Eksploracja Masywnych Danych - Analiza danych

Kajetan Zimniak & Bartosz Górka

02 November, 2019

Contents

Podsumowanie analizy

Przedmiotem analizy było określenie przyczyn zmniejszenia się długości śledzi. Do dyspozycji mieliśmy ponad 52 tysięcy obserwacji dokonanych podczas połowów. Każda obserwacja zawierała dane o stanie środowiska, dostępności pokarmu i eksploatacji łowisk.

Przedstawiono podstawowe charakterystyki zbioru danych dotyczących poszczególnych atrybutów i wyeliminowano obserwacje odstające. Następnie uzupełniono brakujące dane korzystając z filtru Kalmana. Dysponując tak przygotowanym zbiorem danych, sprawdzono korelację pomiędzy cechami. Zauważono znaczny wpływ temperatury przy powierzchni wody na długość śledzi.

Przeprowadzając dalszą analizę, przedstawiono zmiany w długości śledzi, środowisku naturalnym oraz dostępności pokarmu. Określono moment w czasie, od którego długość ryb uległa zmniejszeniu.

Ostatnim krokiem było przygotowanie trzech modeli regresji próbujących przewidzieć długość śledzi. Modele zostały porównane ze sobą oraz przedstawiono ich rozkład ważności atrybutów. Potwierdziło to hipotezę, o zależności długości ryb od temperatury przy powierzchni wody.

Wykorzystane biblioteki

- `knitr`
- `kableExtra`
- `dplyr`
- `tidyverse`
- `ggplot2`
- `gridExtra`
- `imputeTS`
- `corrplot`
- `reshape2`
- `caret`
- `gganimate`
- `gifski`

Ustawienie ziarna generatora

Celem zapewnienia powtarzalności operacji losowania, a co za tym idzie powtarzalności wyników przy każdym uruchomieniu raportu na tych samych danych, zastosowano ziarno generatora o wartości `102019`.

```
set.seed(102019)
```

Charakterystyka obserwacji - zastosowane atrybuty

W ramach analizy mamy do czynienia z obserwacjami opisanymi za pomocą następujących atrybutów:

- **length**: długość złowionego śledzia [cm]
- **cfin1**: dostępność planktonu [zagęszczenie *Calanus finmarchicus* gat. 1]
- **cfin2**: dostępność planktonu [zagęszczenie *Calanus finmarchicus* gat. 2];
- **chel1**: dostępność planktonu [zagęszczenie *Calanus helgolandicus* gat. 1];
- **chel2**: dostępność planktonu [zagęszczenie *Calanus helgolandicus* gat. 2];
- **lcop1**: dostępność planktonu [zagęszczenie *widłonogów* gat. 1];
- **lcop2**: dostępność planktonu [zagęszczenie *widłonogów* gat. 2];
- **fbar**: natężenie połowów w regionie [ułamek pozostawionego narybku];
- **recr**: roczny narybek [liczba śledzi];
- **cumf**: łączne roczne natężenie połowów w regionie [ułamek pozostawionego narybku];
- **totaln**: łączna liczba ryb złowionych w ramach połowu [liczba śledzi];
- **sst**: temperatura przy powierzchni wody [°C];
- **sal**: poziom zasolenia wody [Knudsen ppt];
- **xmonth**: miesiąc połowu [numer miesiąca];
- **nao**: oscylacja północnoatlantycka [mb].

Table 1: Wybrane pomiary

length	cfin1	cfin2	chel1	chel2	lcop1	lcop2	fbar	recr	cumf	totaln	sst	sal	xmonth	nao
23.0	0.02778	0.27785	2.46875	NA	2.54787	26.35881	0.356	482831	0.3059879	267380.8	14.30693	35.51234	7	2.8
22.5	0.02778	0.27785	2.46875	21.43548	2.54787	26.35881	0.356	482831	0.3059879	267380.8	14.30693	35.51234	7	2.8
25.0	0.02778	0.27785	2.46875	21.43548	2.54787	26.35881	0.356	482831	0.3059879	267380.8	14.30693	35.51234	7	2.8
25.5	0.02778	0.27785	2.46875	21.43548	2.54787	26.35881	0.356	482831	0.3059879	267380.8	14.30693	35.51234	7	2.8
24.0	0.02778	0.27785	2.46875	21.43548	2.54787	26.35881	0.356	482831	0.3059879	267380.8	14.30693	35.51234	7	2.8
22.0	0.02778	0.27785	2.46875	21.43548	2.54787	NA	0.356	482831	0.3059879	267380.8	14.30693	35.51234	7	2.8

Wczytanie danych z pliku

Dane zamieszczone na stronie przedmiotu w postaci pliku CSV pobieramy wyłącznie w sytuacji braku pliku w katalogu roboczym. Pozwala to nam na ograniczenie niepotrzebnego transferu danych, jeżeli plik już istnieje.

```
file_name = "sledzie.csv"
source_url = "http://www.cs.put.poznan.pl/alabijak/emd/projekt/sledzie.csv"

if (!file.exists(file_name)) {
  download.file(source_url, destfile = file_name, method = "wget")
}
```

Po zapewnieniu istnienia zbioru danych wczytujemy obserwacje.

```
content =
  file_name %>%
  read_csv(col_names = TRUE, na = c("", "NA", "?")) %>%
  select(-1)
```

Oryginalnie zbiór posiada znaki `?` jako oznaczenie wartości pustej (brakującej). Dzięki wykorzystaniu parametru `na` podczas wywołania funkcji `read_csv` możemy zastąpić znak `?` poprawnym oznaczeniem braku wartości `NA`.

```
content %>%
  head(n = 6) %>%
  kable(align = "c", caption = "Wybrane pomiary") %>%
  kable_styling(latex_options = "scale_down")
```

W tabeli `Wybrane pomiary` zaprezentowano pierwsze sześć obserwacji. Jak możemy zaobserwować, żadna nie ma wartości `?`, która została poprawnie oznaczona jako `NA`.

Podstawowe statystyki zbioru danych

W zbiorze danych mamy do czynienia z 52582 obserwacjami opisanych za pomocą 15 atrybutów. W całym zbiorze mamy do czynienia z 42488 obserwacjami bez ani jednej wartości pustej co stanowi 81 procent całego zbioru.

Statystyka parametrów obserwacji

Table 2: Statystyka zbioru danych

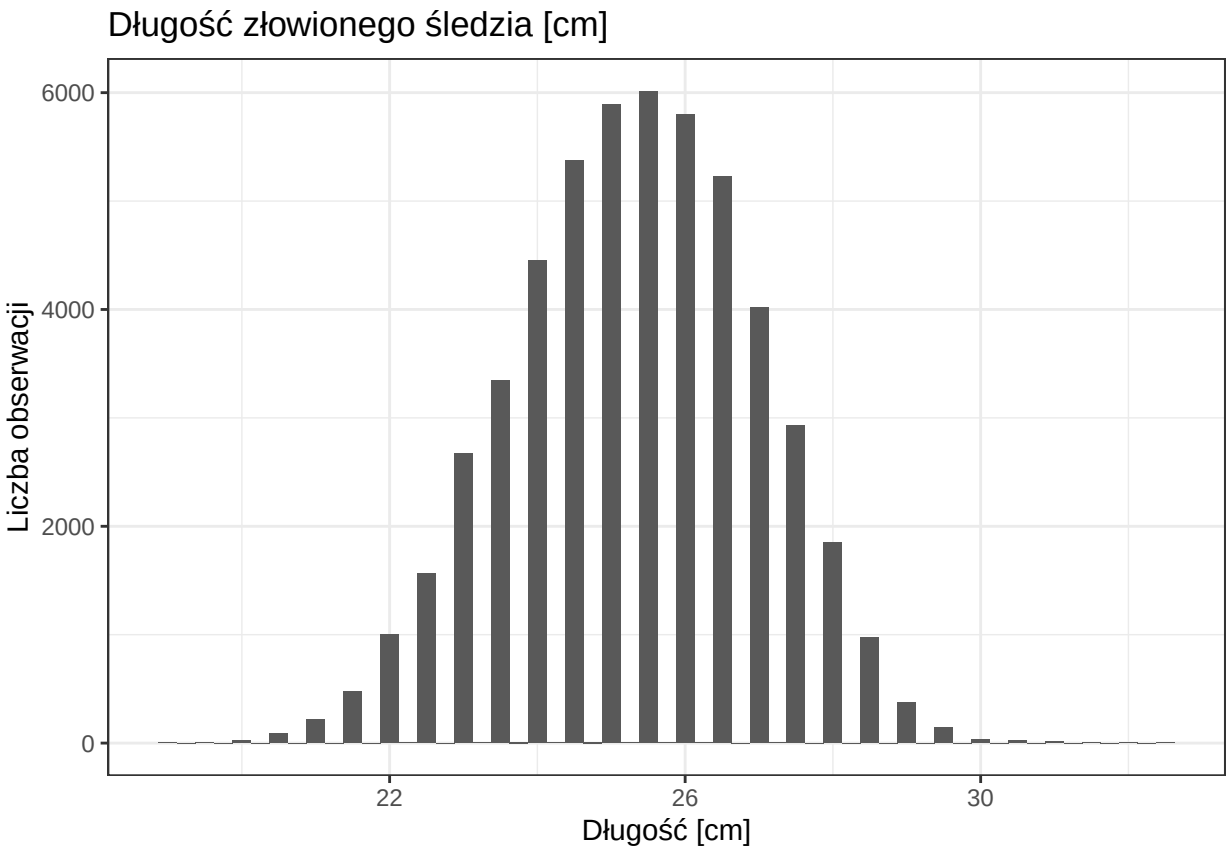
	length	cfin1	cfin2	chel1	chel2	lcop1	lcop2	fbar	recr	cumf	totaln	sst	sal	xmonth	nao
	Min. :19.0	Min. : 0.0000	Min. : 0.0000	Min. : 0.000	Min. : 5.238	Min. : 0.3074	Min. : 7.849	Min. :0.0680	Min. : 140515	Min. :0.06833	Min. : 144137	Min. :12.77	Min. :35.40	Min. : 1.000	Min. :-4.89000
	1st Qu.:24.0	1st Qu.: 0.0000	1st Qu.: 0.2778	1st Qu.: 2.469	1st Qu.:13.427	1st Qu.: 2.5479	1st Qu.:17.808	1st Qu.:0.2270	1st Qu.: 360061	1st Qu.:0.14809	1st Qu.: 306068	1st Qu.:13.60	1st Qu.:35.51	1st Qu.: 5.000	1st Qu.:-1.89000
	Median :25.5	Median : 0.1111	Median : 0.7012	Median : 5.750	Median :21.673	Median : 7.0000	Median :24.859	Median :0.3320	Median : 421391	Median :0.23191	Median : 539558	Median :13.86	Median :35.51	Median : 8.000	Median : 0.20000
	Mean :25.3	Mean : 0.4458	Mean : 2.0248	Mean :10.006	Mean :21.221	Mean : 12.8108	Mean :28.419	Mean :0.3304	Mean : 520366	Mean :0.22981	Mean : 514973	Mean :13.87	Mean :35.51	Mean : 7.258	Mean :-0.09236
	3rd Qu.:26.5	3rd Qu.: 0.3333	3rd Qu.: 1.7936	3rd Qu.:11.500	3rd Qu.:27.193	3rd Qu.: 21.2315	3rd Qu.:37.232	3rd Qu.:0.4560	3rd Qu.: 724151	3rd Qu.:0.29803	3rd Qu.: 730351	3rd Qu.:14.16	3rd Qu.:35.52	3rd Qu.: 9.000	3rd Qu.: 1.63000
	Max. :32.5	Max. :37.6667	Max. :19.3958	Max. :75.000	Max. :57.706	Max. :115.5833	Max. :68.736	Max. :0.8490	Max. :1565890	Max. :0.39801	Max. :1015595	Max. :14.73	Max. :35.61	Max. :12.000	Max. : 5.08000
	NA	NA's :1581	NA's :1536	NA's :1555	NA's :1556	NA's :1653	NA's :1591	NA	NA	NA	NA	NA's :1584	NA	NA	NA

```
content %>%
  summary() %>%
  kable(align = "c", caption = "Statystyka zbioru danych") %>%
  kable_styling(latex_options = "scale_down")
```

W tabeli `Statystyka zbioru danych` zaprezentowano wynik działania funkcji `summary`, która dokonała analizy rozkładu wartości każdego z atrybutów. W zbiorze danych mamy do czynienia z siedmioma atrybutami posiadającymi wartości puste. Analizę rozkładów wartości pozostawiamy czytelnikowi. W obecnej postaci nie jest ona jednakże przydatna w próbie rozwiązania problemu zmniejszenia się wielkości śledzi.

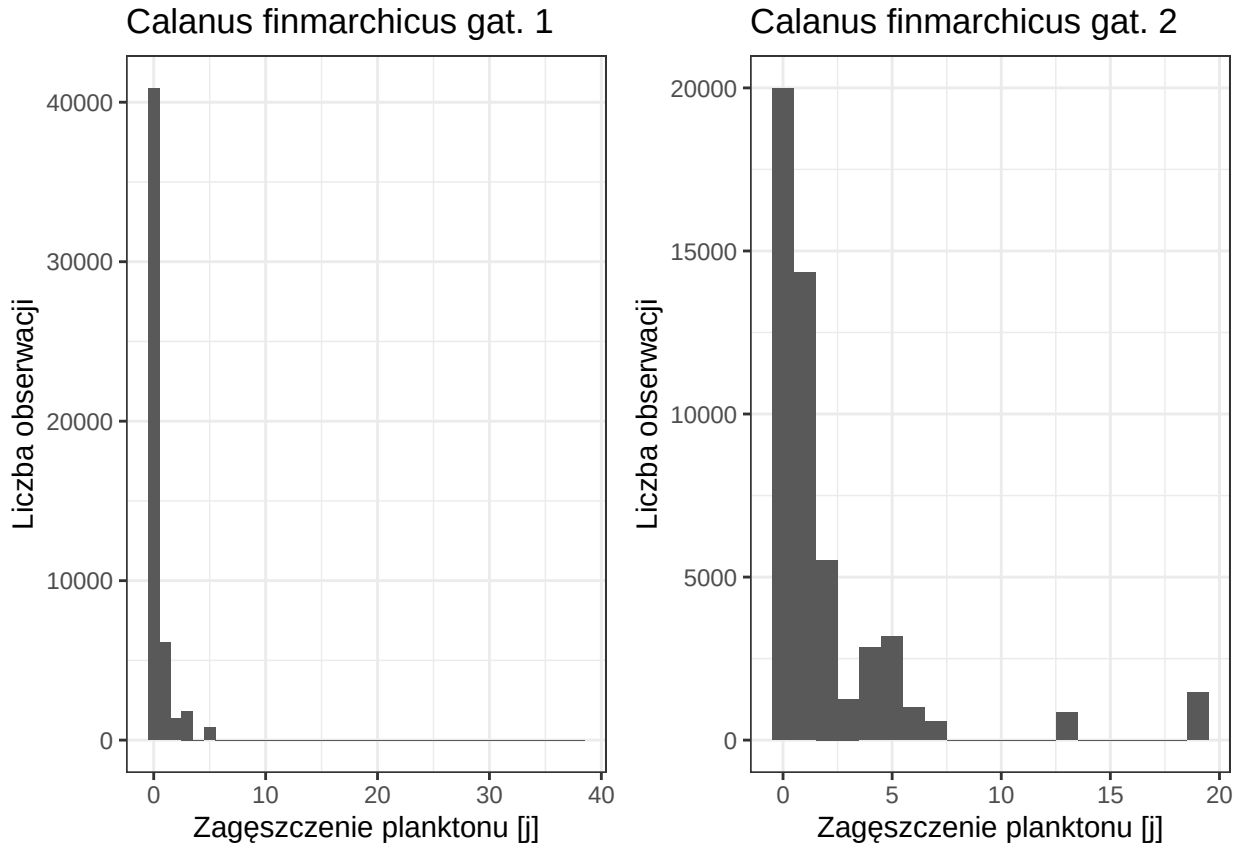
Rozkład wartości cech

```
ggplot(content, aes(x = length)) + geom_histogram(binwidth = 0.25) +
  theme_bw() + ggtitle("Długość złowionego śledzia [cm]") +
  xlab(sprintf("Długość [cm]")) + ylab("Liczba obserwacji")
```



Jak możemy zaobserwować, większość śledzi w połowie ma długość od 23 do 27 centymetrów. Mamy do czynienia z rozkładem bardzo zbliżonym do rozkładu normalnego.

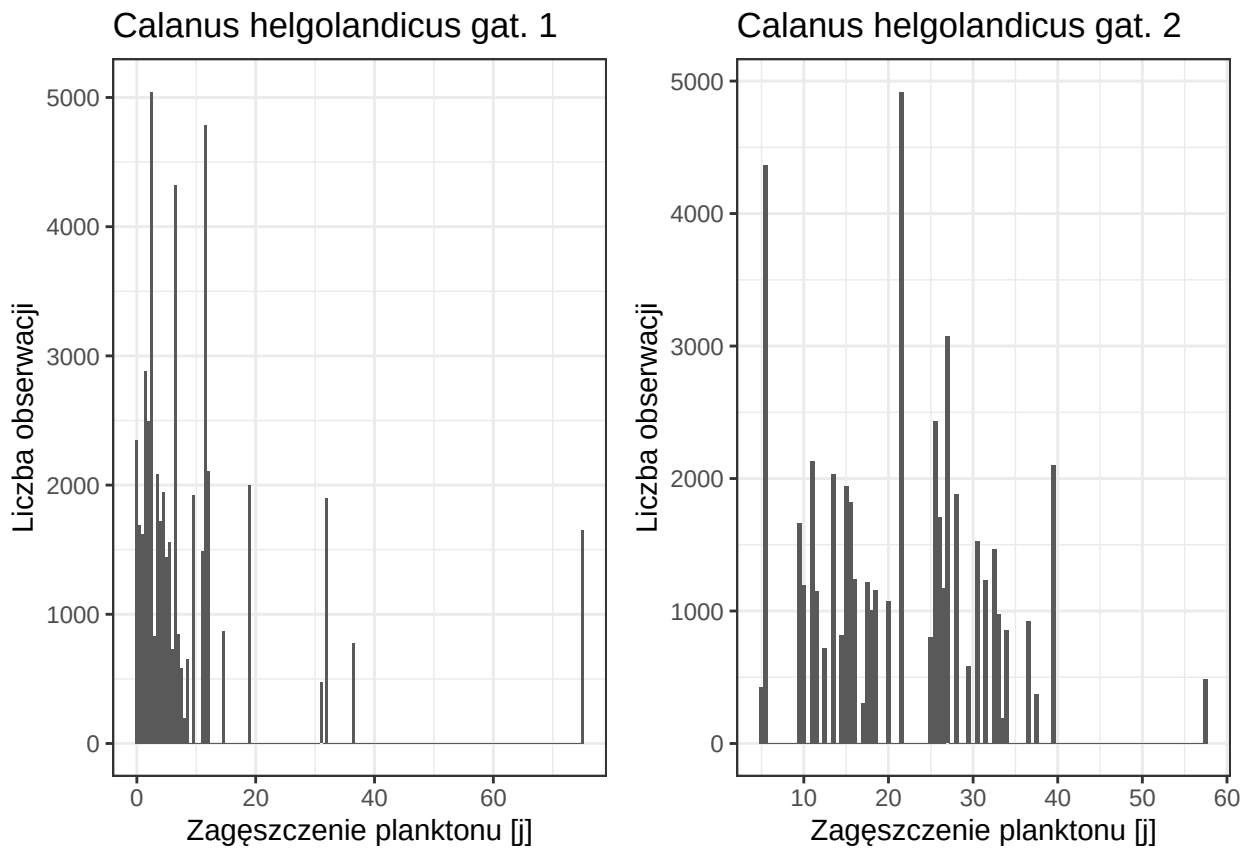
```
plot_cfin1 <- ggplot(content, aes(x = cfin1)) + geom_histogram(binwidth = 1.0) +
  theme_bw() + ggtitle("Calanus finmarchicus gat. 1") +
  xlab(sprintf("Zagęszczenie planktonu [j]")) + ylab("Liczba obserwacji")

plot_cfin2 <- ggplot(content, aes(x = cfin2)) + geom_histogram(binwidth = 1.0) +
  theme_bw() + ggtitle("Calanus finmarchicus gat. 2") +
  xlab(sprintf("Zagęszczenie planktonu [j]")) + ylab("Liczba obserwacji")

grid.arrange(plot_cfin1, plot_cfin2, nrow = 1)
```



Wykres zagęszczenia planktonu *Calanus finmarchicus* wskazuje nam, jak wiele obserwacji jest zbliżonych do siebie. Jest to widoczne szczególnie dla gatunku 1, który kwartyle pierwszy, drugi oraz trzeci osiąga w zakresie `[0; 0,5]`, podczas gdy jego wartość maksymalna wynosi aż 37,67. Wartości odstające powinny zostać wyeliminowane w dalszej analizie.

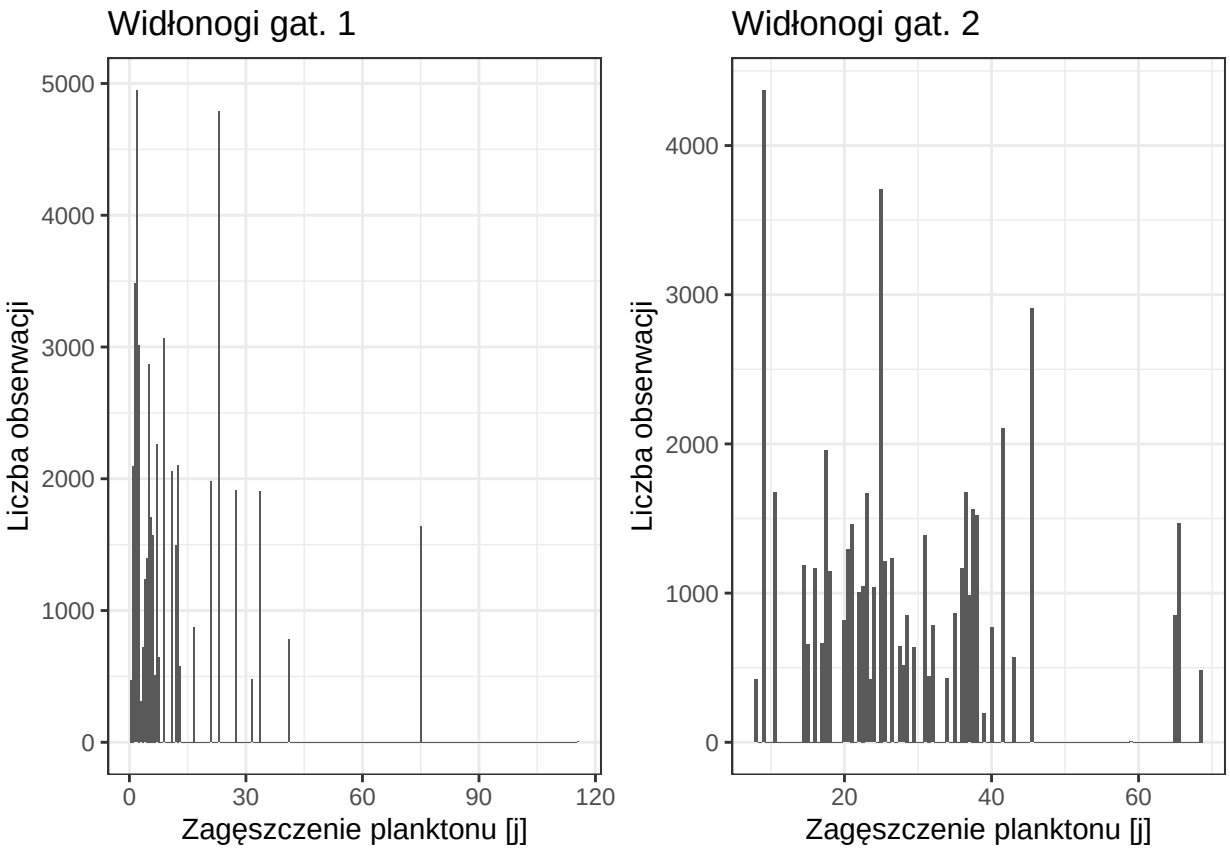
```
plot_chel1 <- ggplot(content, aes(x = chel1)) + geom_histogram(binwidth = 0.5) +
  theme_bw() + ggtitle("Calanus helgolandicus gat. 1") +
  xlab(sprintf("Zagęszczenie planktonu [j]")) + ylab("Liczba obserwacji")

plot_chel2 <- ggplot(content, aes(x = chel2)) + geom_histogram(binwidth = 0.5) +
  theme_bw() + ggtitle("Calanus helgolandicus gat. 2") +
  xlab(sprintf("Zagęszczenie planktonu [j]")) + ylab("Liczba obserwacji")
```

```
grid.arrange(plot_chel1, plot_chel2, nrow = 1)
```



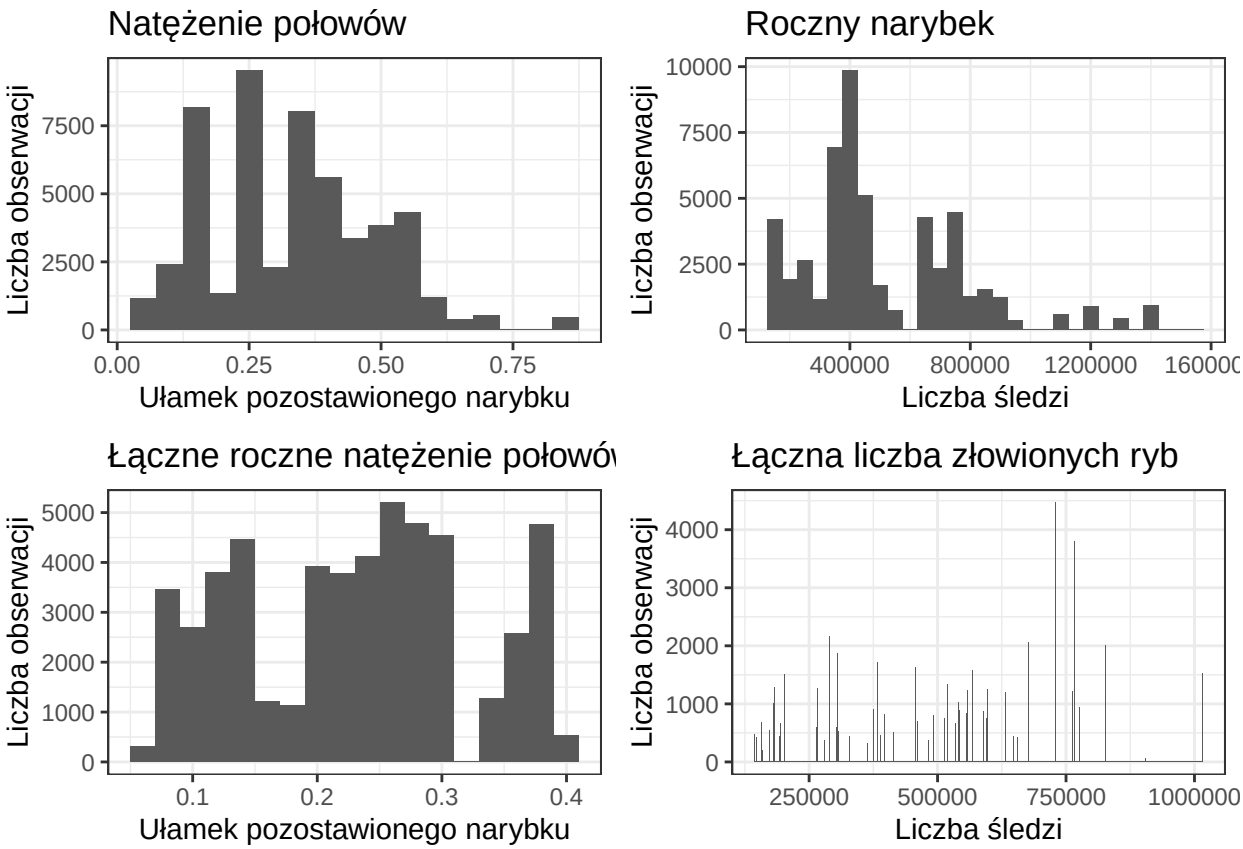
W przypadku zagęszczenia planktonu *Calanus helgolandicus* występuje stosunkowo liczna grupa obserwacji z wysoką wartością. Mogą pochodzić one z lepszego łowiska (łowiska z większą dostępnością pokarmu). Jest to widoczne szczególnie dla gatunku pierwszego. Rozkład wartości jest jednakże mniej skupiony w okolicach zera, a bardziej rozproszony (szczególnie dla gatunku drugiego).

```
plot_lcop1 <- ggplot(content, aes(x = lcop1)) + geom_histogram(binwidth = 0.5) +
  theme_bw() + ggtitle("Widłonogi gat. 1") +
  xlab(sprintf("Zagęszczenie planktonu [j]")) + ylab("Liczba obserwacji")

plot_lcop2 <- ggplot(content, aes(x = lcop2)) + geom_histogram(binwidth = 0.5) +
  theme_bw() + ggtitle("Widłonogi gat. 2") +
  xlab(sprintf("Zagęszczenie planktonu [j]")) + ylab("Liczba obserwacji")

grid.arrange(plot_lcop1, plot_lcop2, nrow = 1)
```

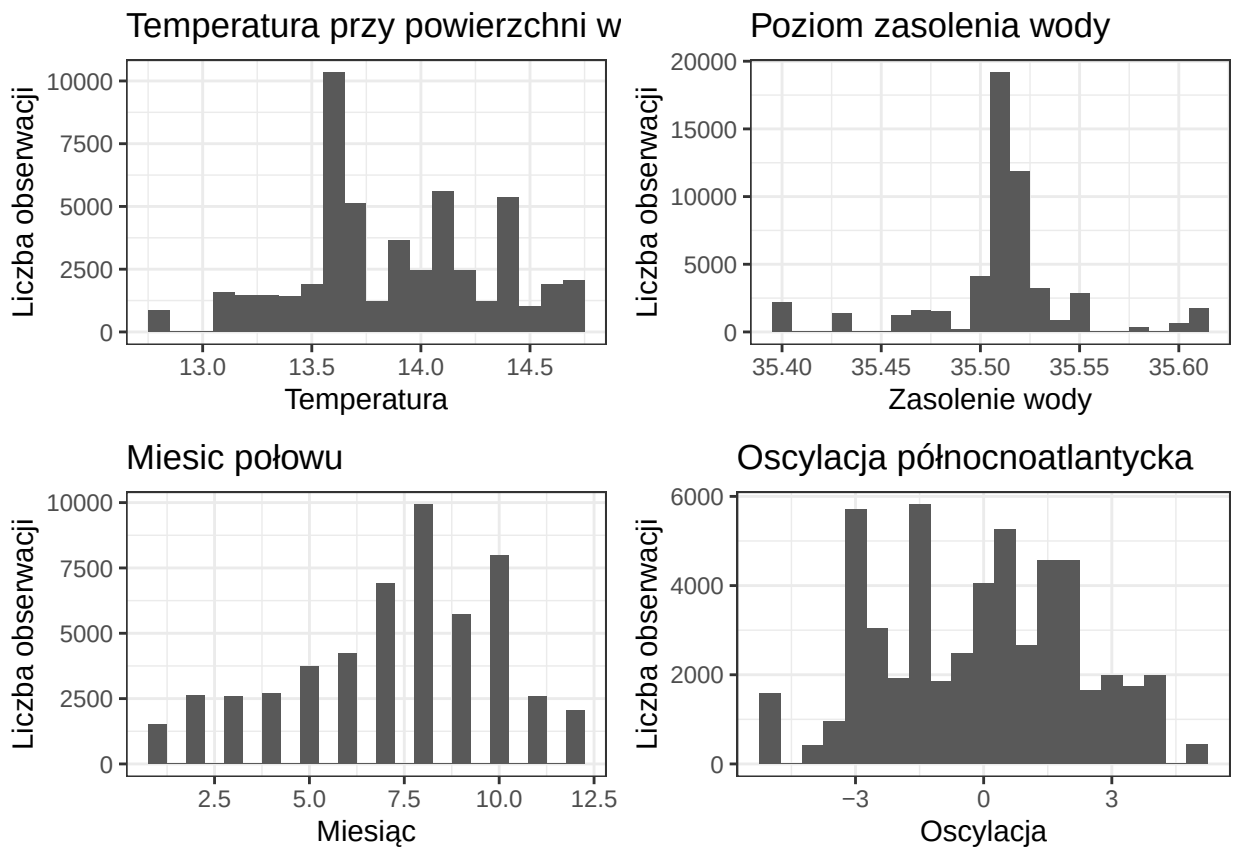


Dokonując analizy *zagęszczenia planktonu: Widłonogów* obserwujemy ponownie obserwacje odstające dla gatunku pierwszego. Gatunek drugi osiąga rozkład mniej skupiony wokół jednej wartości.

```
plot_fbar <- ggplot(content, aes(x = fbar)) + geom_histogram(binwidth = 0.05) +
  theme_bw() + ggtitle("Natężenie połowów") +
  xlab(sprintf("Ułamek pozostawionego narybku")) + ylab("Liczba obserwacji")

plot_recr <- ggplot(content, aes(x = recr)) + geom_histogram(binwidth = 50000.0) +
  theme_bw() + ggtitle("Roczny narybek") +
  xlab(sprintf("Liczba śledzi")) + ylab("Liczba obserwacji")

plot_cumf <- ggplot(content, aes(x = cumf)) + geom_histogram(binwidth = 0.02) +
  theme_bw() + ggtitle("Łączne roczne natężenie połowów") +
  xlab(sprintf("Ułamek pozostawionego narybku")) + ylab("Liczba obserwacji")

plot_totaln <- ggplot(content, aes(x = totaln)) + geom_histogram(binwidth = 1000.0) +
  theme_bw() + ggtitle("Łączna liczba złowionych ryb") +
  xlab(sprintf("Liczba śledzi")) + ylab("Liczba obserwacji")

grid.arrange(plot_fbar, plot_recr, plot_cumf, plot_totaln, nrow = 2)
```




```
plot_sst <- ggplot(content, aes(x = sst)) + geom_histogram(binwidth = 0.1) +
  theme_bw() + ggtitle("Temperatura przy powierzchni wody") +
  xlab(sprintf("Temperatura")) + ylab("Liczba obserwacji")

plot_sal <- ggplot(content, aes(x = sal)) + geom_histogram(binwidth = 0.01) +
  theme_bw() + ggtitle("Poziom zasolenia wody") +
  xlab(sprintf("Zasolenie wody")) + ylab("Liczba obserwacji")

plot_xmonth <- ggplot(content, aes(x = xmonth)) + geom_histogram(binwidth = 0.5) +
  theme_bw() + ggtitle("Miesic połowu") +
  xlab(sprintf("Miesiąc")) + ylab("Liczba obserwacji")

plot_nao <- ggplot(content, aes(x = nao)) + geom_histogram(binwidth = 0.5) +
  theme_bw() + ggtitle("Oscylacja północnoatlantycka") +
  xlab(sprintf("Oscylacja")) + ylab("Liczba obserwacji")

grid.arrange(plot_sst, plot_sal, plot_xmonth, plot_nao, nrow = 2)
```



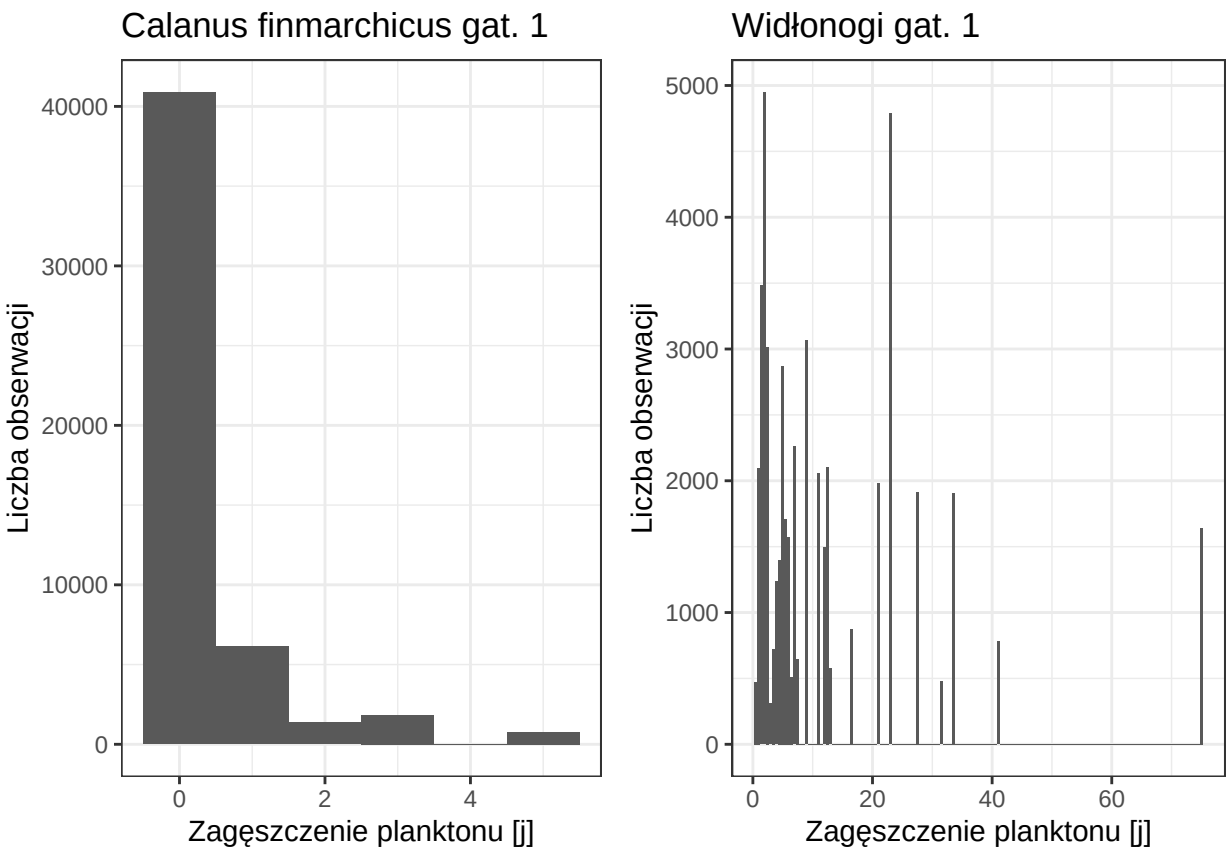
Rozkłady parametrów opisujących cechy środowiska naturalnego są zbliżone do rozkładu normalnego bądź go przypominają. Jest to szczególnie widoczne, jeżeli chodzi o miesiące połowu. Dla temperatury przy powierzchni wody możemy zaobserwować skupienie wartości w okolicy temperatury 13,8°C oraz występowanie rozbudowanej prawej części co może wskazywać na wzrost temperatury w ciągu połowów.

W przypadku parametrów dostępności planktonu *Calanus finmarchicus gat. 1* oraz *Widłonogów gat. 1* obserwujemy występowanie drobnej próbki danych odbierających znacząco od reszty. Na potrzeby dalszego przetwarzania dane zostaną oczyszczone z tych obserwacji odstających.

```
without_outliers =
  content %>%
  filter(cfin1 <= 10 | is.na(cfin1)) %>%
  filter(lcop1 <= 90 | is.na(lcop1))
```

Po operacji w zbiorze obserwacji pozostało 52576 próbek (usunięto 6 obserwacji).

```
plot_cfin1_clear <- ggplot(without_outliers, aes(x = cfin1)) + theme_bw() +
  geom_histogram(binwidth = 1.0) + xlab(sprintf("Zagęszczenie planktonu [j]")) +
  ggtitle("Calanus finmarchicus gat. 1") + ylab("Liczba obserwacji")

plot_lcop1_clear <- ggplot(without_outliers, aes(x = lcop1)) + theme_bw() +
  geom_histogram(binwidth = 0.5) + xlab(sprintf("Zagęszczenie planktonu [j]")) +
  ggtitle("Widłonogi gat. 1") + ylab("Liczba obserwacji")

grid.arrange(plot_cfin1_clear, plot_lcop1_clear, nrow = 1)
```



Rozkład wartości po usunięciu wartości odstających opisujących dostępność planktonu *Calanus finmarchicus gat. 1* oraz *Widłonogów gat. 1* wskazano powyżej.

Przetwarzanie brakujących danych

Korzystajac z pakietu `imputeTS` i funkcji `statsNA` możemy przeprowadzić analizę wartości pustych w poszczególnych obserwacjach.

```
without_outliers %>%
  colnames() %>%
  sapply(function(attr) {
    statsNA(without_outliers[[attr]], printOnly = FALSE)
  }) %>%
  kable(align = "c", caption = "Statystyka atrybutów pod względem wartości NA") %>%
  kable_styling(latex_options = "scale_down")
```

Analizując zaprezentowane podsumowania dla wszystkich atrybutów, możemy zauważyć, że wartości puste stanowią mniej niż 3.5% całego zbioru obserwacji. Ponadto ich rozkład ma charakter losowy oraz są równomierne. W danych nie występują długie serie wartości pustych (sekwencje liczące dwie oraz trzy wartości puste są rzadkie). Wykorzystując wiedzę o charakterystyce danych, możemy wykonać interpolację z wykorzystaniem filtru Kalmana, aby pozbyć się wartości pustych.

```
without_outliers$cfin1 <- na_kalman(without_outliers$cfin1)
without_outliers$cfin2 <- na_kalman(without_outliers$cfin2)
```

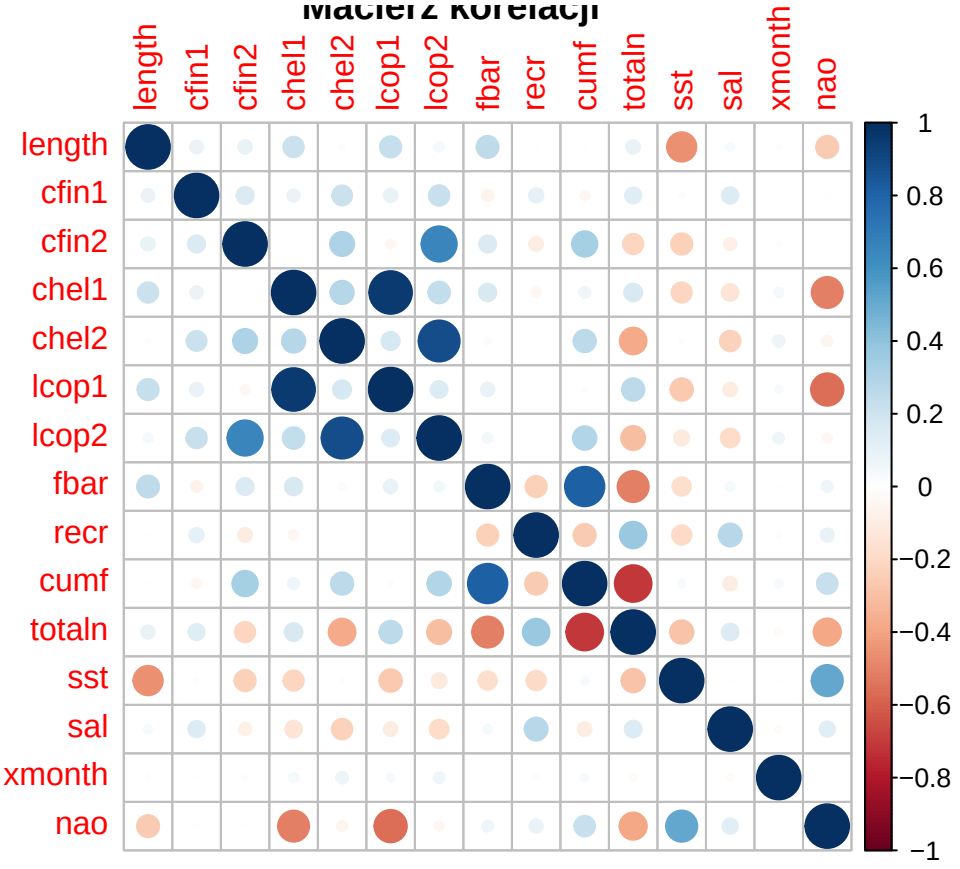
Table 3: Statystyka atrybutów pod względem wartości NA

	length	cfin1	cfin2	chel1	chel2	lcop1	lcop2	fbar	recr	cumf	totaln	sst	sal	xmonth	nao
lengthTimeSeries	52576	52576	52576	52576	52576	52576	52576	52576	52576	52576	52576	52576	52576	52576	52576
numberNAs	0	1581	1536	1555	1556	1653	1591	0	0	0	0	1584	0	0	0
percentageNAs	0%	3.01%	2.92%	2.96%	2.96%	3.14%	3.03%	0%	0%	0%	0%	3.01%	0%	0%	0%
naGapLongest	NA	3	3	3	3	2	3	NA	NA	NA	NA	3	NA	NA	NA
naGapMostFrequent	52576	1	1	1	1	1	1	52576	52576	52576	52576	1	52576	52576	52576
naGapMostOverallNAs	52576	1	1	1	1	1	1	52576	52576	52576	52576	1	52576	52576	52576

```
without_outliers$chel1 <- na_kalman(without_outliers$chel1)
without_outliers$chel2 <- na_kalman(without_outliers$chel2)
without_outliers$lcop1 <- na_kalman(without_outliers$lcop1)
without_outliers$lcop2 <- na_kalman(without_outliers$lcop2)
without_outliers$sst <- na_kalman(without_outliers$sst)
```

Korelacja atrybutów

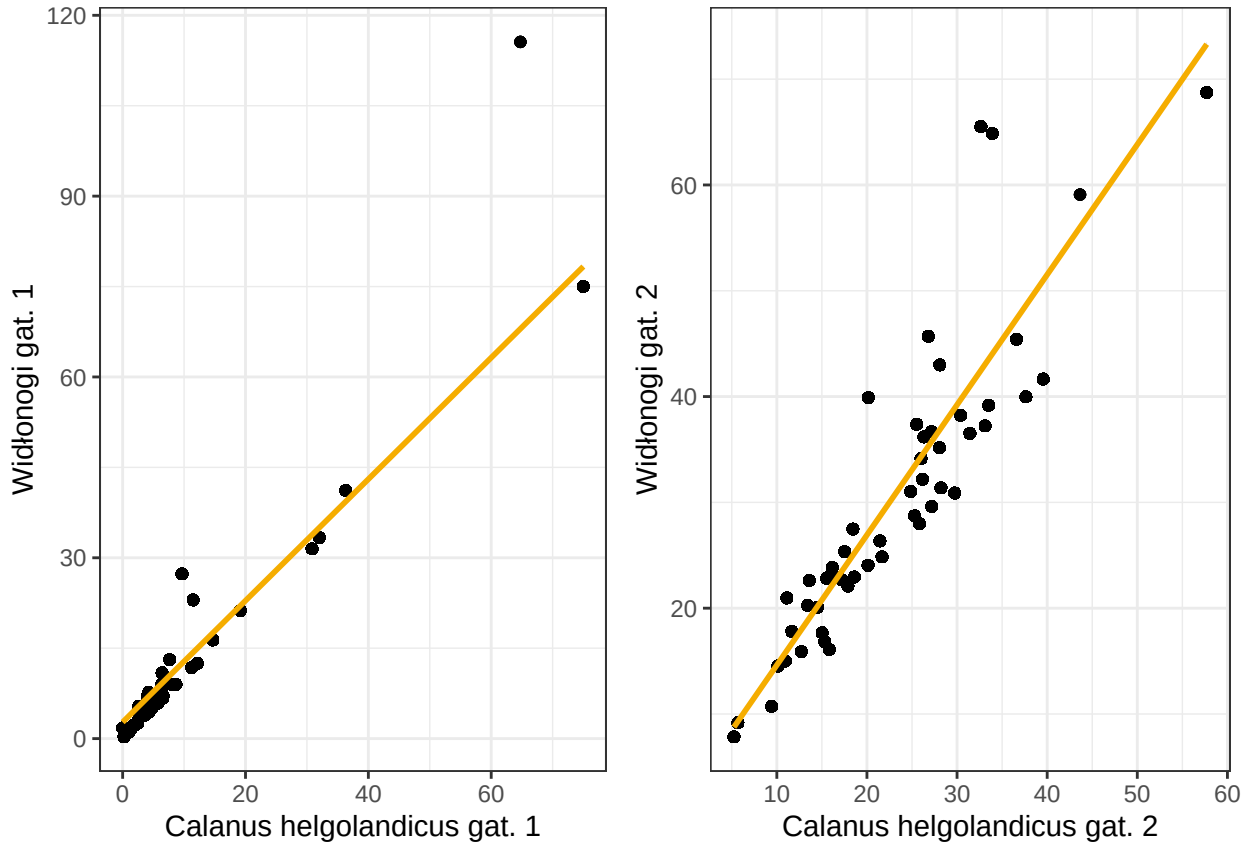
```
corelation_matrix <- cor(without_outliers)
corrplot(corelation_matrix, method = "circle", title = "Macierz korelacji")
```



Na wykresie powyższym została przedstawiona macierz korelacji pomiędzy poszczególnymi atrybutami. Jak możemy zaobserwować, istnieje bardzo silna pozytywna korelacja pomiędzy parametrem opisującym dostępność *Calanus helgolandicus gat. 1* oraz *zagęszczenie widłonogów gat. 1* wynosząca w przybliżeniu `0,96`.

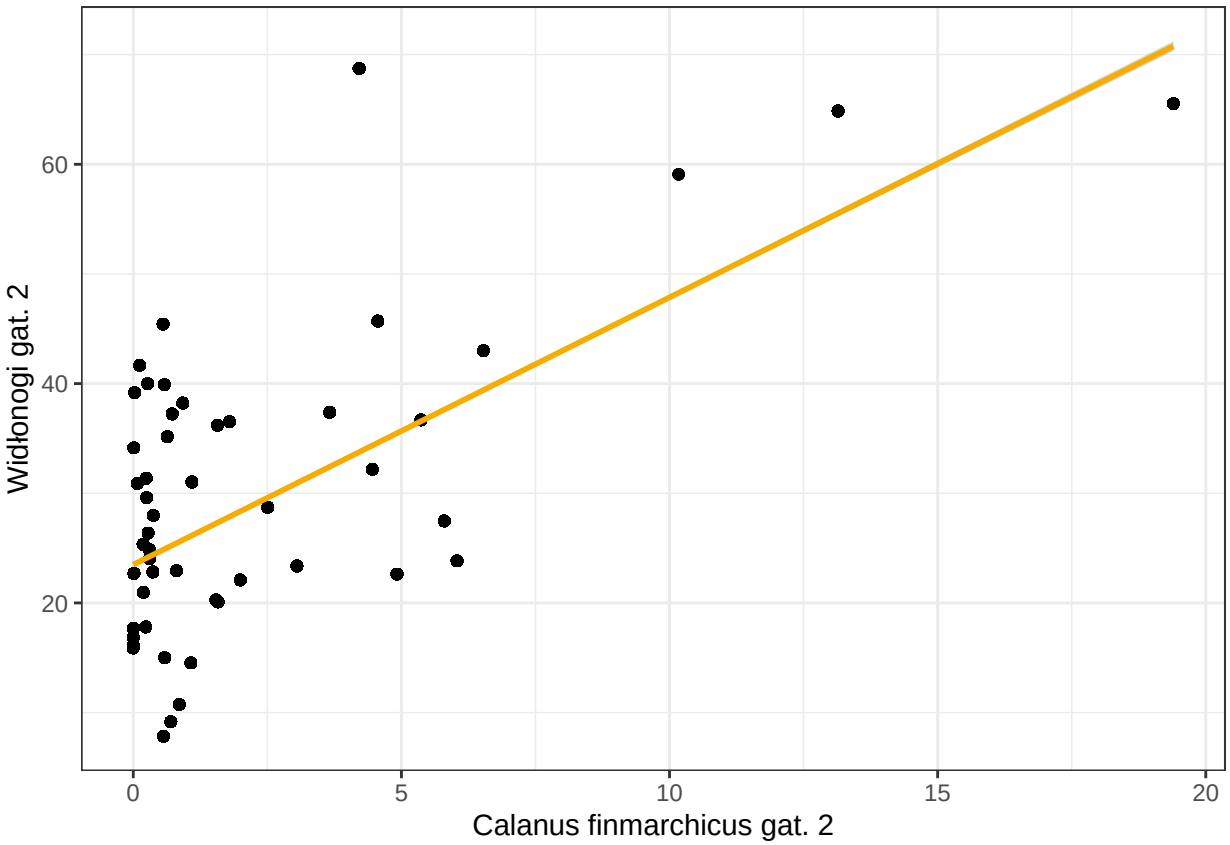
Także pomiędzy zagęszczeniem *Calanus helgolandicus gat. 2* oraz *zagęszczenie widłonogów gat. 2* możemy zaobserwować korelację wynoszącą `0,88`. Wynika z tego, że występowanie planktonu *Calanus helgolandicus gat. 1* związane jest z obecnością *widłonogów gat. 1* i vice versa. Podobnie w przypadku planktonów drugiego gatunku, czyli pary *Calanus helgolandicus gat. 2* oraz *widłonogów gat. 2*.

```
plot_chel1_lcop1 <- ggplot(content, aes(chel1, lcop1)) + geom_point() +  theme_bw() +
  geom_smooth(color = "#f5ad00", method = "lm") + ylab(sprintf("Widłonogi gat. 1")) +
  xlab("Calanus helgolandicus gat. 1")

plot_chel2_lcop2 <- ggplot(content, aes(chel2, lcop2)) + geom_point() +  theme_bw() +
  geom_smooth(color = "#f5ad00", method = "lm") + ylab(sprintf("Widłonogi gat. 2")) +
  xlab("Calanus helgolandicus gat. 2")

grid.arrange(plot_chel1_lcop1, plot_chel2_lcop2, nrow = 1)
```



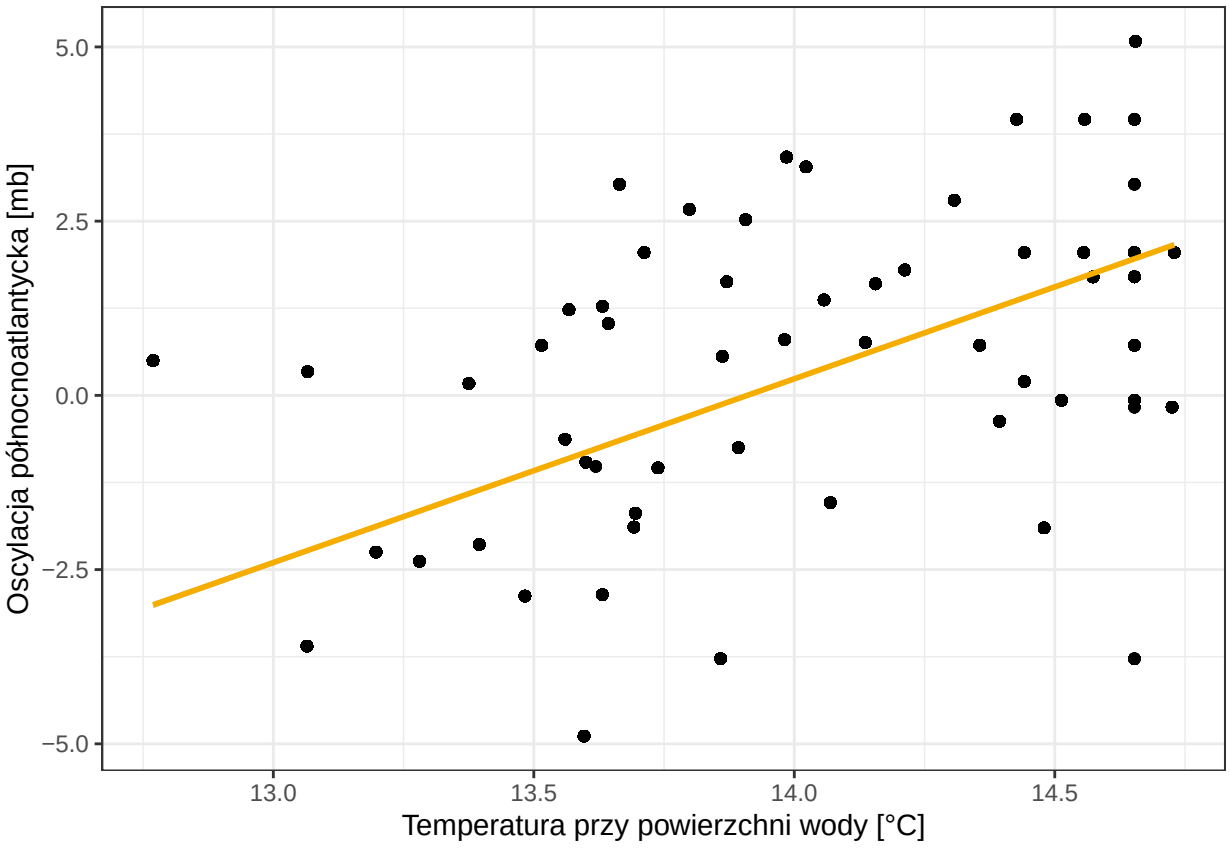
Analizując dalej macierz korelacji, możemy zaobserwować pozytywną zależność pomiędzy parametrami `cfin2` i `lcop2` wynosząca `0,65` - zagęszczenie *Calanus finmarchicus gat. 2* ma powiązanie w obecności *widłonogów gat. 2*.

```
ggplot(content, aes(cfin2, lcop2)) + geom_point() +  theme_bw() +
  geom_smooth(color = "#f5ad00", method = "lm") +
  ylab(sprintf("Widłonogi gat. 2")) + xlab("Calanus finmarchicus gat. 2")
```



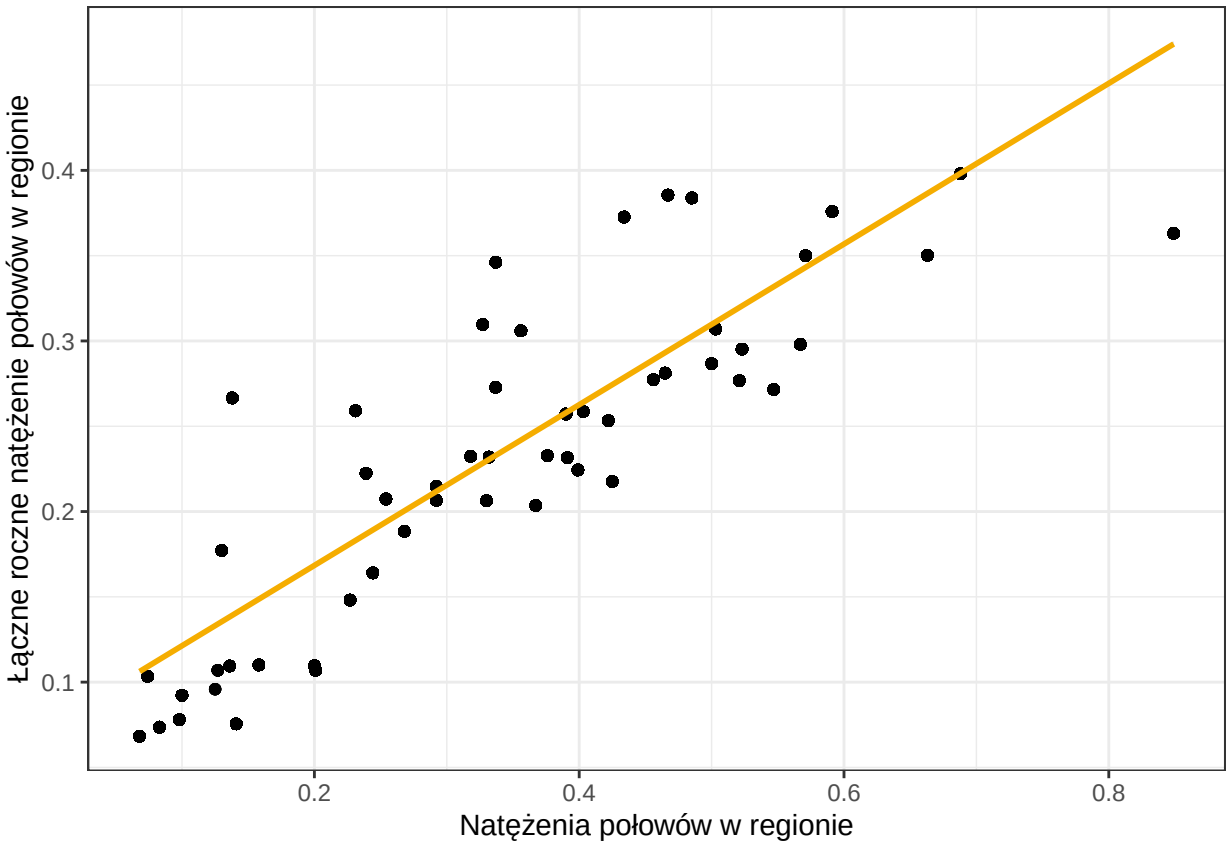
Ciekawą zależnością jest przypadek parametrów `sst` oraz `nao`. Korzystając z opisu `oscylacji północnoatlantyckiej` na stronie encyklopedii Wikipedia mamy do czynienia ze zjawiskiem meteorologicznym wpływającym na klimat, co manifestuje się między innymi zmianą temperatury. Podkreśla to wiarygodność naszych obserwacji, gdyż doszło do odwzorowania zjawiska fizycznego w naszych danych.

```
ggplot(content, aes(sst, nao)) + geom_point() +  theme_bw() +
  geom_smooth(color = "#f5ad00", method = "lm") +
  ylab(sprintf("Oscylacja północnoatlantycka [mb]")) +
  xlab("Temperatura przy powierzchni wody [°C]")
```



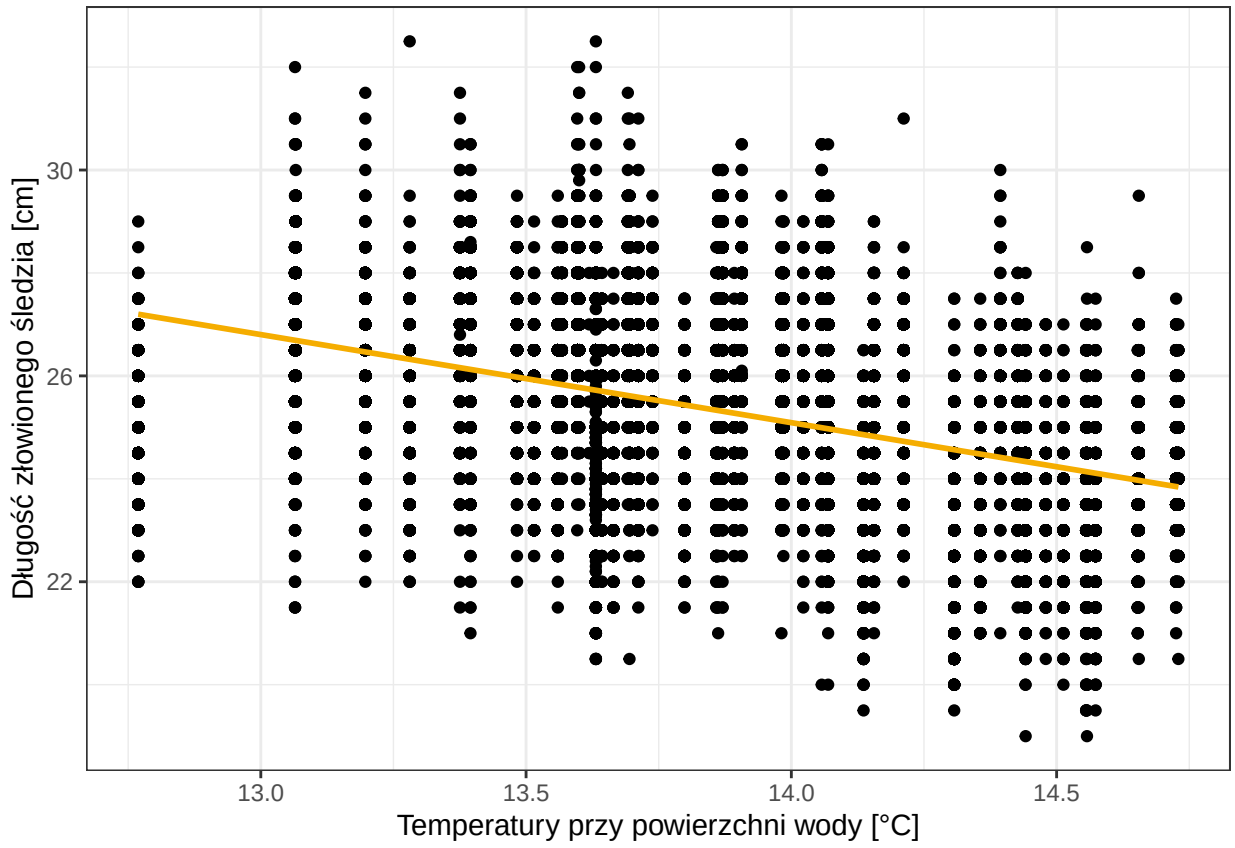
Wysoką wartość zależności `fbar` oznaczającej *natężenia połowów w regionie* oraz `cumf` czyli *łączne roczne natężenie połowów w regionie* wynoszącej `0,82` można łatwo wyjaśnić. Łowienie w danym miejscu przez długi czas sumarycznie wpłynie na wysoką wartość drugiego parametru.

```
ggplot(content, aes(fbar, cumf)) + geom_point() +  theme_bw() +
  geom_smooth(color = "#f5ad00", method = "lm") +
  ylab(sprintf("Łączne roczne natężenie połowów w regionie")) +
  xlab("Natężenia połowów w regionie")
```



Interesującą z punktu widzenia tematu analizy, jest zależność *temperatury przy powierzchni wody* i *długości złowionego śledzia*. Wynosi ona `-0,45`. **Większa temperatura ma odzwierciedlenie w mniejszych rozmiarach śledzi**.

```
ggplot(content, aes(sst, length)) + geom_point() + theme_bw() +
  geom_smooth(color = "#f5ad00", method = "lm") +
  ylab(sprintf("Długość złowionego śledzia [cm]")) +
  xlab("Temperatury przy powierzchni wody [°C]")
```

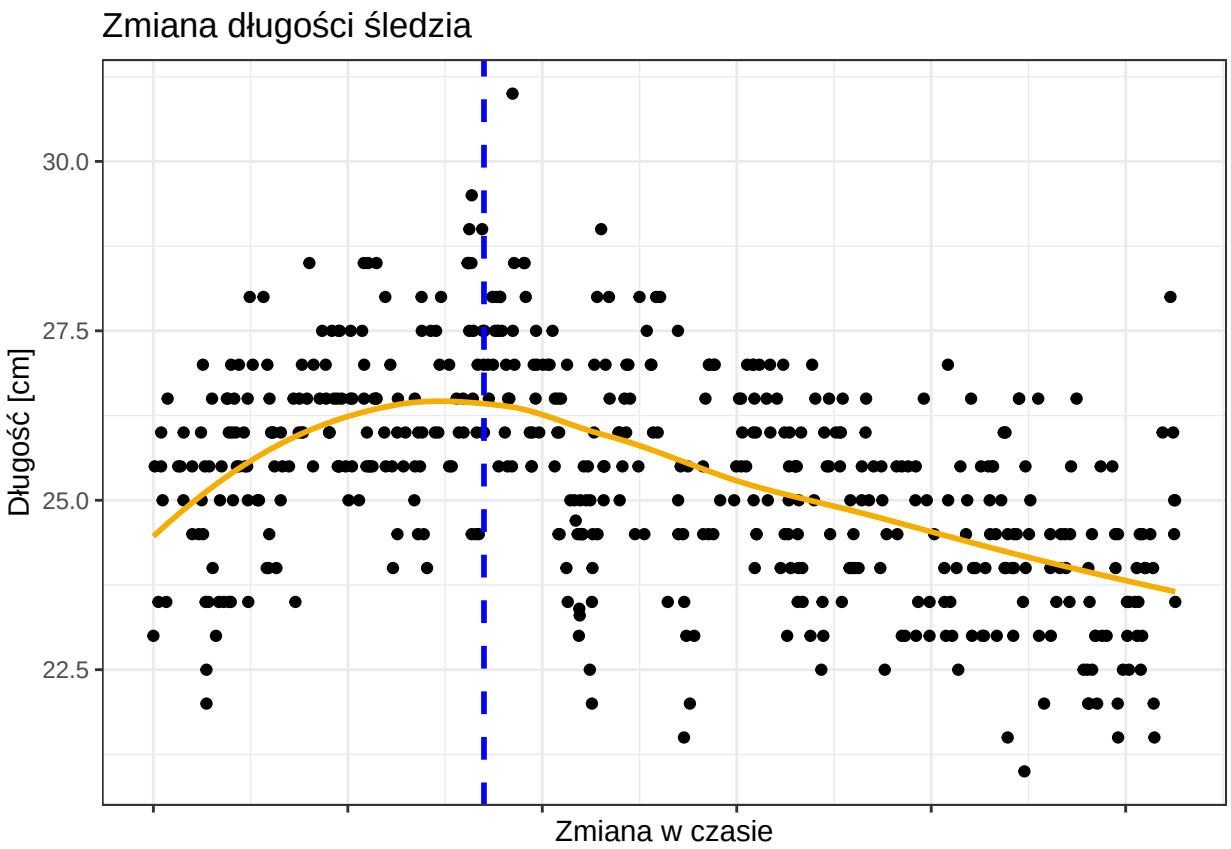


Zmienność cech w ramach następujących po siebie połowów

W kolejnych podrozdziałach zostanie przeanalizowana zmienność cech. Naszym celem jest wykrycie przyczyny spadku długości śledzi w połowach.

Długość śledzi

```
ggplot(sampled_data, aes(x=id, y=length)) + theme_bw() + geom_point() +
  theme(axis.text.x=element_blank()) + ylab("Długość [cm]") + xlab("Zmiana w czasie") +
  geom_smooth(method = "loess", formula = y ~ x, se = FALSE, colour = "#f5ad00",
              size = 1.0) + ggtitle("Zmiana długości śledzia") +
  geom_vline(xintercept = 17000, colour="blue", linetype = 2, size = 1.0)
```



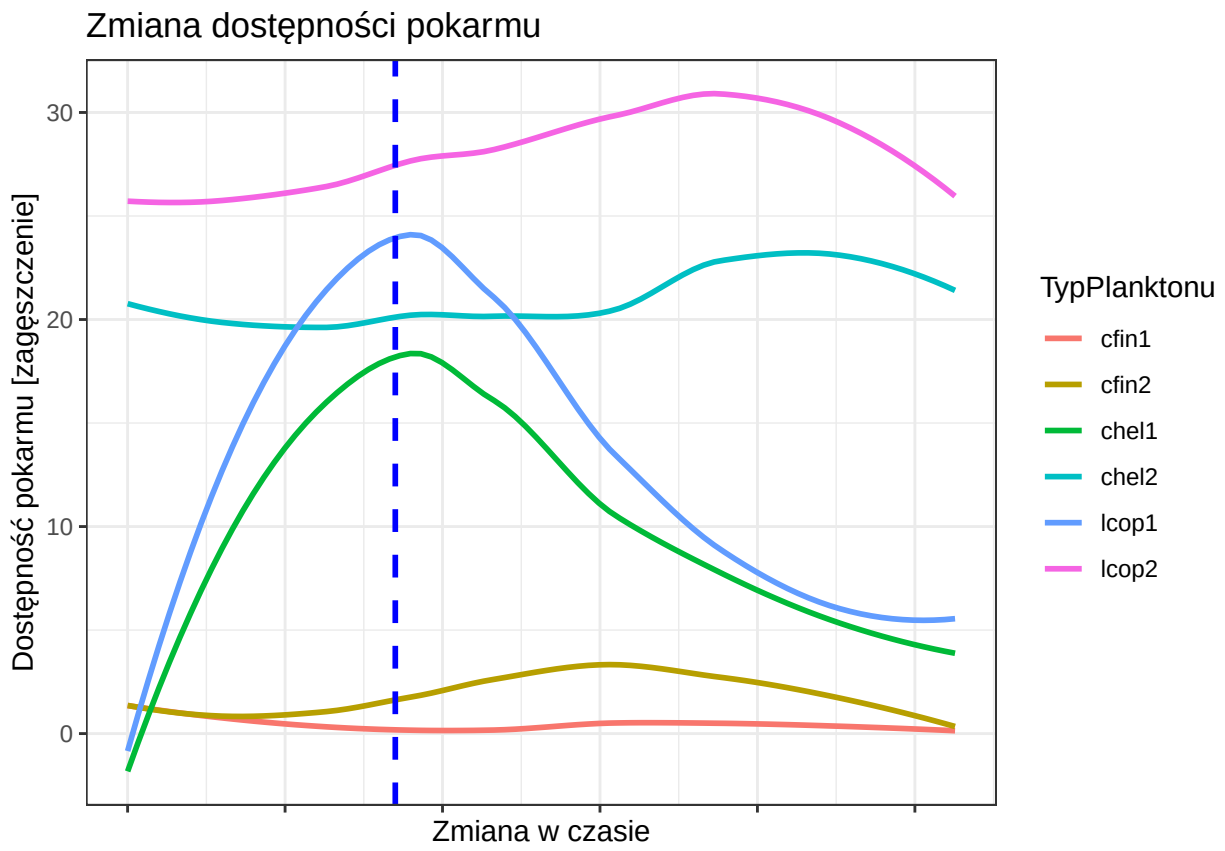
Z wykresu przestawiającego zmianę długości śledzi w czasie, możemy zaobserwować odwrócenie tendencji. Na początku rozmiar wzrastał z około `24,5 cm` do `26 cm` aby następnie spaść poniżej `23,5 cm`. Za pomocą niebieskiej linii oznaczono punkt przed rozpoczęciem spadku. Moment w czasie (na podstawie historii obserwacji) zostanie wykorzystany jako punkt referencyjny w kolejnych wykresach.

```
ggplot(
  sampled_data,
  group = xmonth,
  aes(x=id, y=length)
) + theme_bw() +
  geom_line() + transition_reveal(id)
```

TODO: W MD i HTML animacja, w PDF bez

Dostępność pokarmu

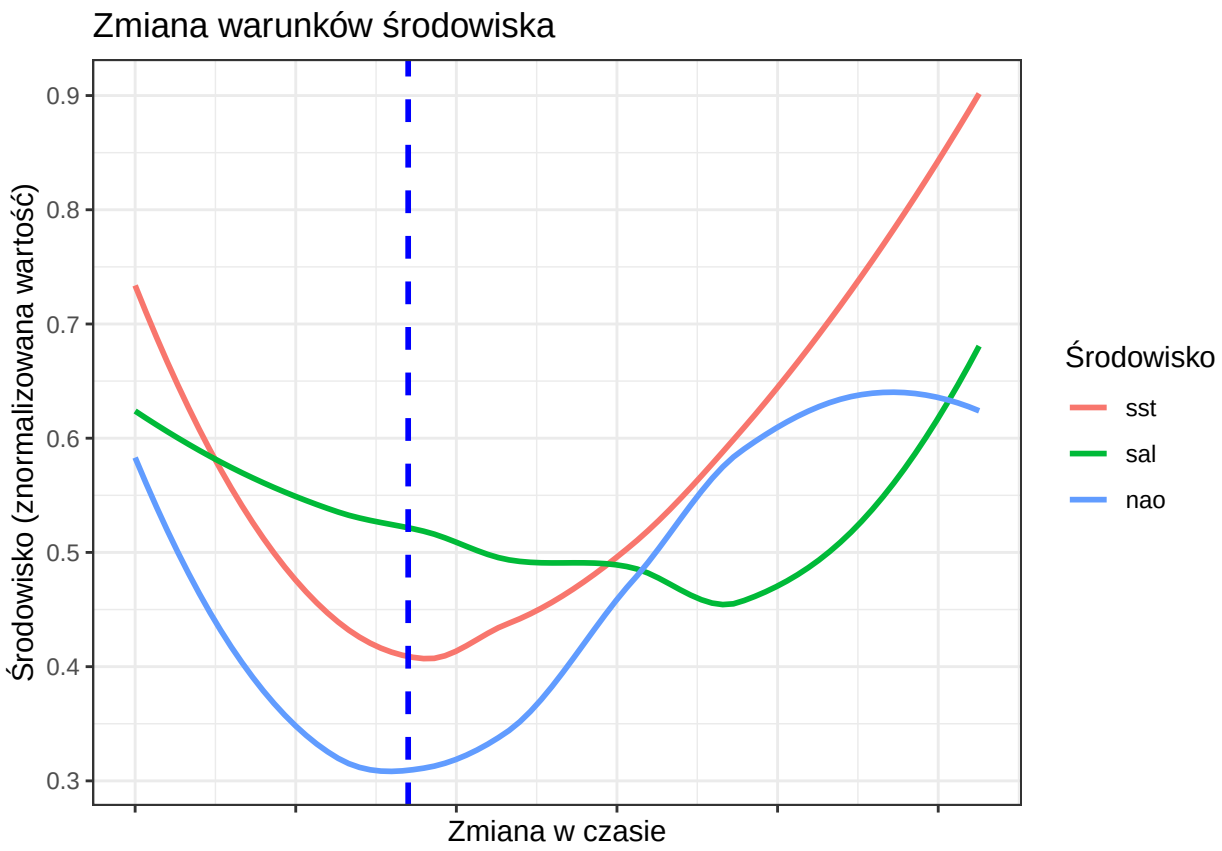
```
dostepnosc_planktonu <- melt(sampled_data[, c(16, 2:7)], id.vars = c("id"),
                             variable.name = "TypPlanktonu", value.name = "Values")
ggplot(dostepnosc_planktonu, aes(id, Values, color = TypPlanktonu)) + theme_bw() +
  theme(axis.text.x=element_blank()) + geom_smooth(se = FALSE) +
  ggtitle("Zmiana dostępności pokarmu") +  xlab("Zmiana w czasie") +
  ylab("Dostępność pokarmu [zagęszczenie]") +
  geom_vline(xintercept = 17000, colour="blue", linetype = 2, size = 1.0)
```



Analizując zestawienie dostępności pokarmu na łowiskach, obserwujemy znaczną zmianę dwóch parametrów. Są to *zagęszczenie widłonogów gat. 1* oraz *zagęszczenie Calanus helgolandicus gat. 1*. W przypadku pozostałych nie obserwujemy znacznych zmian wartości, jedynie drobne fluktuacje.

Parametry środowiska

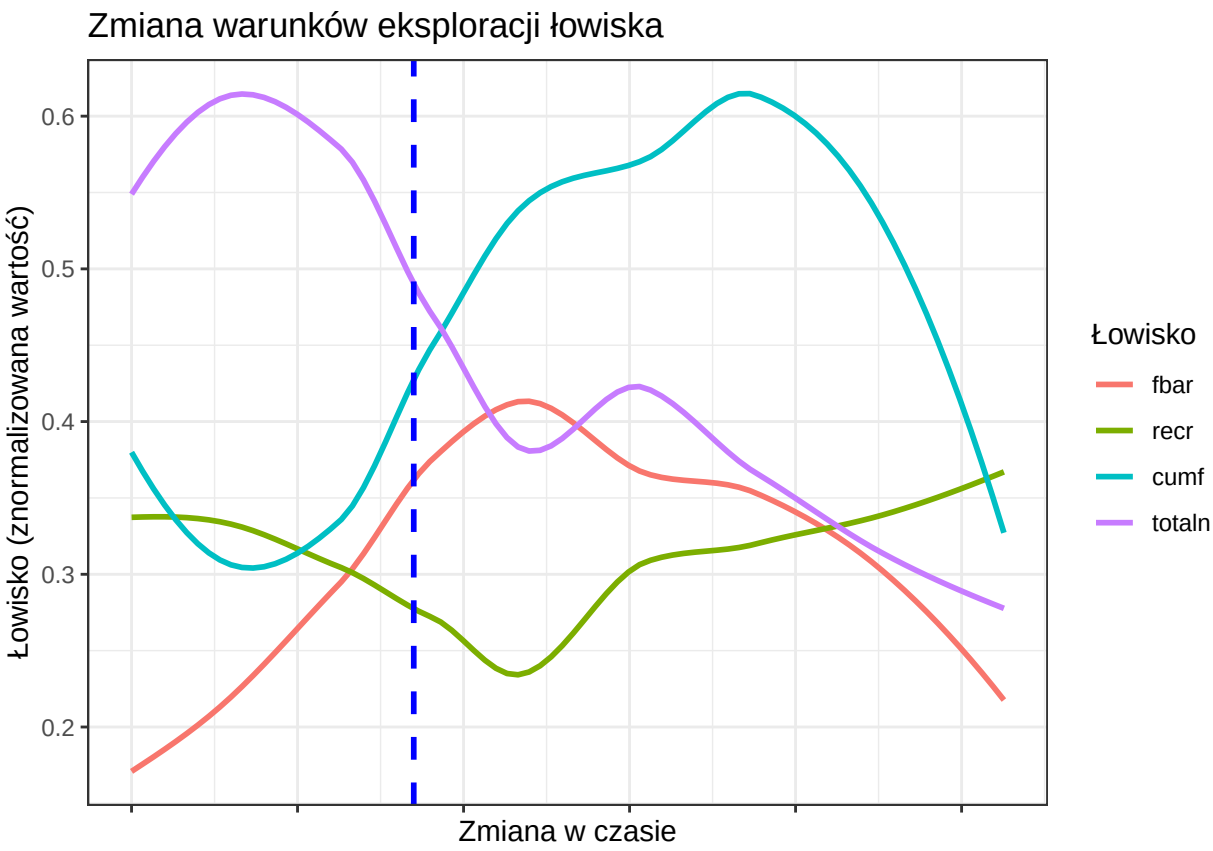
```
parametry_srodowiska <- sampled_data[, c(12, 13, 15)]
normalized_environment <- as.data.frame(lapply(parametry_srodowiska, function(x) {
  (x - min(x)) / (max(x) - min(x))
}))
normalized_environment["id"] <- sampled_data[, 16]

melt(normalized_environment, id.vars = c("id"), variable.name = "Środowisko",
     value.name = "Values") %>% ggplot(aes(id, Values, color = Środowisko)) +
  theme_bw() + theme(axis.text.x=element_blank()) + geom_smooth(se = FALSE) +
  ggtitle("Zmiana warunków środowiska") + xlab("Zmiana w czasie") +
  ylab("Środowisko (znormalizowana wartość)") +
  geom_vline(xintercept = 17000, colour="blue", linetype = 2, size = 1.0)
```



Zmiana środowiska dotyczy głównie parametrów *oscylacji północnoatlantyckiej* oraz *temperatury przy powierzchni wody*.

Eksploatacja łowiska

```
parametry_lowiska <- sampled_data[, c(8:11)]
normalized_lowisko <- as.data.frame(lapply(parametry_lowiska, function(x) {
  (x - min(x)) / (max(x) - min(x))
}))
normalized_lowisko["id"] <- sampled_data[, 16]

melt(normalized_lowisko, id.vars = c("id"), variable.name = "Łowisko",
     value.name = "Values") %>% ggplot(aes(id, Values, color = Łowisko)) +
  theme_bw() + theme(axis.text.x=element_blank()) + geom_smooth(se = FALSE) +
  ggtitle("Zmiana warunków eksploracji łowiska") + xlab("Zmiana w czasie") +
  ylab("Łowisko (znormalizowana wartość)") +
  geom_vline(xintercept = 17000, colour="blue", linetype = 2, size = 1.0)
```



Regresor - predykcja

W ramach naszego eksperymentu pragniemy przygotować, poza analizą wpływu czynników otoczenia na długość śledzi również regresor pozwalający przewidywać owy rozmiar na podstawie parametrów środowiska. W tym celu wykorzystamy wiedzę zdobytą na wcześniejszych etapach analizy problemu.

Zbiór danych podzielimy na część uczącą oraz testową, aby zminimalizować ryzyko przeuczenia naszego modelu. W ramach zbioru uczącego wykorzystamy zbiór walidujący (działanie kontrolowane przez bibliotekę `caret`).

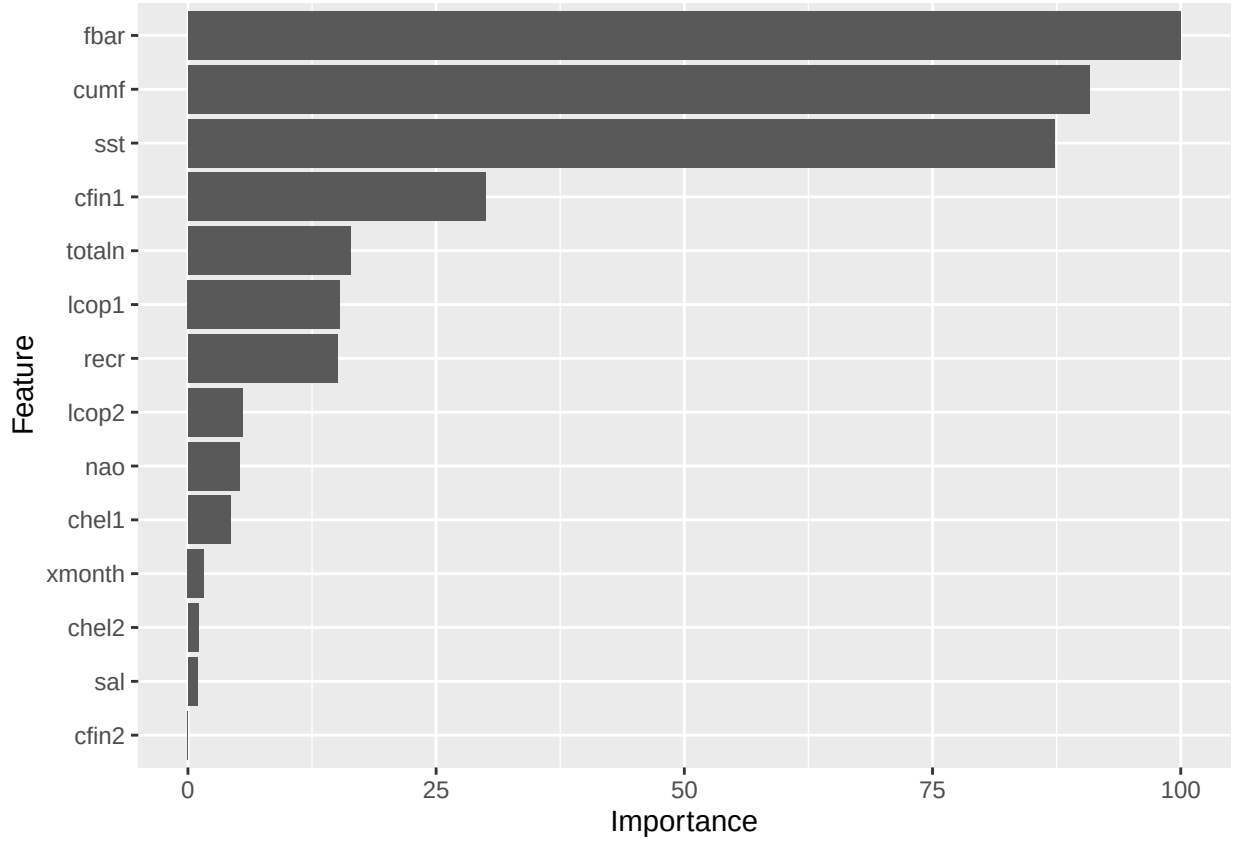
```
indexesInTraningSet <- createDataPartition(y = without_outliers$length,
                                           p = 0.75, list = FALSE)
trainingSet <- without_outliers[indexesInTraningSet, ]
testSet <- without_outliers[-indexesInTraningSet, ]

ctrl <- trainControl(method = "repeatedcv", number = 10, repeats = 10,
                     allowParallel = TRUE)
```

Zastosujemy uczenie z wykorzystaniem powtórzonej oceny krzyżowej (ang. `repeated cross validation`) z dziesięcioma podziałami oraz dziesięciokrotnym powtórzeniem.

Naszym pierwszym regresorem będzie regresja liniowa.

```
model_linear_regression <- train(length ~ ., data = trainingSet, method = "lm",
                                 trControl = ctrl)

predicted_linear_regression <- predict(model_linear_regression,
                                       newdata = testSet)
predicted_linear_regression <- sapply(predicted_linear_regression,
                                      round, digits = 0)

expected_values <- sapply(testSet$length, round, digits = 0)
levels <- unique(c(expected_values, predicted_linear_regression))

result <- confusionMatrix(data = factor(predicted_linear_regression, levels = levels),
                factor(expected_values, levels = levels))
```

	x
Accuracy	0.2877577
Kappa	0.0985870
AccuracyLower	0.2800276
AccuracyUpper	0.2955820
AccuracyNull	0.3240508
AccuracyPValue	1.0000000
McnemarPValue	NaN

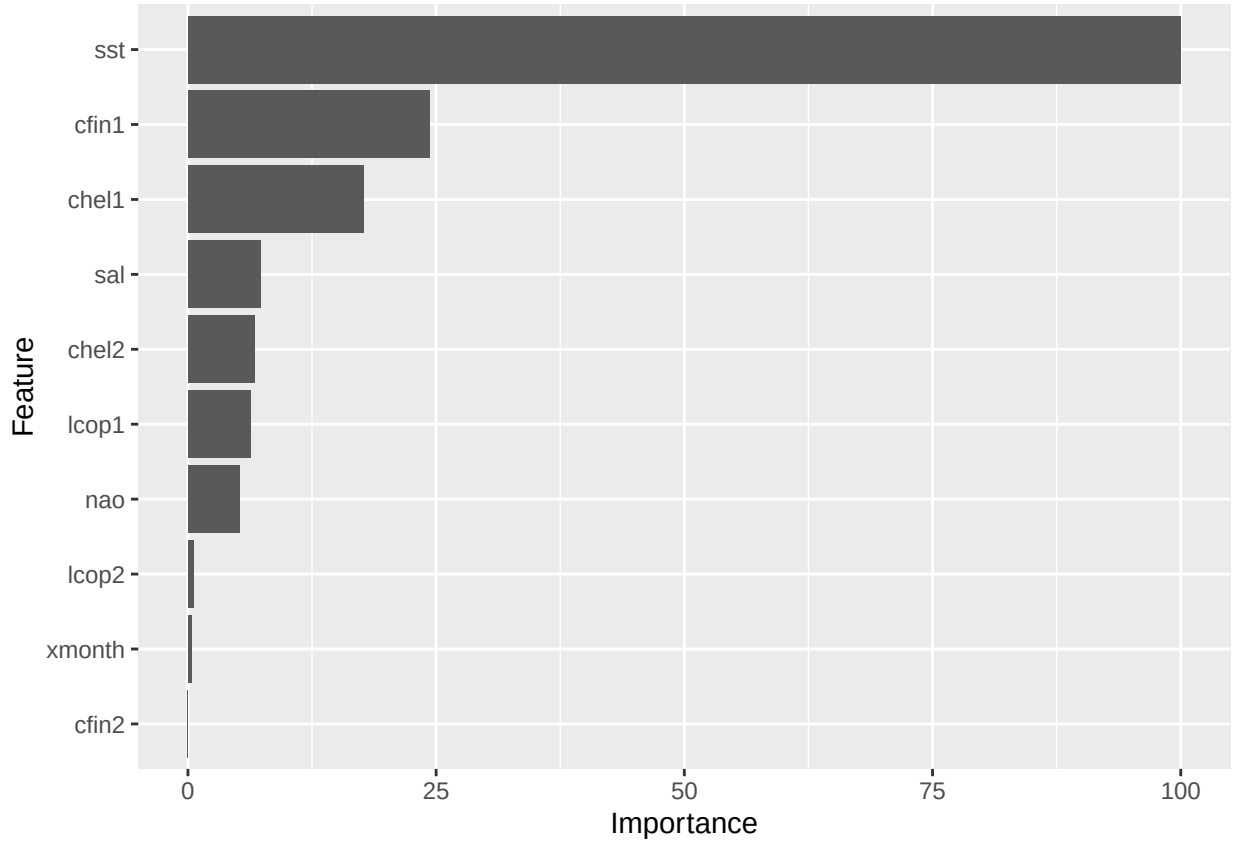


W ramach eksperymentu, jako zbiór danych zastosujemy oryginalny zbiór danych z pominięciem charakterystyk dotyczących połowu na łowiskach:

- **length**: długość złowionego śledzia [cm]
- **cfin1**: dostępność planktonu [zagęszczenie *Calanus finmarchicus* gat. 1]
- **cfin2**: dostępność planktonu [zagęszczenie *Calanus finmarchicus* gat. 2];
- **chel1**: dostępność planktonu [zagęszczenie *Calanus helgolandicus* gat. 1];
- **chel2**: dostępność planktonu [zagęszczenie *Calanus helgolandicus* gat. 2];
- **lcop1**: dostępność planktonu [zagęszczenie *widłonogów* gat. 1];
- **lcop2**: dostępność planktonu [zagęszczenie *widłonogów* gat. 2];
- **sst**: temperatura przy powierzchni wody [°C];
- **sal**: poziom zasolenia wody [Knudsen ppt];
- **xmonth**: miesiąc połowu [numer miesiąca];
- **nao**: oscylacja północnoatlantycka [mb].

```
model_linear_preproc <- train(length ~ ., data = trainingSet[, -c(8:11)],
                              method = "lm", trControl = ctrl)

predicted_linear_preproc <-
  model_linear_preproc %>%
  predict(newdata = testSet[, -c(8:11)]) %>%
  sapply(round, digits = 0)

expected_values <- sapply(testSet$length, round, digits = 0)
levels <- unique(c(expected_values, predicted_linear_preproc))

result <- confusionMatrix(data = factor(predicted_linear_preproc, levels = levels),
                factor(expected_values, levels = levels))
```

	x
Accuracy	0.3028228
Kappa	0.0930971
AccuracyLower	0.2949737
AccuracyUpper	0.3107592
AccuracyNull	0.3240508
AccuracyPValue	0.9999999
McnemarPValue	NaN

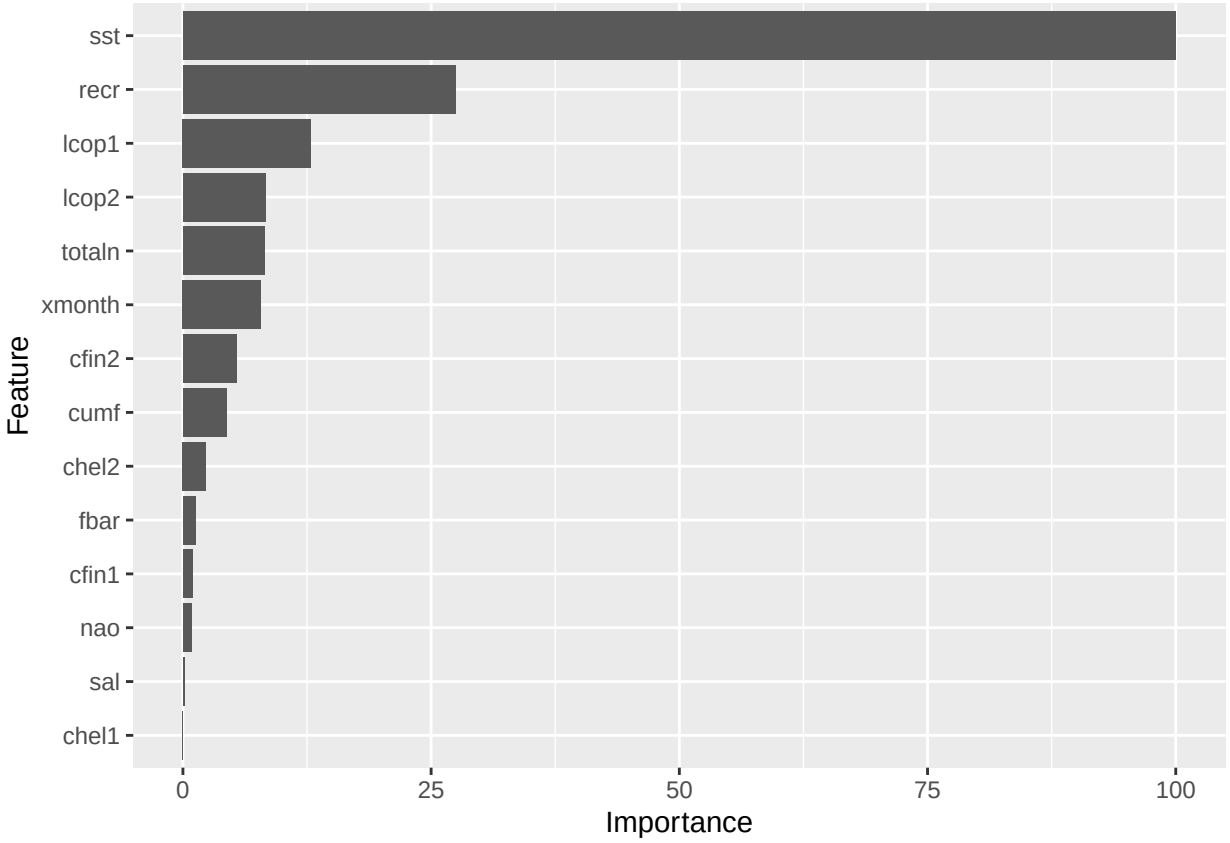


Ostatnim modelem będzie `eXtreme Gradient Boosting`. W ramach uczenia zastosujemy macierz parametrów.

```
grid = expand.grid(
  nrounds = c(10, 20, 50, 100),
  alpha = c(1, 0.7, 0.3, 0.1, 0),
  lambda = c(1, 0.7, 0.3, 0.1, 0),
  eta = 0.3
)


model_xgb <- train(length ~ ., data = trainingSet, method = "xgbLinear",
                   trControl = ctrl, tuneGrid = grid, max_depth = 5)

predicted_xgb <- predict(model_xgb, newdata = testSet)
predicted_xgb <- sapply(predicted_xgb, round, digits = 0)

expected_values <- sapply(testSet$length, round, digits = 0)
levels <- unique(c(expected_values, predicted_xgb))

result <- confusionMatrix(data = factor(predicted_xgb, levels = levels),
                factor(expected_values, levels = levels))
```

	x
Accuracy	0.3059423
Kappa	0.1594843
AccuracyLower	0.2980699
AccuracyUpper	0.3139008
AccuracyNull	0.3240508
AccuracyPValue	0.9999960
McnemarPValue	NaN

	nrounds	lambda	alpha	eta
97	100	0.1	1	0.3



Porównanie modeli

```
resampled_models <-
  list(linear = model_linear_regression,
       linear_preprocess = model_linear_preproc,
       xgb = model_xgb) %>% resamples()
stats <- summary(resampled_models)

stats$statistics %>% kable(align = "c", caption = "Porównanie regresorów")
```

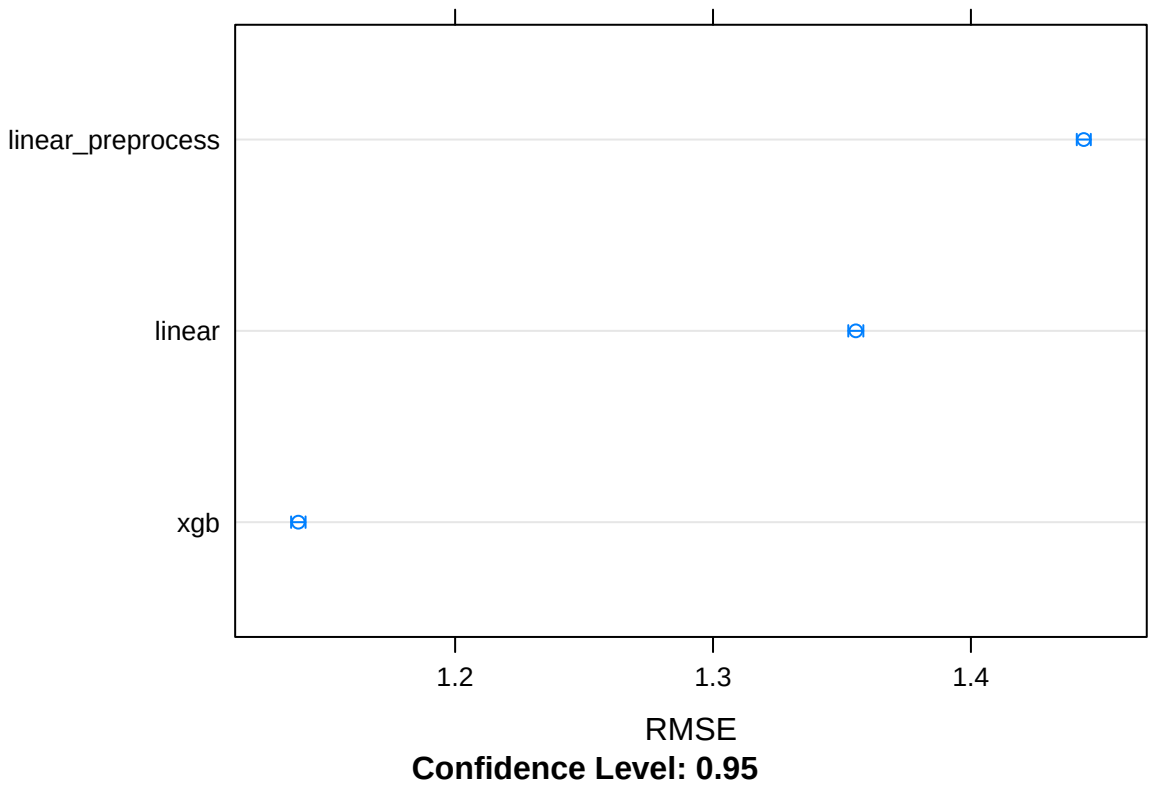
Table 4: Porównanie regresorów

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
linear	1.0503160	1.0687982	1.0770041	1.0779126	1.086004	1.1051476	0
linear_preprocess	1.1264511	1.1489647	1.1573950	1.1564836	1.164328	1.1842091	0
xgb	0.8741322	0.8914763	0.8973731	0.8979285	0.905810	0.9298612	0

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
linear	1.324775	1.344401	1.354436	1.355377	1.365470	1.387657	0
linear_preprocess	1.408171	1.434654	1.444064	1.443766	1.453226	1.476833	0
xgb	1.107851	1.129791	1.138798	1.139204	1.148444	1.174290	0

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
linear	0.2927245	0.3168098	0.3229482	0.3233796	0.3324098	0.3477401	0
linear_preprocess	0.2080042	0.2239755	0.2319444	0.2322573	0.2405701	0.2568149	0
xgb	0.4989756	0.5135854	0.5220140	0.5219858	0.5294956	0.5486606	0

```
dotplot(resampled_models, metric = "RMSE")
```



Do porównania regresorów użyto miary `RMSE`, której im mniejsza wartość, tym lepiej. Najlepszym z regresorów okazał się `xgbLinear`. Biorąc pod uwagę wskazane najważniejsze parametry modelu, można podtrzymać wcześniejszą obserwację. Zmiana temperatury przy powierzchni wody ma znaczący wpływ na wielkość śledzi. W przypadku zmian dostępności planktonu możemy mieć do czynienia z reakcją na zmianę środowiska (zmianę temperatury), co w bezpośredni sposób wpłynęło na rozmiar złowionych śledzi.