1^-} P(X_2 > F_2^{-1}(u) | X_1 > F_1^{-1}(u)) \\ &= \lim_{u \to 1^-} P(U_2 > u | U_1 > u) \\ &= \lim_{u \to 1^-} \frac{1 - 2u + C(u, u)}{1 - u}, \end{aligned} \tag{5}$$

gdzie $U_1 = F(X_1)$ i $U_2 = F(X_2)$. W zależności od tego, czy współczynnik jest dodatni czy równy zero, mówimy o zależności lub niezależności w górnym ogonie rozkładu. Wartość dodatnia wskazuje na tendencję kopuły do generowania wspólnych ekstremalnych zdarzeń.

Z rysunku 9 odczytujemy, że dla większości stacji w Polsce, zależności między ekstremalnymi zdarzeniami nie występują. Jest jednak jedna stacja, dla której wartość $\lambda \approx 0.028 > 0$ - jest to stacja w Głuchołazach. Dla tej stacji występuje zależność między ekstremalnymi temperaturami i wiatrami.



Rysunek 9: Mapa Polski ilustrująca górny współczynnik zależności ekstremalnej λ.

2 Poziomy zwrotu

Aby wyznaczyć poziomy zwrotu dla każdej ze stacji korzystamy z poniżej opisanej metody. Z łącznego rozkładu maksymalnych dobowych temperatur i wiatrów

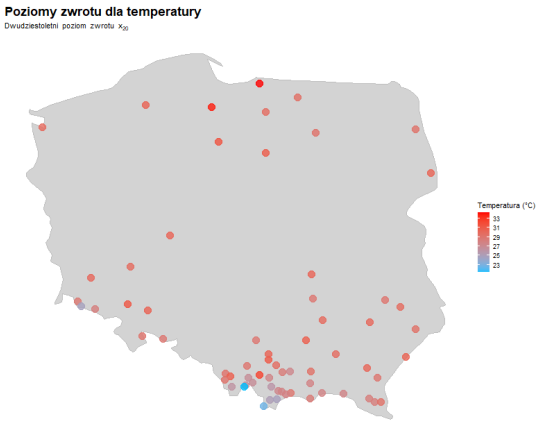
$$C_i(\hat{F}_1, \hat{F}_2) \tag{6}$$

w stacji s_i (gdzie C_i jest wcześniej dopasowaną kopułą), generujemy metodą nieparametryczną, próbę liczności n, gdzie n to liczba dni w badanym okresie. Otrzymujemy w ten sposób realizację n możliwych maksymalnych dobowych temperatur i maksymalnych prędkości wiatrów w stacji s_i. Przyjmujemy, że $n = 92 \cdot 11$, gdyż zakładamy, że w badanych okresach mamy 92 dni, na przestrzeni 11 lat. Następnie grupujemy je w 11 bloków, oddzielnie dla temperatur i dla wiatrów oraz wyznaczamy maksimum z danego bloku. Ten proces powtarzamy 100 razy i w ten sposób otrzymujemy 1100 maksimów rocznych dla stacji $s_1, s_2, \ldots, s_d$. Aby, na podstawie wygenerowanych danych, wyznaczyć poziomy zwrotu korzystamy z metody maksimów blokowych.

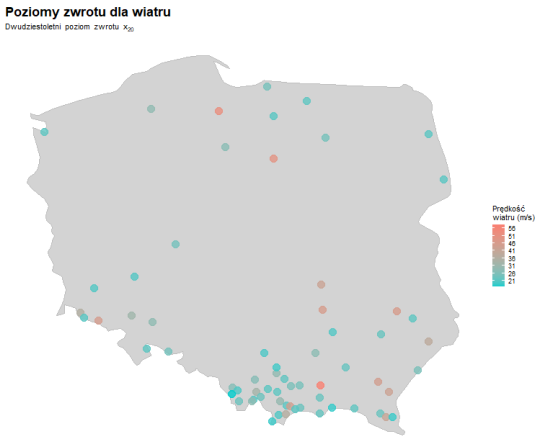
Rysunki 10 i 11 przedstawiają dwudziestoletnie poziomy zwrotu, odpowiednio dla temperatur i wiatrów. Otrzymane podczas analizy wartości temperatur sięgają od około $21°C$ do ponad $34°C$, z czego najniższe wartości znajdują się głównie na południu kraju. Natomiast prędkości wiatrów, których możemy się spodziewać średnio raz na dwadzieścia lat wynoszą w zależności od stacji od $\approx 16\frac{m}{s}$ do $\approx 60\frac{m}{s}$.

Na wykresach 12 i 13 możemy przyjrzeć się pięćdziesięcioletnim poziomom zwrotu dla polskich stacji. Na pierwszym z nich możemy zauważyć, że przedział wartości spodziewanych temperatur nieznacznie się zwiększył i oczekiwane średnio raz na pięćdziesiąt lat temperatury wynoszą między $21°C$ do prawie $36°C$. Dla wiatrów różnica między dwudziestoletnimi i pięćdziesięcioletnimi poziomami zwrotu jest bardziej zauważalna. Oczekiwane raz na pięćdziesiąt lat prędkości wiatru osiągają między $16\frac{m}{s}$ do prawie $96\frac{m}{s}$.

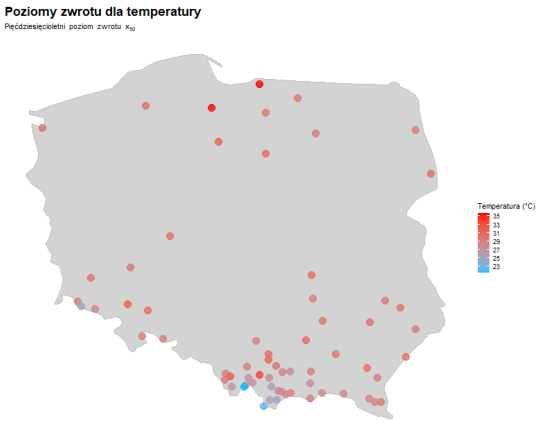
Nie zauważamy ewidentnych schematów zachowań przewidywanych temperatur ani wiatrów w zależności od położenia geograficznego. Pewną zależność zauważamy dla temperatur - niższe temperatury przewidujemy dla południa oraz południowego zachodu Polski, natomiast zależność dla wiatrów trudno jest sprecyzować.



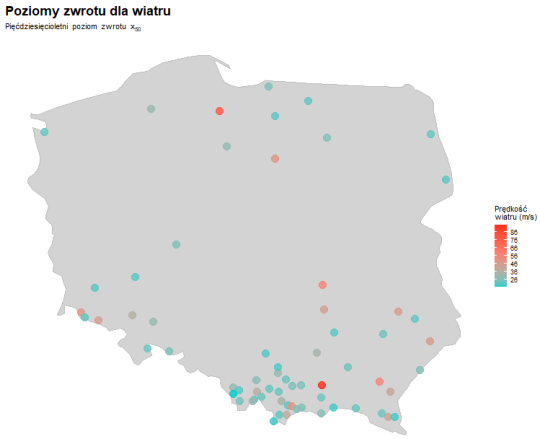
Rysunek 10: Dwudziestoletni poziom zwrotu dla temperatur.



Rysunek 11: Dwudziestoletni poziom zwrotu dla wiatru.



Rysunek 12: Pięćdziesięcioletni poziom zwrotu dla temperatur.



Rysunek 13: Pięćdziesięcioletni poziom zwrotu dla wiatru.

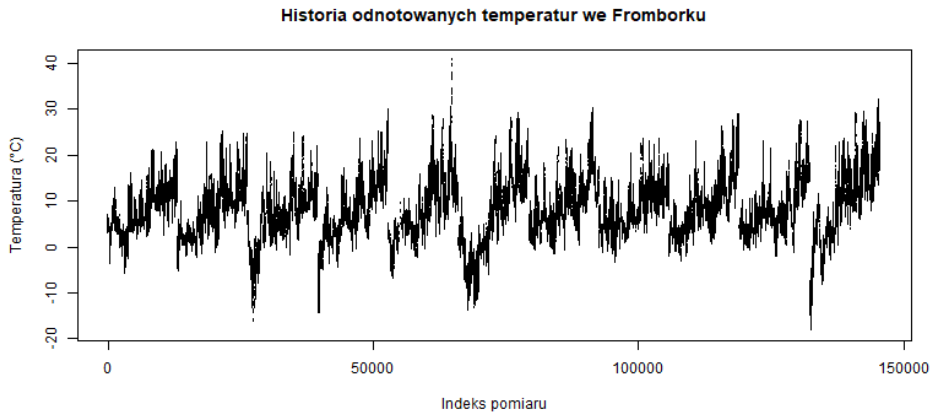
Tabela 2 przedstawia maksymalne i minimalne wartości dwudziestoletnich i pięćdziesięcioletnich poziomów zwrotu zarówno dla temperatur, jak i dla wiatrów.

Tabela 2: Maksymalne i minimalne wartości poziomów zwrotu.

Temperatura			
	Wartość x_{20}	Wartość x_{50}	Odpowiadająca stacja
Maksimum	$34.19°C$	$35.71°C$	Frombork
Minimum	$21.38°C$	$21.49°C$	Pilsko
Wiatr			
	Wartość x_{20}	Wartość x_{50}	Odpowiadająca stacja
Maksimum	$58.32\frac{m}{s}$	$94.67\frac{m}{s}$	Jastrzębia
Minimum	$16.76\frac{m}{s}$	$16.79\frac{m}{s}$	Wisła

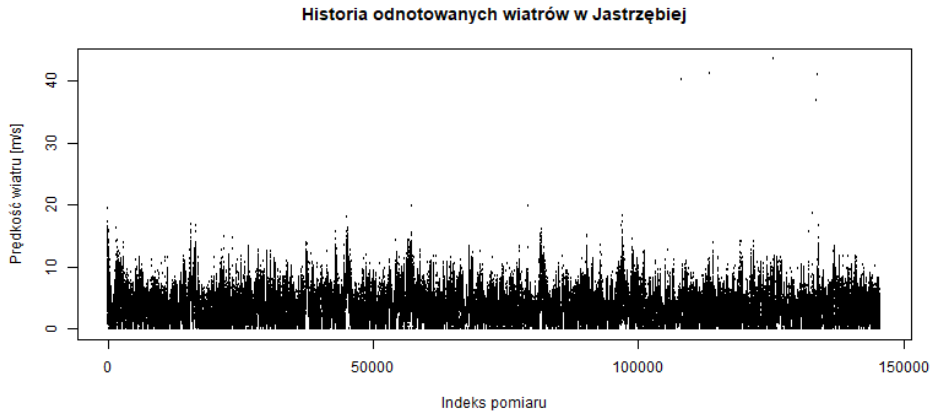
Z tabeli odczytujemy, że najwyższe temperatury odnotowujemy dla Fromborka, natomiast najniższe

dla Pilska. Odnosząc się do rysunku 4, który przedstawia wykres pudełkowy temperatur, możemy wysunąć wniosek, że poziomy zwrotu dla stacji we Fromborku są tak wysokie poprzez odstające obserwacje rzędu $30°C$ i $40°C$. Można by się zastanowić czy nie warto usunąć tych obserwacji do dalszej analizy. Na wykresie 14 widzimy, że nie nastąpił jednak nieoczekiwany skok temperatury, a wręcz przeciwnie, temperatura danego dnia wzrastała aż osiągnęła około 40 stopni. Odrzucenie tych pomiarów temperatur nie byłoby w tym przypadku uzasadnione. Przyglądając się wykresowi pudełkowemu dla Pilska widzimy, że dla tej stacji średnia temperatura z całego analizowanego okresu jest najniższa, więc nie dziwią nas niskie wartości poziomów zwrotu.



Rysunek 14: Historia odnotowanych temperatur we Fromborku na przestrzeni 11 lat.

Odnosząc się do prędkości wiatru, najsilniejsze wiatry zarówno średnio raz na 20 jak i 50 lat będziemy odnotowywać w Jastrzębiej, a najsłabsze w Wiśle. Przewidywana prędkość wiatru w Jastrzębiej klasyfikuje się jako huragan dewastujący i niszczycielski ($\geqslant 70\frac{m}{s}$), który powoduje niewyobrażalne szkody: zrywa dachy i niszczy budynki o wzmocnionej konstrukcji, przewraca pociągi i samochody ciężarowe, porywa i przenosi samochody osobowe, wyrywa i łamie drzewa na całych połaciach lasów oraz powoduje lewitowanie ciężkich przedmiotów. Wymagana jest wówczas ewakuacja ludności. Gdy spojrzymy na wykres 15 widzimy, że maksymalne wiatry odnotowane przez tę stację w przeciągu 11 lat wahają się wokół $40\frac{m}{s}$. Dana stacja odnotowała takie silne wiatry 5 razy w ciągu okresu obserwacji. Dodatkowo widać, że w tych przypadkach nie widzimy ciągłego wzrostu prędkości, a odnotowujemy skok od około $20\frac{m}{s}$ do $40\frac{m}{s}$ i powrót do $20\frac{m}{s}$. W tej sytuacji warto by było zapytać specjalisty, aby upewnić się czy takie skoki prędkości wiatrów są możliwe. Jeśli nie - powinniśmy usunąć te pomiary.



Rysunek 15: Historia odnotowanych prędkości wiatru w Jastrzębiej na przestrzeni 11 lat.

3 Porównanie z wcześniejszymi wnioskami

Stacje, dla których otrzymaliśmy najwyższe i najniższe poziomy zwrotu, są te same jak we wcześniejszej analizie z wykorzystaniem metody maksimów blokowych. Porównanie wartości możemy zobaczyć w tabeli 3.

Tabela 3: Maksymalne i minimalne wartości poziomów zwrotu.

x_{20}			
	Wyznaczony w analizie nr 3	Wyznaczony w analizie nr 2	Odpowiadająca stacja
Maksimum	$34.19°C$	$36.77°C$	Frombork
Minimum	$21.38°C$	$21.51°C$	Pilsko
x_{50}			
	Wyznaczony w analizie nr 3	Wyznaczony w analizie nr 2	Odpowiadająca stacja
Maksimum	$35.71°C$	$38.14°C$	Frombork
Minimum	$21.49°C$	$21.62°C$	Pilsko

Wartości wyznaczonych poziomów zwrotu są niższe gdy korzystamy z kopuł. Różnica dla temperatur we Fromborku, dla którego otrzymaliśmy największe ich wartości, wynosi około $2.5°C$ zarówno dla dwudziesto- i pięćdziesięcioletnich poziomów zwrotu. Natomiast różnica pomiędzy najniższymi poziomami zwrotu dla obu analiz, które zostały wyznaczone dla stacji w Pilsku, nie przekracza $0.15°C$.

Dodatkowo możemy stwierdzić, że poziomy zwrotu dla temperatur z wcześniej analizowanych stacji rozkładają się podobnie na terenie Polski. Najniższe temperatury przewidywaliśmy dla południa oraz południowego zachodu, analogicznie jak przy analizie korzystającej z kopuł.

Tabela 4: Wyznaczone wartości zwrotu w zależności od metody.

Stacja	Poziom zwrotu	Metoda 1	Metoda 2	Metoda 3
Olewin	x_{20}	32.10	28.96	28.79
	x_{50}	36.75	29.20	28.92
Pszenno	x_{20}	31.38	31.71	31.23
	x_{50}	32.50	32.29	31.86

W tabeli 4 widzimy przedstawione wartości poziomów zwrotu wyznaczone metodą maksimów blokowych. W pierwszej metodzie dopasowujemy rozkład GEV korzystając z maksimów rocznych (liczba statystyk $r = 1$). W drugiej metodzie korzystamy z dziesięciu maksimów rocznych ($r = 10$) i do tych danych dopasowujemy rozkład GEV. Trzecia metoda polega na dopasowaniu kopuły oraz wygenerowaniu danych z wyznaczonego rozkładu łącznego by następnie wyznaczyć poziomy zwrotu.

Literatura

[1] Z. Gródek-Szostak, G. Malik. *Modeling the Dependency between Extreme Prices of Selected Agricultural Products on the Derivatives Market Using the Linkage Function.* 2019, strona: 5