ALEKSANDRA PEŁKA - 404407
GEOINFORMATYKA

PROJEKT

ANALIZA I INTERPRETACJA MODELI REGRESYJNYCH

KRAKÓW, 19.06.2022

SPIS TREŚCI:

CEL **2**

1. PREPROCESSING **2**
- Import danych, sprawdzenie struktury danych 2
- Wykonanie wykresów sprawdzających występowanie wartości błędnych i ekstremalnych 2
- Sprawdzenie czy w danych występują wartości ekstremalne 4
- Sprawdzenie występowania składowych szeregu czasowego 4
- Korelacja 6
- Wybór i uzasadnienie modelu 6

2. MODEL CART **7**
- Wykonanie modelu 7
- Ocena jakości modelu 8
- Ocena jakości reszt 9
- Ocena jakości prognoz 9

3. MODEL REGRESJI WIELORAKIEJ **10**
- Wykonanie modelu 10
- Ocena jakości modelu 10
- Ocena jakości reszt 11
- Ocena jakości prognoz 11

4. WNIOSKI **12**

5. LITERATURA **12**

CEL: Wykonanie modeli regresyjnych, zgodnie z metodyką CRISP-dm, na podstawie danych 257440.txt przedstawiających stosunek izotopu tlenu na przestrzeni ok. 2,6 mln lat, a także dokonanie analizy i interpretacji modeli.

1. PREPROCESSING

➢ Import danych, sprawdzenie struktury danych

Pierwszym wykonanym krokiem było wczytanie i wyświetlenie danych w celu sprawdzenia, czy zostały poprawnie zaimportowane. Podczas importu usunięto białe wiersze oraz wiersze zawierające odwołanie do literatury, a także opis poszczególnych kolumn. Następnie, w programie RStudio, nadano pomocnicze nazwy kolumn.

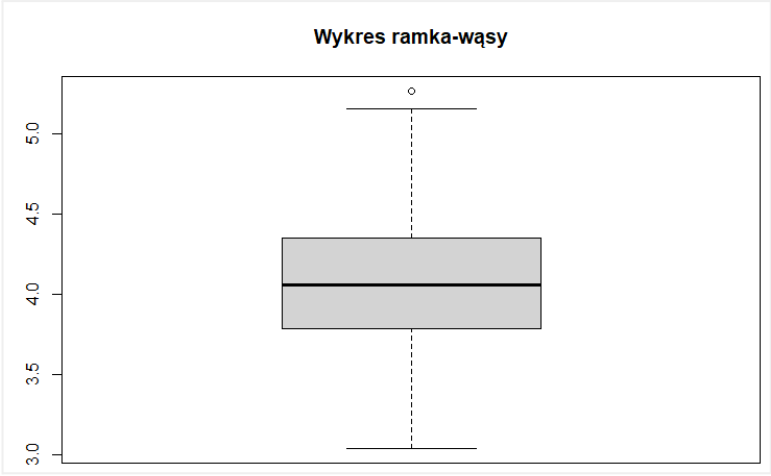
Zaimportowane dane dotyczą stosunku izotopu tlenu pomierzonego w równych odstępach czasu, co 3000 lat. Kolejno sprawdzono strukturę danych, liczbę obserwacji, wyświetlono pierwsze i ostatnie rekordy. Na podstawie otrzymanych rezultatów stwierdzono, że dane to szereg czasowy zawierający 866 obserwacji dla 2 zmiennych numerycznych: pierwsza kolumna to odwrócony czas z krokiem 3000 lat, natomiast druga kolumna prezentuje stosunek izotopu tlenu (na podstawie książki West & Harrison, "Bayesian Forecasting & Dynamic Models", chapter 15.3.4). Dodatkowo, po wyświetleniu kilku pierwszych, czy ostatnich rekordów, w danych nie zaobserwowano żadnych niepokojących wartości.

➢ Wykonanie wykresów sprawdzających występowanie wartości błędnych i ekstremalnych

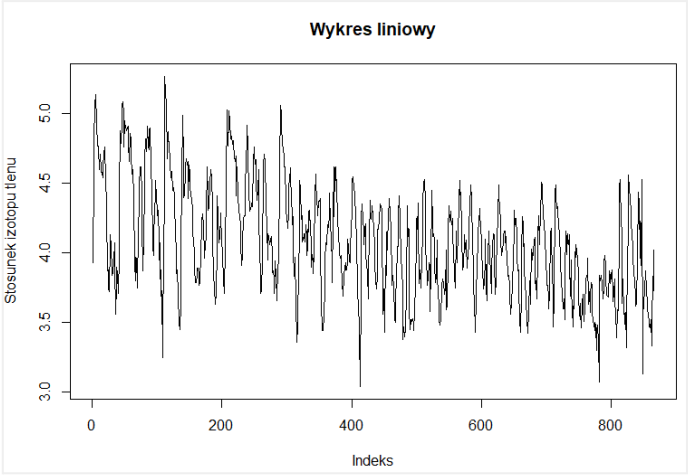
W celu wizualizacji niepokojących wartości wykonano wykres pudełkowy, liniowy oraz histogram, a także obliczono podstawowe statystyki. Wyniki zilustrowano poniżej (**Rys. 1** - **Rys. 4**).

```
> summary(dane$isotope)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  3.040   3.790   4.060   4.088   4.350   5.270 
```

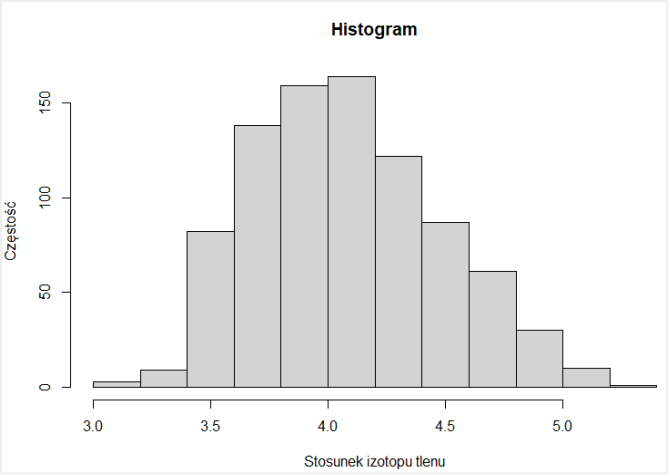
Rys. 1 *Podstawowe statystyki opisowe.*



Rys. 2 *Wykres ramka-wąsy.*



Rys. 3 *Wykres liniowy.*



Rys. 4 *Histogram.*

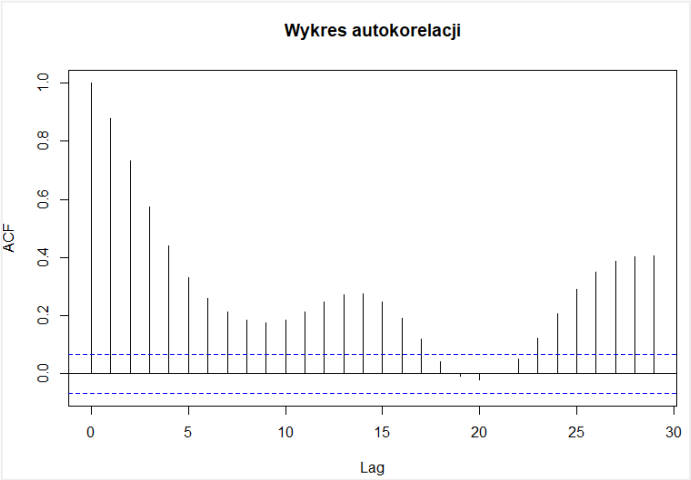
Analizując powyższe wykresy (**Rys 2 - Rys. 4**), można zauważyć, że w danych nie występują błędy grube. Jednakże na pierwszym z wykonanych wykresów, można dostrzec jedną obserwację odstającą od pozostałych (o tym, czy wspomniana wartość jest wartością ekstremalną zadecydowano w następnym kroku). Wykres liniowy prezentuje przebieg uzyskiwanych wartości izotopu na przestrzeni lat, a wartość odstająca nie wyróżnia się znacząco od innych, szczególnie, że w tym okresie występowało wiele wyższych wartości niż w przypadku obserwacji od około 350 indeksu. Histogram przedstawia rozkład stosunku izotopu tlenu, zbliżony do rozkładu normalnego. Ponadto, na histogramie, można zaobserwować także delikatne zaburzenie - szerszy rozkład z prawej strony informujący o prawostronnej skośności. Na podstawie obliczonych prostych statystyk opisowych (**Rys. 1**), można stwierdzić, że wartości izotopu mieszczą się w przedziale od 3,04 do 5,27, wartość średnia wynosi 4,088, a dane pozbawione są wartości NA.

➢ Sprawdzenie czy w danych występują wartości ekstremalne

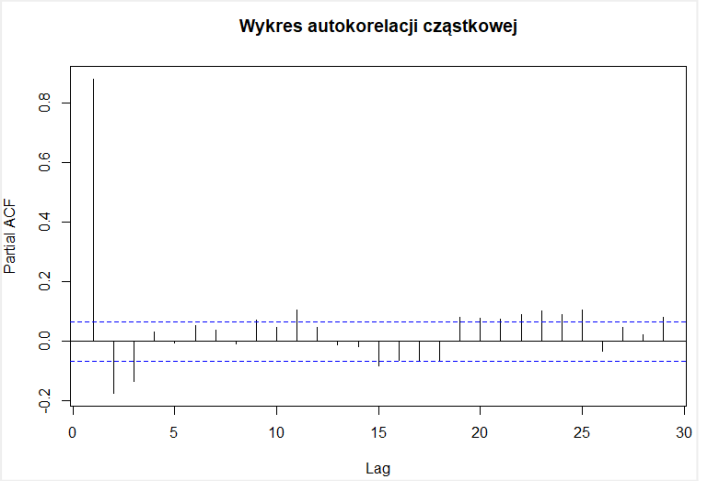
Na podstawie uzyskanych wyników podjęto decyzję, że wykonanie preprocessingu pod kątem czyszczenia danych nie jest konieczne. W danych nie występują błędy grube, natomiast wartość odstająca okazała się nie być wartością ekstremalną, na podstawie wykonanych obliczeń (reguła 3 sigm), w związku z czym nie została zastąpiona wartością NA, a dane uznano za czyste.

➢ Sprawdzenie występowania składowych szeregu czasowego

Z uwagi na fakt, iż szereg czasowy posiada informację o kroku czasowym, a poszczególne obserwacje są ze sobą powiązane, wykonano analizę jego składowych: trendu i sezonowości. W tym celu wygenerowano wykresy autokorelacji ACF i autokorelacji cząstkowej PACF. Otrzymane rezultaty zaprezentowano na **Rys. 5** i **Rys. 6**.



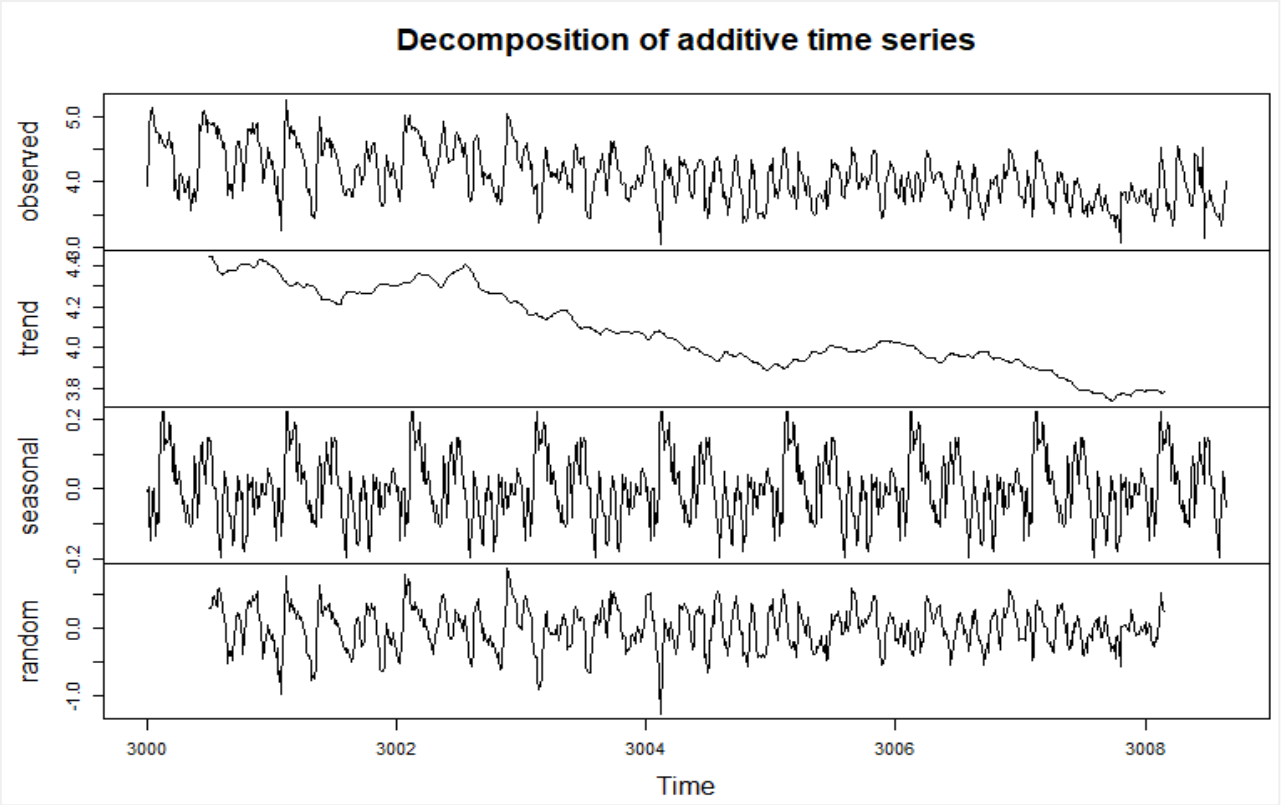
Rys. 5 *Wykres autokorelacji.*



Rys. 6 *Wykres autokorelacji.*

Na podstawie wykresu autokorelacji (**Rys. 5**), można stwierdzić, że w danych występuje zarówno trend, jak i sezonowość. Charakterystyczny wachlarz - wolno malejąca korelacja dla pierwszych opóźnień świadczy o występowaniu trendu, natomiast sinusoida - o sezonowości. Wysoko wysunięte ponad przedział ufności, pierwsze opóźnienie na wykresie autokorelacji cząstkowej (**Rys. 6**), oznacza mocną korelację wskazującą na silny trend; natomiast pierwsze dodatnie, wyraźnie istotne statystycznie opóźnienie, świadczące o sezonowości, to opóźnienie 11. Ponadto, w późniejszych istotnie statystycznych opóźnieniach, można zauważyć przeciek widma do sąsiednich opóźnień - charakterystyczne rozmycie wskazujące na podobieństwo kilku obserwacji w sąsiedztwie.

Ponadto, w celu wizualizacji składowych szeregu wykonano jego dekompozycję, a wyniki zilustrowano poniżej (**Rys. 7**).



***Rys.** 7 Dekompozycja szeregu czasowego na składowe.*

Geologiczne zmiany czasu w tlenie i inne pomiary izotopów z głębokich rdzeni oceanicznych odnoszą się do wzorców zmienności globalnej objętości lodu i temperatury oceanu (Shackleton i Hall 1989; Park i Maasch 1993). Widoczny na powyższym wykresie (**Rys. 7**) trend rosnący oznacza wyraźny wzrost poziomów izotopu tlenu w czasach współczesnych (odwrócona skala czasu), co odzwierciedla wzrost średniej temperatury na świecie oraz mniejsze średnie masy lodu. Sezonowość jest natomiast związana z precesją (krótsze cykle) i nachyleniem orbity Ziemi (dominujący, tzw. 100-tysięczny cykl epoki lodowcowej).

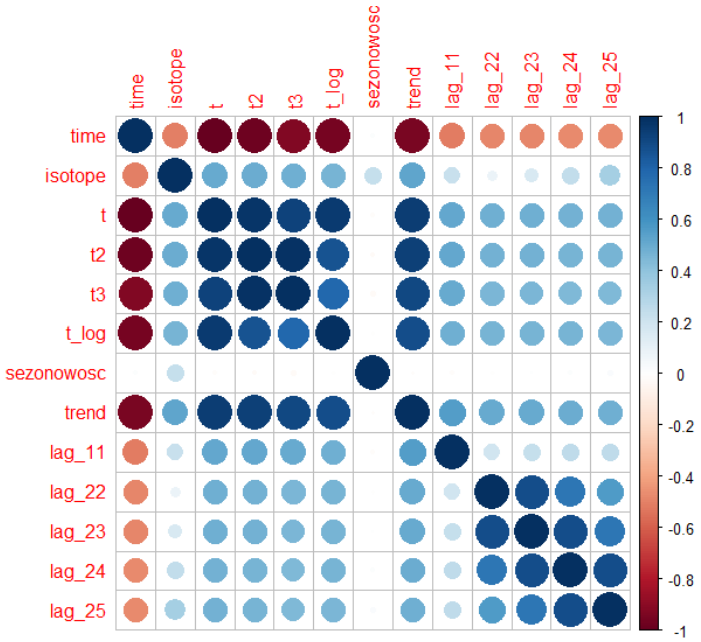
W kolejnym kroku dodano nowe zmienne, do już istniejącej ramki danych, odzwierciedlające

- ➢ trend liniowy (t - odwrócona sekwencja indeksów tj. od 866 do 1),
- ➢ paraboliczny (wektor t podniesiony do kwadratu),
- ➢ hiperboliczny (wektor t podniesiony do sześcianu)
- ➢ i logarytmiczny (logarytm z t).

Dodano również zmienną opóźnioną o 11 okresów, a także zmienne opóźnione o 22, 23, 24 i 25 okresów, wskazujące na przeciek widma.

➢ Korelacja

Sprawdzono korelację między zmiennymi opóźnionymi, a wartościami izotopu. Stosunkowo niskie wartości korelacji liniowej Pearsona mogą oznaczać, iż zmienne słabo pokażą kształt sezonowości (najwyższą wartość korelacji uzyskano dla opóźnienia 25 tj. 0.3101872). Wykonano również macierz korelacji (**Rys. 8**), na podstawie której można zauważyć, że wszystkie dodane zmienne odzwierciedlające trend dobrze korelują z samym trendem.



Rys. 8 *Macierz korelacji.*

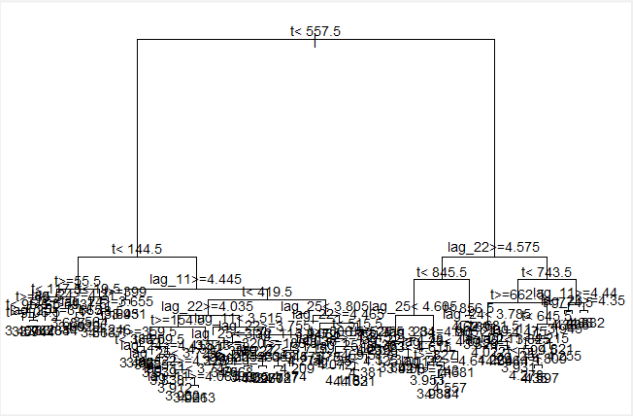
➢ Wybór i uzasadnienie modelu

Założono, że pomiędzy zmienną objaśnianą (zawartość izotopu tlenu), a zmiennymi objaśniającymi (utworzone zmienne) występują zależności. Przewidywanie wartości izotopu tlenu na podstawie utworzonych zmiennych umożliwiły modele regresyjne, stąd do zamodelowania wartości izotopu wykorzystano dwa, ze wspomnianej grupy modeli: model drzewa regresyjnego (CART - wylosowany model) oraz model regresji wielorakiej (wybrany model). W tym celu zastosowano różne postacie trendu i wyznaczone opóźnienia, do przewidywania wartości zmiennej zależnej.

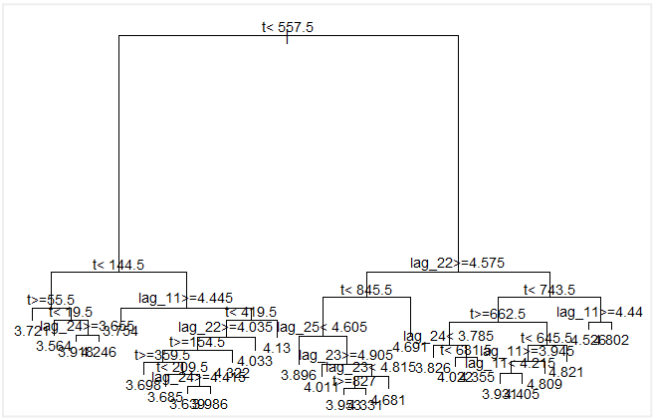
2. MODEL CART

➢ Wykonanie modelu

Na początku dane podzielono na zbiór uczący i testowy - do zbioru uczącego trafiły 862 obserwacje, natomiast do zbioru testowego, ostatnie 4 obserwacje. Jako zmienne do modelu zostały wybrane wszystkie postacie trendu oraz wykorzystano jednakże wszystkie utworzone zmienne opóźnione. Wygenerowano wykres drzewa przed przycięciem, a otrzymany wynik przedstawiono na **Rys. 9**, oraz po przycięciu - **Rys. 10**.

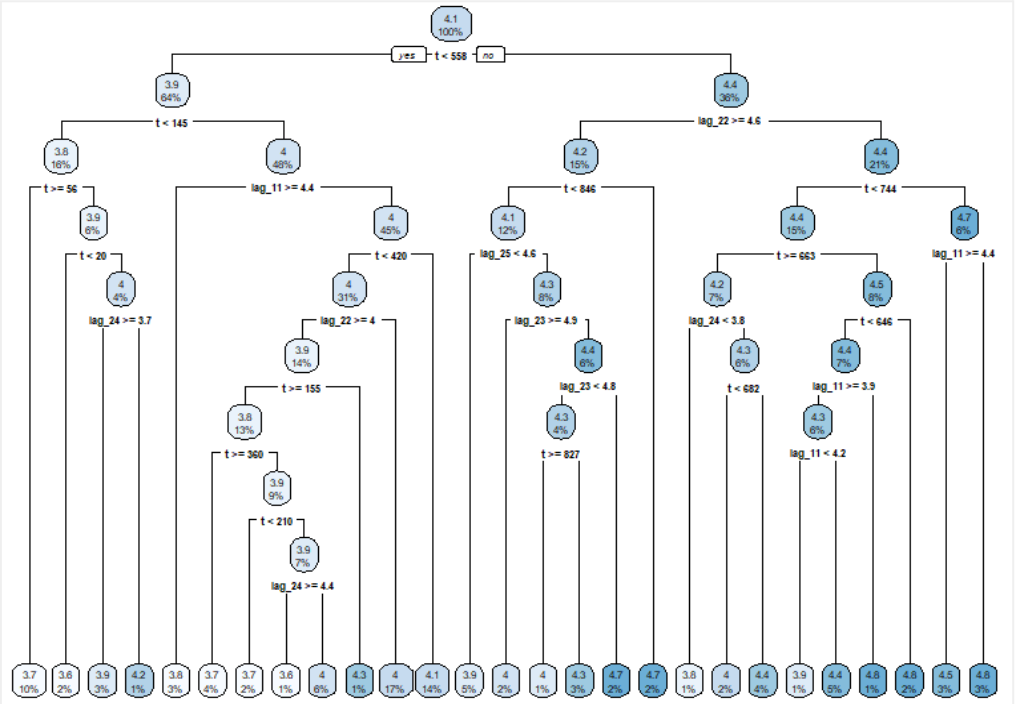


Rys. 9 *Wykres drzewa przed przycięciem.*



Rys. 10 *Wykres drzewa po przycięciu.*

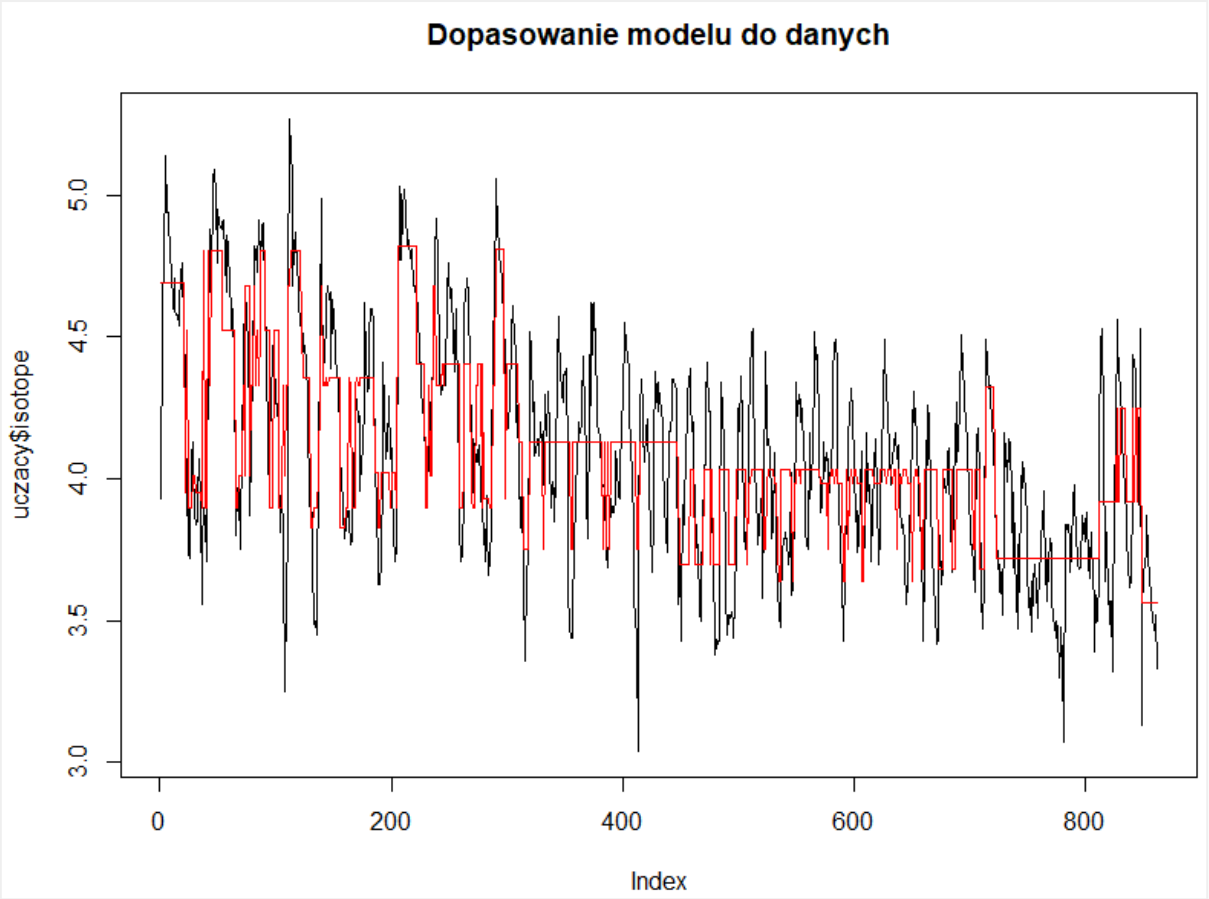
Dodatkowo wykres drzewa zaprezentowano na ilustracji poniżej (**Rys. 11**). Otrzymane drzewo jest dosyć rozbudowane, jednak takie przycięcie pozwoliło uzyskać lepszą jakość modelu niż w przypadku większego przycięcia drzewa.



Rys. 11 *Wykres otrzymanego drzewa.*

➢ Ocena jakości modelu

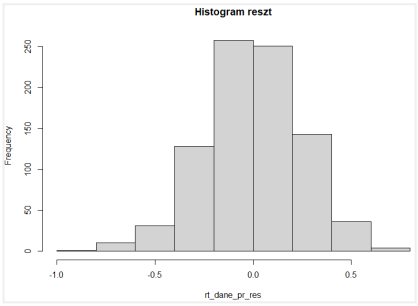
Na podstawie otrzymanej wartości współczynnika determinacji R^2, sprawdzono, czy uzyskany model ma wystarczająco dobrą jakość. Obliczona wartość R^2 (równa 0.6252548), świadczy o tym, iż około 63% zmienności danych rzeczywistych zostało wytłumaczonych za pomocą modelu. Błąd standardowy estymacji wyniósł natomiast 2.503794e-16, co oznacza że model średnio myli się o stosunkowo niewiele. W celu wizualizacji dopasowania modelu do danych rzeczywistych wykonano poniższy wykres (**Rys. 12**). Na jego podstawie można zaobserwować, iż w większości model dobrze przewiduje rzeczywiste wartości, poza najniższymi wartościami, szczególnie dla obserwacji o indeksie między 760 a 800.



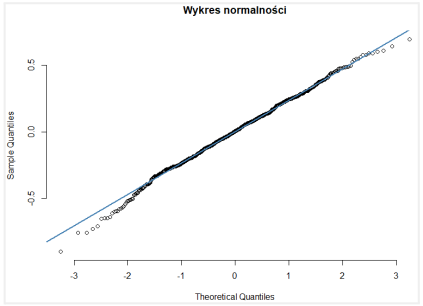
Rys. 12 *Dopasowanie modelu do danych rzeczywistych.*

➢ Ocena jakości reszt

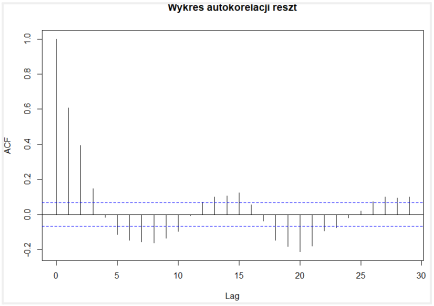
W celu oceny jakości reszt, został wykonany histogram reszt, wykres autokorelacji i autokorelacji cząstkowej oraz wykres rozrzutu. Otrzymane rezultaty zilustrowano na **Rys. 13 - Rys. 15**.



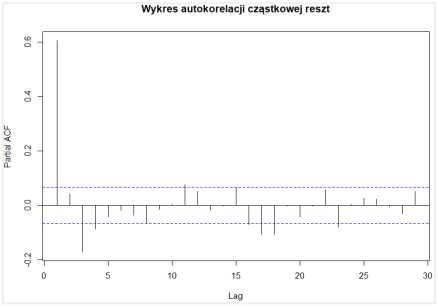
Rys. 13 *Histogram reszt.*



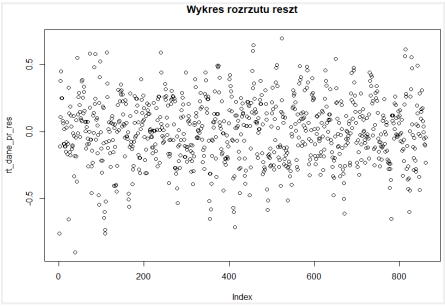
Rys. 14 *Wykres normalności.*



Rys. 15 *Wykres autokorelacji.*



Rys. 16 *Wykres autokorelacji cząstkowej.*



Rys. 17 *Wykres rozrzutu reszt.*

Analizując powyższe wykresy (**Rys. 13 - Rys. 16**), można stwierdzić, iż reszty spełniają założenia procesu białoszumowego. W resztach pozostała sama składowa szumowa, co potwierdza histogram oraz wykres normalności (**Rys. 13**, **Rys. 14**), na podstawie których można zauważyć, że reszty posiadają rozkład bardzo zbliżony do rozkładu Gaussa (na histogramie widoczny charakterystyczny dzwonowaty kształt, na wykresie normalności - zdecydowana większość punktów leży na linii teoretycznego rozkładu). Na podstawie wykresu autokorelacji i autokorelacji cząstkowej (**Rys. 15**, **Rys. 16**), można zaobserwować, iż między kolejnymi opóźnieniami nie występują zależności, opóźnienia nie są istotne statystycznie (**Rys. 16,** poza jednym nieznacznie wysuwającym się ponad przedział ufności). Na ostatnim ze wspomnianych wykresów (**Rys. 17**), można zauważyć, iż reszty mają charakter zupełnie losowy - widoczna bezładna chmura punktów. Na podstawie przedstawionych wizualizacji, można stwierdzić, że model został prawidłowo dopasowany do danych, zatem poprawnie wytłumaczono zmienność danych rzeczywistych, które zostały zamodelowane.

➢ Ocena jakości prognoz

Ocenę jakości prognoz dokonano na podstawie wyliczonej wartości MAPE, która wyniosła 0.062148. Jest to wartość zadowalająca, średnio w okresie predykcji prognozy odchylają się o 6%. Obliczony został także współczynnik Theila, który wyniósł wartość równą 0, świadczącą o idealnie trafnych prognozach.

3. MODEL REGRESJI WIELORAKIEJ

➢ Wykonanie modelu

W przypadku drugiego, wybranego modelu, dane również podzielono na zbiór uczący i testowy. Analogicznie jak w poprzednim modelu, do zbioru uczącego trafiły 862 obserwacje, natomiast do zbioru testowego, ostatnie 4 obserwacje. Zmiennymi, za pomocą których zamodelowano wartości izotopu, były te same zmienne objaśniające, które użyto w modelu drzewa regresyjnego.

➢ Ocena jakości modelu

Na podstawie przedstawionego poniżej **Rys. 18** i **Rys. 19**, została oceniona jakość wykonanego modelu.

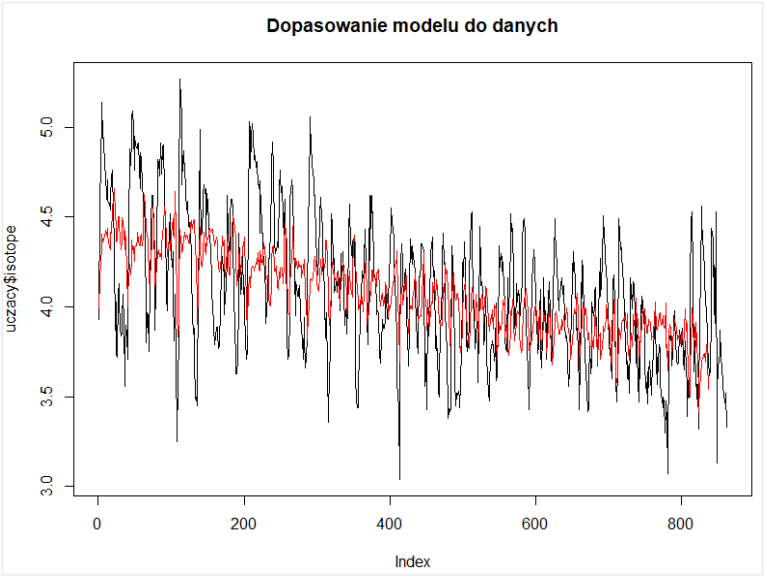
```
Call:
lm(formula = isotope ~ t + t2 + t3 + t_log + lag_11 + lag_22 +
    lag_23 + lag_24 + lag_25, data = uczacy)

Residuals:
     Min       1Q   Median       3Q      Max
-1.06864 -0.21368  0.00575  0.22236  0.92459

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.117e+00  2.976e-01  13.832  < 2e-16 ***
t           -2.021e-03  1.202e-03  -1.681   0.0931 .
t2           5.587e-06  2.286e-06   2.444   0.0147 *
t3          -3.481e-09  1.516e-09  -2.296   0.0219 *
t_log        1.339e-01  6.960e-02   1.924   0.0546 .
lag_11      -7.170e-02  3.282e-02  -2.184   0.0292 *
lag_22      -3.493e-01  6.502e-02  -5.372 1.01e-07 ***
lag_23       2.284e-02  9.314e-02   0.245   0.8063
lag_24      -6.709e-02  9.305e-02  -0.721   0.4711
lag_25       2.785e-01  6.445e-02   4.321 1.74e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3189 on 827 degrees of freedom
  (25 obserwacji zostało skasowanych z uwagi na braki w nich zawarte)
Multiple R-squared:  0.3285,    Adjusted R-squared:  0.3212
F-statistic: 44.96 on 9 and 827 DF,  p-value: < 2.2e-16
```

Rys. 18 Ocena jakości modelu regresji wielorakiej.

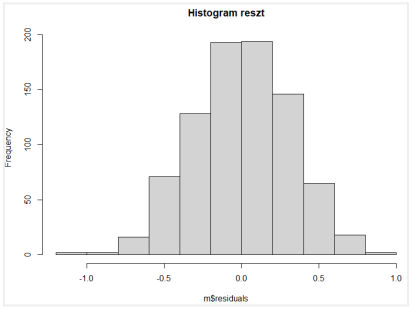


Rys. 19 Dopasowanie modelu do danych.

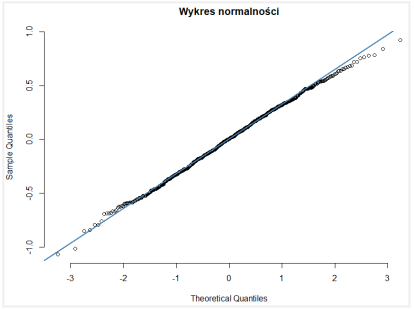
Analizując dane zgromadzone powyżej (**Rys. 18**), można stwierdzić, że otrzymany model jest istotny statystycznie (p-value < 0.05), mimo że nie wszystkie zmienne użyte w modelu okazały się istotne statystycznie. Takimi zmiennymi były opóźnienie 23 oraz 24. Zmiennymi na granicy istotności były natomiast dwie zmienne przedstawiające trend tj. odzwierciedlające trend liniowy i zlogarytmowany. Najbardziej istotne w modelu okazały się opóźnienie 22 i 25. Jednakże, wartość współczynnika determinacji, wynosząca ok. 33%, jest znacznie niższa niż wartość otrzymana w przypadku pierwszego wykonanego modelu (**punkt 2**). Dodatkowo błąd standardowy estymacji, także jest znacznie wyższy (0.3189) niż dla modelu CART. Dopasowanie modelu do danych rzeczywistych przedstawiono na powyższym wykresie (**Rys. 19**), na podstawie którego można zaobserwować, iż model słabiej radzi sobie z przewidywaniem wyższych wartości, szczególnie w okresie dla pierwszych 300 obserwacji, i z najniższymi wartościami - w okolicach 800 obserwacji.

➢ Ocena jakości reszt

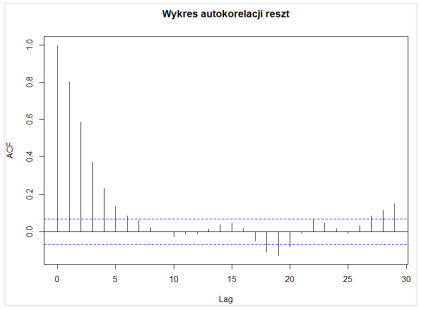
Ocenę jakości reszt wykonano analogicznie jak w poprzednim przypadku. W tym celu wygenerowane zostały poniższe wykresy (**Rys. 20** - **Rys. 24**).



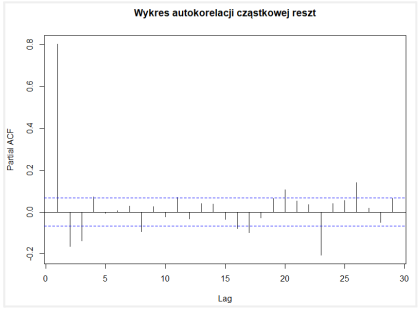
Rys. 20 *Histogram reszt.*



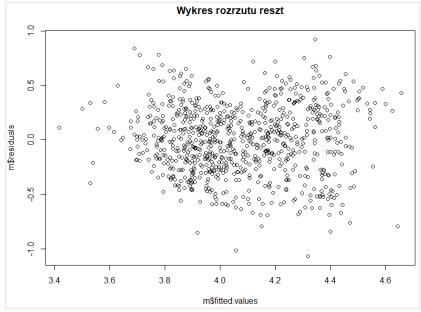
Rys. 21 *Wykres normalności.*



Rys. 22 *Wykres autokorelacji.*



Rys. 23 *Wykres autokorelacji cząstkowej.*



Rys. 24 *Wykres rozrzutu reszt.*

Interpretując wyniki otrzymane na powyższych wykresach, można stwierdzić, na podstawie histogramu oraz wykresu normalności (**Rys. 20**, **Rys. 21**), że reszty mają rozkład zbliżony do rozkładu normalnego. Na wykresie autokorelacji i autokorelacji cząstkowej (**Rys. 22**, **Rys. 23**), można zauważyć niewielkie zależności, niedużo wysunięte ponad przedział ufności dwa dalsze opóźnienia. Wykres rozrzutu reszt nie przedstawia konkretnego kształtu, ale punkty są bardziej skupione niż na wykresie rozrzutu reszt (**Rys. 17**) dla poprzedniego modelu. Ostatecznie uzyskane reszty można uznać za spełniające założenia procesu białoszumowego.

➢ Ocena jakości prognoz

Aby ocenić jakość prognoz, został obliczony MAPE oraz współczynnik Theila, które odpowiednio wyniosły 0.1061917 i 0.0003338739. Wartość MAPE nieco wyższa niż w poprzednim modelu, ale nieznacznie przekraczająca 10% przez co jakość prognoz można uznać za zadowalającą. Potwierdza to również uzyskana wartość współczynnika Theila bliska zeru.

4. WNIOSKI

Zastosowane metody pozwoliły na zamodelowanie wartości izotopu tlenu od dodanych zmiennych objaśniających. Na podstawie wykonanych modeli oraz ich oceny można stwierdzić, że lepszym jakościowo modelem okazał się wylosowany model drzewa regresyjnego CART. Zarówno pod względem oceny samego modelu, reszt oraz prognoz, model drzewa regresyjnego wykazał się lepszymi rezultatami. Mimo gorszych wyników, model regresji wielorakiej również okazał się istotny statystycznie, w związku z czym oba modele można by wdrożyć do przemysłu, wykorzystać do prognozowania, czy wyjaśniania związków, przy czym model CART sprawdzi się znacznie lepiej dla tego typu danych.

5. LITERATURA

➢ https://bayanbox.ir/view/5561099385628144678/Bayesian-forecasting-and-dynamic-models-West-Harison.pdf