Zaawansowane elementy
stosowania pakietów statystycznych

Kamila Bogusz 126083
Wojciech Rybacki 126084
Karol Korniat 123453

1.Model ekonometryczny

Do modelu za zmienna objaśnianą wybrano liczbę osób bezrobotnych w Polsce zarejestrowanych w latach 2015-2023 (*Rys.1*). Natomiast za zmienne objaśniające przyjęliśmy:

- X1: Średnie zarobki
- X2: Średnią temperaturę powietrza
- X3: Stopę oprocentowania
- X4: Saldo obrotu towarów (różnica pomiędzy wartością eksportu,a wartością importu)
- X5: Przeciętne zatrudnienie w sekcji przedsiębiorstwa

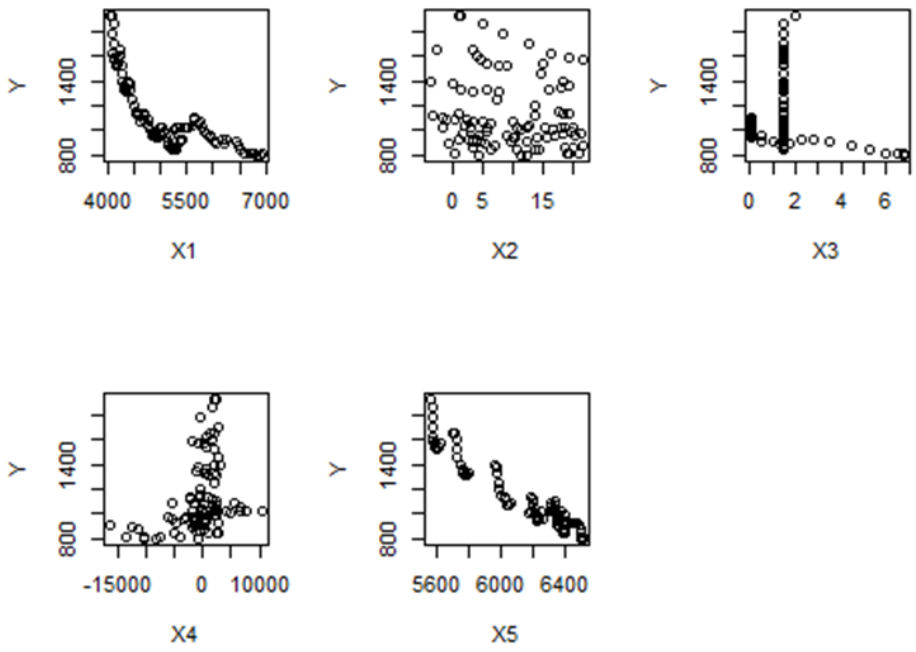
Pierwszym etapem był podział naszych danych na dane testowe oraz zestaw uczący. Dane testowe stanowiły ostatni rok zakresu naszych danych natomiast zestaw uczący - pierwsze 7 lat.

```
##         stopa     rok   miesiac   sr_zarob   sr.temp   stopa_oprocentowania   saldo_obrotow_tow     pzws
## 91  810.2  2022          7   6794.47    19.10                        6.50               -7414.6   6508
## 92  806.9  2022          8   6700.36    20.80                        6.50              -10243.1   6503
## 93  801.7  2022          9   6809.83    12.40                        6.75              -10389.0   6494
## 94  796.0  2022         10   6812.62    11.31                        6.75               -8322.1   6501
## 95  800.2  2022         11   6903.92     4.36                        6.75                -823.3   6507
## 96 812.3 2022      12  6936.27    0.22                 6.75          -13579.9  6505
```

Rys.1 Przykładowe wartości z danych testowych (stopa oznacza liczbę osób bezrobotnych).

1.1 Wykresy rozrzutu

Po podzieleniu danych wykonaliśmy wykresy, które miały na celu wstępne oszacowanie czy dane mogą być problematyczne. Dla modelowania ekonometrycznego, takie wykresy są pomocne w identyfikacji rodzaju zależności między zmiennymi.



Rys.2 Wykresy rozrzutów dla zmiennych objaśniających: średnich zarobków (X1), średniej temperatury powietrza (X2), stopy oprocentowania (X3), salda obrotu towarów (X4), przeciętnego zatrudnienia w sekcji przedsiębiorstwa (X5).

Analiza wykresów (*Rys.2*)**:**

- X1vsY: Wykres wskazuje na nieliniową, potencjalnie ekspotencjalną lub hiperboliczną zależność między X1 a Y. Wartości Y spadają gwałtownie przy niższych wartościach X1, a spadek wydaje się zwalniać przy wyższych wartościach X1.

-X2vsY: Wykres nie pokazuje wyraźnej zależności między X2 a Y. Punkty są rozrzucone i nie wykazują jednoznacznego trendu, co może sugerować słabą lub brak liniowej korelacji.

-X3vsY: Podobnie jak przy X2, punkty są dość równomiernie rozrzucone wzdłuż osi X3, co sugeruje, że nie ma silnej liniowej zależności między X3 a Y.

-X4vsY: Na tym wykresie punkty są rozrzucone w taki sposób, że nie widać wyraźnej zależności między X4 a Y. Widoczne są dwa skupienia punktów, co może wskazywać, że inna zmienna lub wpływa na tę relację w sposób, który nie został uchwycony przez proste porównanie tych dwóch zmiennych.

-X5vsY: Istnieje wyraźna nieliniowa zależność między X5 a Y. Zależność może być negatywnie nieliniowa lub może przybierać postać wykładniczego spadku.

1.2 Korelacje

Kolejnym etapem było sprawdzenie korelacji pomiędzy zmiennymi objaśniającymi (*Rys.4*) oraz między zmiennymi objaśniającymi, a zmienną objaśnianą (*Rys.3*). Oczekuje się, że wartości między zmienną objaśnianą, a zmiennymi objaśniającymi będą oddalone jak najbardziej od zera (bliskie -1 lub 1) co świadczy o mocnej korelacji zmiennych. Zjawisko to jest najlepiej zauważalne dla korelacji Y i X5 oraz Y i X1 (*Rys.3*). Natomiast wśród korelacji między zmiennymi objaśniającymi spodziewa się wartości bliskich zeru, które będą świadczyły o braku korelacji. Niestety zauważalna jest wysoka korelacja wynosząca 0.86 pomiędzy zmienną X1 a zmienną X5 (*Rys.4*). Może to świadczyć o autokorelacji w danych.

```
##                cor
## Y i X1: 0.00 -0.79
## Y i X2: 0.14 -0.15
## Y i X3: 0.02 -0.24
## Y i X4: 0.00  0.34
## Y i X5: 0.00 -0.95
```

Rys.3. Wartości korelacji między zmienną objaśnianą oraz zmiennymi objaśniającymi.

```
##                 cor
## X1 i X2: 0.81  0.02
## X1 i X3: 0.00  0.49
## X1 i X4: 0.00 -0.49
## X1 i X5  0.00  0.86
## X2 i X3: 0.82  0.02
## X2 i X4: 0.80  0.03
## X2 i X5: 0.95  0.01
## X3 i X4: 0.00 -0.67
## X3 i X5  0.01  0.25
## X4 i X5  0.00 -0.33
```

Rys.4 Wartości korelacji między zmiennymi objaśniającymi.

1.3 Współczynnik zmienności

W tym etapie sprawdziliśmy zmienności naszych zmiennych, aby zapobiec sytuacji, gdzie wartości mogą być praktycznie stałe. Niska wartość współczynnika zmienności oznacza, że wartości tej zmiennej w bardzo małym stopniu różnicują co wpływać będzie na słabe dopasowanie modelu. Zmienna X5 wykazała zmienność wynoszącą jedynie ~ 5% (*Rys.5*). Oznacza to, że jest bliska wartości stałej. Dlatego nie będziemy jej uwzględniać w modelu, ponieważ mogłaby zaburzyć poprawność predykcji. Tym samym rozwiązaliśmy problem z wysoką korelacją pomiędzy zmiennymi X1 i X5.

```
## Zmienna X1:  0.1591279

## Zmienna X2:  0.7600385

## Zmienna X3:  0.927567

## Zmienna X4:  -11.00709

## Zmienna X5:  0.04996765
```

Rys.5 Wyliczone wartości współczynników zmienności dla poszczególnych zmiennych objaśniających.

1.4 Tworzenie modeli

Do wybrania poszczególnych zmiennych objaśniających wykorzystano takie metody jak: metoda pojemności informacyjnej, metoda grafowa, metodę automatycznej funkcji doboru zmiennych, metodę analizy współczynników korelacji. Do dalszych analiz otrzymaliśmy cztery modele, gdzie model 2 był tożsamy z modelem 4:

- Model1: *y = 2593.2182 - 0.277X1 - 5.2023X2 + e*
- Model2: *y = 2549.1476 - 0.2782X1 + e*
- Model3: *y = 2706.728 - 0.311X1 - 5.287X2 + 35.181X3 + e*
- Model5: *y = 7089.329 +0.0687X1 - 5.948X2 + 0.005X4 -1.018X5 + e*

```
##       [,1]
## [1,] -0.79
## [2,] -0.15
## [3,] -0.24
## [4,]  0.34

##       [,1] [,2]  [,3]  [,4]
## [1,]  1.00 0.02  0.49 -0.49
## [2,]  0.02 1.00  0.02  0.03
## [3,]  0.49 0.02  1.00 -0.67
## [4,] -0.49 0.03 -0.67  1.00
```

Rys.6 Przedstawienie dwóch macierzy. Macierz zawierająca wartości korelacji między zmienną objaśnianą, a zmiennymi objaśniającymi (góra). Macierz zawierająca wartości korelacji między zmiennymi objaśniającymi (dół).

1.4.1 Metoda pojemności informacyjnej

Jest to metoda, która pozwala na wyznaczenie pojemności integralnej kombinacji nośników informacji na podstawie indywidualnych pojemności nośników. Do ich wyliczenia wykorzystano wcześniej utworzone macierze (*Rys.6*). Po ich wyznaczeniu wybierana jest kombinacja, dla której pojemność informacji (h) jest największa. Z obliczeń wynika, że największą wartość równą 0.63 (*Rys.7*) posiada h12 czyli kombinacja zmiennych X1 i X2. Możemy dzięki temu stworzyć model o postaci: *y= a0 + a1X1 + a2X2 + e*. Zmienne są istotne statystycznie (*Rys.8*), dlatego mogą zostać wykorzystane do tworzenia finalnego modelu. Model pierwszy przyjmuje formę *y = 2593.2182 - 0.277X1 - 5.2023X2*.

```
##             [,1]
## h1    0.62410000
## h2    0.02250000
## h3    0.05760000
## h4    0.11560000
## h12   0.63392157
## h13   0.45751678
## h14   0.49644295
## h23   0.07852941
## h24   0.13407767
## h34   0.10371257
## h123  0.46635762
## h124  0.51079246
## h234  0.12351141
## h1234 0.41228544
```

Rys.7 Wartości indywidualnych pojemności nośników.

```
## Call:
## lm(formula = Y ~ X1 + X2, data = dane_ucz)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -247.82 -176.21   10.38  118.87  453.16
##
## Coefficients:
##              Estimate Std. Error t value            Pr(>|t|)
## (Intercept) 2593.2182   115.8877  22.377 <0.0000000000000002 ***
## X1            -0.2770     0.0221 -12.534 <0.0000000000000002 ***
## X2            -5.2023     2.4603  -2.115              0.0371 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 174.6 on 93 degrees of freedom
## Multiple R-squared:  0.6367, Adjusted R-squared:  0.6289
## F-statistic: 81.48 on 2 and 93 DF,  p-value: < 0.00000000000000022
```

Rys.8 Informacje o modelu *y= a0 + a1X1 + a2X2 + e* uzyskanego na podstawie najwyższej wartości indywidualnych nośników (h12).

1.4.2 Metoda grafowa

Przy zastosowaniu tej metody wybraliśmy do modelu zmienną z największą liczbą wiązań z najbardziej znaczącą korelacją ze zmienną objaśnianą. Otrzymany model przyjął postać *y = a0 + a1X1 + e* , gdzie X1 jest istotny statycznie (*Rys.9*). Ostateczna forma drugiego modelu *y = 2549.1476 - 0.2782X1 + e*.

```
##
## Call:
## lm(formula = Y ~ X1, data = dane_ucz)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -249.28 -151.41    5.77  121.97  497.47
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept) 2549.1476   116.0840   21.96 <0.0000000000000002 ***
## X1            -0.2782     0.0225  -12.36 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 177.8 on 94 degrees of freedom
## Multiple R-squared:  0.6192, Adjusted R-squared:  0.6152
## F-statistic: 152.9 on 1 and 94 DF,  p-value: < 0.00000000000000022
```

Rys.9 Informacje o modelu *y= a0 + a1X1 + e* uzyskanego na podstawie metody grafowej.

1.4.3 Metoda automatycznej funkcji doboru zmiennych

Przy pomocy funkcji step() porównano wartości AIC przed jak i po wykluczeniu każdej ze zmiennych. Im mniejsza wartość AIC tym świadczy to, że dany model jest lepiej dopasowany. Po pierwszym etapie funkcja wykluczyła zmienną X4, jako że jej odjęcie obniżyło wartość AIC z 989 na 988 (*Rys.10*). Kolejny etap pokazuje, że dalsze odejmowanie zmiennych nie ma sensu. Otrzymano model *y = a0 + a1X1 + a2X2 + a3X3 + e,* który następnie poddano analizie. Wykazano, że każda z trzech zmiennych objaśniających jest istotna statystycznie (*Rys.11*). Model trzeci przyjmuję postać *y = 2706.728 - 0.311X1 - 5.287X2 + 35.181X3 + e.*

```
## Start:  AIC=989.26
## Y ~ X1 + X2 + X3 + X4
##
##        Df Sum of Sq     RSS     AIC
## - X4    1     25445 2609826  988.20
## <none>              2584380  989.26
## - X2    1    147654 2732034  992.60
## - X3    1    233900 2818281  995.58
## - X1    1   4122123 6706503 1078.81
##
## Step:  AIC=988.2
## Y ~ X1 + X2 + X3
##
##        Df Sum of Sq     RSS     AIC
## <none>              2609826  988.20
## - X2    1    140825 2750651  991.25
## - X3    1    226439 2836264  994.19
## - X1    1   4580838 7190664 1083.50
```

Rys.10 Dobór zmiennych objaśniających na podstawie wartości AIC uzyskanych z funkcji step() na podstawie modelu z czterema zmiennymi objaśniającymi (X1,X2,X3,X4).

```
## Call:
## lm(formula = Y ~ X1 + X2 + X3, data = dane_ucz)
##
## Coefficients:
## (Intercept)           X1           X2           X3
##    2706.728       -0.311       -5.287       35.181
## Call:
## lm(formula = Y ~ X1 + X2 + X3, data = dane_ucz)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -238.68 -166.80   28.54  108.99  407.13
##
## Coefficients:
##               Estimate Std. Error t value             Pr(>|t|)
## (Intercept) 2706.72761  118.76957  22.790 < 0.0000000000000002 ***
## X1            -0.31102    0.02448 -12.708 < 0.0000000000000002 ***
## X2            -5.28723    2.37301  -2.228              0.02831 *
## X3            35.18096   12.45215   2.825              0.00579 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 168.4 on 92 degrees of freedom
## Multiple R-squared:  0.6657, Adjusted R-squared:  0.6548
## F-statistic: 61.06 on 3 and 92 DF,  p-value: < 0.00000000000000022
```

Rys.11 Analiza zmiennych objaśniających modelu *y = a0 + a1X1 + a2X2 + a3X3 + e* uzyskanego na podstawie automatycznej funkcji doboru zmiennych

1.4.4 Metoda analizy współczynników korelacji

Za pomocą tej metody uzyskano model tożsamy z modelem drugim, mianowicie *y = a0 + a1X1 +e*. Zmienna objaśniająca była istotna statystycznie (Rys.12). Przyjęto model równy *y = 2549.1476 - 0.2782X1 + e.*

```
## Call:
## lm(formula = Y ~ X1, data = dane_ucz)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -249.28 -151.41    5.77  121.97  497.47
##
## Coefficients:
##               Estimate Std. Error t value            Pr(>|t|)
## (Intercept) 2549.1476   116.0840   21.96 <0.0000000000000002 ***
## X1            -0.2782     0.0225  -12.36 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 177.8 on 94 degrees of freedom
## Multiple R-squared:  0.6192, Adjusted R-squared:  0.6152
## F-statistic: 152.9 on 1 and 94 DF,  p-value: < 0.00000000000000022
```

Rys.11 Przedstawienie parametrów modelu *y = a0 + a1X1 + e* uzyskanego na podstawie analizy współczynników korelacji.

1.4.5 Metoda regresji krokowej

W tej metodzie uwzględniliśmy również X5 który wcześniej został odrzucony na poziomie określania współczynnika zmienności. Analiza modelu z pięcioma zmiennymi wykazała brak istotności zmiennej X3 oraz X4 gdyż wartość p-value wyniosła powyżej 0.05 (*Rys.12*). Jednakże najpierw usunięto zmienną o większej wartości p-value (czyli X3) w celu sprawdzenia czy faktycznie należy także wyeliminować drugą zmienną. Drugi model pokazał, że wszystkie zmienne modelu są istotne statystycznie (*Rys.13*). Uzyskano model o postaci *y = 7089.329 +0.0687X1 - 5.948X2 + 0.005X4 -1.018X5 + e* ma on najwyższy współczynnik determinacji wynoszący 0.926 co świadczy o jego najlepszym dopasowaniu.

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + X4 + X5, data = dane_ucz)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -169.555  -54.002    7.269   47.186  224.720
##
## Coefficients:
##                Estimate  Std. Error t value             Pr(>|t|)
## (Intercept) 7125.199547  252.005028  28.274 < 0.0000000000000002 ***
## X1             0.073166    0.023877   3.064              0.00288 **
## X2            -5.931744    1.105987  -5.363          0.000000628 ***
## X3            -3.327610    7.439636  -0.447              0.65575
## X4             0.004495    0.002553   1.760              0.08176 .
## X5            -1.026669    0.056412 -18.200 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 78.33 on 90 degrees of freedom
## Multiple R-squared:  0.9293, Adjusted R-squared:  0.9253
## F-statistic: 236.5 on 5 and 90 DF,  p-value: < 0.00000000000000022
```

Rys.12 Analiza zmiennych objaśniających na podstawie metody regresji krokowej.

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X4 + X5, data = dane_ucz)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -171.047  -53.419    7.597   47.613  222.392
##
## Coefficients:
##                Estimate  Std. Error t value             Pr(>|t|)
## (Intercept) 7089.329264  237.851581  29.806 < 0.0000000000000002 ***
```

```
## X1             0.068729    0.021623   3.179              0.00202 **
## X2            -5.948104    1.100513  -5.405          0.000000517 ***
## X4             0.005101    0.002154   2.368              0.02003 *
## X5            -1.018009    0.052752 -19.298 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 77.98 on 91 degrees of freedom
## Multiple R-squared:  0.9291, Adjusted R-squared:  0.926
## F-statistic: 298.2 on 4 and 91 DF,  p-value: < 0.00000000000000022
```

Rys.13 Analiza zmiennych objaśniających na podstawie metody regresji krokowej.

1.5 Porównanie modeli ANOVA

Za pomocą analizy wariancji przetestowano cztery modele:

- Model1: *y = 2593.2182 - 0.277X1 - 5.2023X2 + e*
- Model2: *y = 2549.1476 - 0.2782X1 + e*
- Model3: *y = 2706.728 - 0.311X1 - 5.287X2 + 35.181X3 + e*
- Model5: *y = 7089.329 +0.0687X1 - 5.948X2 + 0.005X4 -1.018X5 + e*

Porównując wartości p-value (*Rys.14*) wywnioskowano, że istotnie lepszy jest model5.

```
## Analysis of Variance Table
##
## Model 1: Y ~ X1 + X2
## Model 2: Y ~ X1
##   Res.Df     RSS Df Sum of Sq      F  Pr(>F)
## 1     93 2836264
## 2     94 2972625 -1   -136361 4.4712 0.03715 *
## ---
## Signif. Codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ‘ 1

## Analysis of Variance Table
##
## Model 1: Y ~ X1 + X2
## Model 3: Y ~ X1 + X2 + X3
##   Res.Df     RSS Df Sum of Sq      F   Pr(>F)
## 1     93 2836264
## 3     92 2609826  1    226439 7.9823 0.005793 **
## ---
## Signif. Codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ‘ 1

## Analysis of Variance Table
##
## Model 1: Y ~ X1 + X2
## Model 5: Y ~ X1 + X2 + X4 + X5
##   Res.Df     RSS Df Sum of Sq      F               Pr(>F)
## 1     93 2836264
## 5     91  553416  2   2282849 187.69 < 0.00000000000000022 ***
```

```
## ---
## Signif. Codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

## Analysis of Variance Table
##
## Model 2: Y ~ X1
## Model 3: Y ~ X1 + X2 + X3
##   Res.Df     RSS Df Sum of Sq      F  Pr(>F)
## 2     94 2972625
## 3     92 2609826  2    362799 6.3946 0.00251 **
## ---
## Signif. Codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

## Analysis of Variance Table
##
## Model 2: Y ~ X1
## Model 5: Y ~ X1 + X2 + X4 + X5
##   Res.Df     RSS Df Sum of Sq     F                Pr(>F)
## 2     94 2972625
## 5     91  553416  3   2419209 132.6 < 0.00000000000000022 ***
## ---
## Signif. Codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

## Analysis of Variance Table
##
## Model 3: Y ~ X1 + X2 + X3
## Model 5: Y ~ X1 + X2 + X4 + X5
##   Res.Df     RSS Df Sum of Sq      F                Pr(>F)
## 3     92 2609826
## 5     91  553416  1   2056410 338.14 < 0.00000000000000022 ***
## ---
## Signif. Codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Rys.14 Analiza wariancji model1, model2, model3, model5.

1.6 Badanie właściwości składników losowych

Modele ekonometryczne opierają się na założeniach dotyczących danych takich jak: symetria, losowość, homoskedastyczność, brak autokorelacji, rozkład normalny. Niestety wybrane dane nie spełniły żadnego z warunków. Jedynie w modelu5 wykryto rozkład normalny, ale mimo to reszta warunków nie została spełniona.

1.6.1 Rozkład normalny

Hipotezy Shapiro-Wilk:

H_0: Dane mają rozkład normalny

H_a: Dane nie mają rozkładu normalnego

Analiza reszt wykazała, że p-value wyniosło wartość poniżej 0.05 (*Rys.15*) w model1, model2, oraz model3 co oznacza, że odrzucamy H0. Reszty wykazały rozkład normalny jedynie w model5.

```
##
##  Shapiro-Wilk normality test
##
## data:  model1$residuals
## W = 0.94286, p-value = 0.0003922

##
##  Shapiro-Wilk normality test
##
## data:  model2$residuals
## W = 0.94402, p-value = 0.0004614

##
##  Shapiro-Wilk normality test
##
## data:  model3$residuals
## W = 0.93919, p-value = 0.0002359

##
##  Shapiro-Wilk normality test
##
## data:  model5$residuals
## W = 0.97659, p-value = 0.08332
```

Rys.15 Test Shapiro-Wilka dla reszt model1, model2, model3, model5.

1.6.2 Autokorelacja

Hipotezy test Durbin-Watson:

H_0: Nie ma autokorelacji reszt, czyli autokorelacja reszt pierwszego rzędu jest równa zero (ρ=0).
H_a: Istnieje autokorelacja reszt, czyli autokorelacja reszt pierwszego rzędu jest różna od zera

We wszystkich analizowanych modelach wykryto autokorelacje dodatnią (wartość DW bliska zeru), odrzucono hipotezę zerową (*Rys.16*). Zastosowano przekształcenie Cochrana-Orcutta i powtórzono test Durbina-Watsona aby sprawdzić czy autokorelacja reszt zostanie wyeliminowana. Po przekształceniu p-value nadal jest poniżej 0.05 (*Rys.17*) co świadczy o autokorelacji reszt modeli.

```
##
##  Durbin-Watson test
##
## data:  model1
```

```
## DW = 0.033168, p-value < 0.00000000000000022
## alternative hypothesis: true autocorrelation is not 0

##
##  Durbin-Watson test
##
## data:  model2
## DW = 0.038272, p-value < 0.00000000000000022
## alternative hypothesis: true autocorrelation is not 0

##
##  Durbin-Watson test
##
## data:  model3
## DW = 0.037505, p-value < 0.00000000000000022
## alternative hypothesis: true autocorrelation is not 0

##
##  Durbin-Watson test
##
## data:  model5
## DW = 0.41986, p-value < 0.00000000000000022
## alternative hypothesis: true autocorrelation is not 0
```

Rys.16 Test Durbin-Watson dla reszt model1, model2, model3, model5.

```
##
##  Durbin-Watson test
##
## data:  model1_co
## DW = 0.9627, p-value = 0.00000002141
## alternative hypothesis: true autocorrelation is not 0

##
##  Durbin-Watson test
##
## data:  model2_co
## DW = 0.79336, p-value = 0.00000000002673
## alternative hypothesis: true autocorrelation is not 0

##
##  Durbin-Watson test
##
## data:  model3_co
## DW = 0.98311, p-value = 0.00000002779
## alternative hypothesis: true autocorrelation is not 0

##
##  Durbin-Watson test
##
## data:  model5_co
## DW = 0.89585, p-value = 0.000000002244
## alternative hypothesis: true autocorrelation is not 0
```

Rys.17 Powtórzenie testu Durbin-Watson z przekształceniem Cochrana-Orcutta dla reszt model1, model2, model3, model5.

1.6.3 Homoskedastyczność

Hipotezy testu Breusch-Pagan:

H_0: Brak heteroskedastyczności, co oznacza, że wariancje reszt są stałe i nie zależą od wartości zmiennych niezależnych w modelu.
H_a: Istnieję heteroskedastyczność, co oznacza, że wariancje reszt zmieniają się w zależności od wartości zmiennych niezależnych w modelu.

W każdym modelu p-value uzyskało wartość poniżej 0.05 (*Rys.18*) co oznacza, że odrzucono H_0 na rzecz H_a, wykryto heteroskedastyczność.

```
##
##  studentized Breusch-Pagan test
##
## data:  model1_co
## BP = 7.5926, df = 2, p-value = 0.02245

##
##  studentized Breusch-Pagan test
##
## data:  model2_co
## BP = 4.6612, df = 1, p-value = 0.03085

##
##  studentized Breusch-Pagan test
##
## data:  model3_co
## BP = 8.5406, df = 3, p-value = 0.03607

##
##  studentized Breusch-Pagan test
##
## data:  model5_co
## BP = 18.049, df = 4, p-value = 0.001207
```

Rys.18 Wykonanie testu Breusch-Pagan w celu sprawdzenia homoskedastyczności reszt w model1, model2, model3, model5.

1.6.4 Losowość

Hipotezy testu liczby serii:

H_0: Obserwacje są niezależne i identycznie rozłożone. To oznacza, że nie ma żadnego systematycznego wzorca zmiany obserwacji, co wskazuję na losowość sekwencji.

H_a: Obserwacje nie są niezależne lub nie są identycznie rozłożone. Oznacza to, że liczba serii jest zbyt mała lub zbyt duża w porównaniu z oczekiwaną liczbą serii dla sekwencji generowanych losowo, co sugeruje istnienie pewnego wzorca lub zależności w danych.

W każdym modelu otrzymano p-value poniżej 0.05 (*Rys.19*) w wyniku czego odrzucono H_0 na rzecz H_a. Oznacza to, że w modelach istnieje pewien wzorzec lub zależność, która odbiega od tego, co można by oczekiwać w przypadku sekwencji generowanej w sposób całkowicie losowy. Może to być spowodowane wcześniej wykrytą autokorelacją reszt modeli.

```
Model1:
## [1] 0
Model2:
## [1] 0.00000000000000002220446
Model3:
## [1] 0.00000000000000003108624
Model5:
## [1] 0.00000000000000002220446
```

Rys.19 Wartości p-value dla każdego z modeli uzyskanego z testu liczby serii.

1.6.5 Symetria

Przeprowadzono testy symetrii reszt dla czterech modeli regresji liniowej, mających na celu ocenę symetrii rozkładu reszt wokół zera, co jest jednym z założeń standardowych testów statystycznych. Hipoteza zerowa H_0: p = 0.5, zakładająca symetrię rozkładu, została skonfrontowana z hipotezą alternatywną H_a: p ≠ 0.5, sugerującą niesymetrię. Wartość krytyczna dla testu została ustalona na ≈1.985 na poziomie istotności α = 0.05 (test dwustronny). Z otrzymanych wyników dla każdego z modeli wywnioskowano, że wartości statystyk testowych dla wszystkich modeli przekroczyły wartość krytyczną (*Rys.20*), co prowadzi do odrzucenia hipotezy zerowej H_0 na rzecz hipotezy alternatywnej H_1 we wszystkich testowanych przypadkach.

```
Model1:
## [1] 8.717798
Model2:
## [1] 5.320966
Model3:
```

```
## [1] 7.803681

Model5:

## [1] 10.3853
```

Rys.20 Wartości p-value dla każdego z modeli uzyskanego za pomocą testu serii.

1.7 Ocena modeli ekonometrycznych

Opis oraz wnioski miar dokładności (*Rys.21*):

- Współczynnik ***I2***: Wartości współczynnika I2 są wskaźnikiem dopasowania modelu, gdzie niższa wartość wskazuje na lepsze dopasowanie modelu do danych. Model 2 ma najniższą wartość I2 równą 0.088, co sugeruje, że ma najlepsze dopasowanie spośród prezentowanych modeli.

- Błąd średni (***ME***): Wszystkie wartości są ujemne, co wskazuje na to, że modele mają tendencję do niedoszacowania wartości.

- Średni błąd bezwzględny (***MAE***): Model 5 ma najniższą wartość MAE równą 99.642, co sugeruje, że w przeciętnym przypadku ma on najmniejszy błąd prognozy.

- Średni błąd kwadratowy (***MSE***) i Pierwiastek średniego błędu kwadratowego (**RMSE**): Im niższa wartość MSE i RMSE, tym lepsze jest dopasowanie modelu. Model 5 wyróżnia się najniższymi wartościami MSE równe 10516.53 i RMSE równe 102.55, co sugeruje, że ma on najmniejsze rozproszenie błędów.

- Średni procentowy błąd (***MPE***): Pokazuje procentowy błąd w stosunku do rzeczywistych wartości. Ujemne wartości wskazują na niedoszacowanie. Model 5 ma najniższą wartość MPE, co wskazuje na mniejszy średni błąd procentowy w porównaniu z innymi modelami.

- Średni bezwzględny procentowy błąd (***MAPE***): Jest to miara dokładności prognozy w formie procentu i jest często używana do oceny jakości modeli prognozowania. Model 5 ma najniższą wartość MAPE równą 12.5, co sugeruje, że ma on największą dokładność prognoz spośród wszystkich modeli.

- Współczynnik zmienności ***RMSE*** (***VRMSE***): Jest to RMSE podzielony przez średnią wartość zmiennej prognozowanej i jest miarą relatywnej zmienności błędu. Niskie wartości wskazują na mniejszą zmienność błędu w stosunku do średniej wartości. Model 3 i model 5 mają identyczne i jednocześnie najniższe wartości VRMSE równe 0.254, co wskazuje na ich relatywną niezmienność w prognozach.

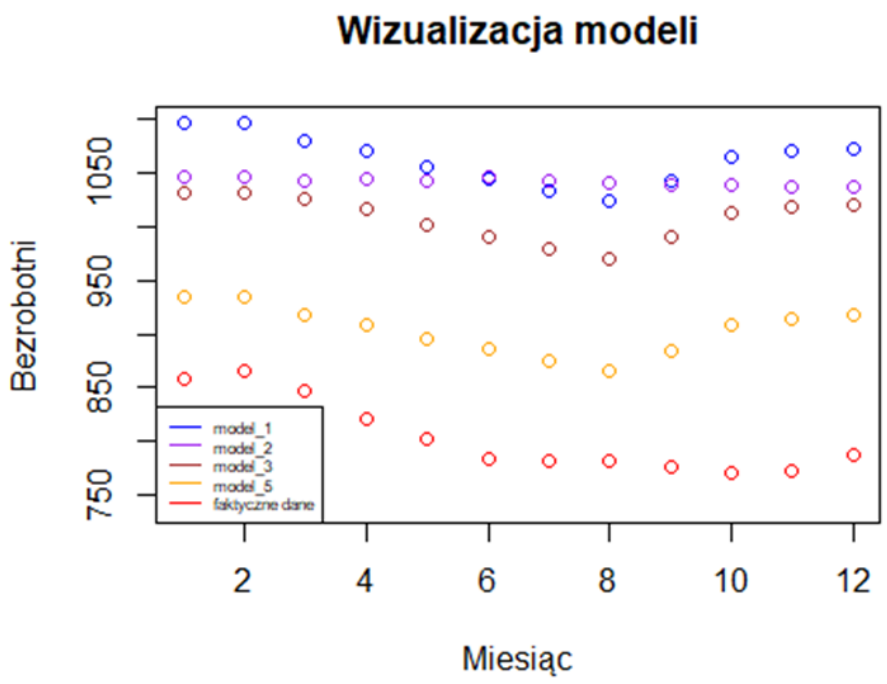
```
##          Model 1    Model 2    Model 3    Model 5
## I2        0.1037     0.0888     0.0646     0.0162
## ME     -258.2236  -237.8657  -203.0125   -99.6425
## MAE     258.2236   237.8657   203.0125    99.6425
## MSE   67151.1359 57521.3424 41834.6507 10516.5373
## MPE     -32.2451   -29.7805   -25.3872   -12.5044
## MAPE     32.2451    29.7805    25.3872    12.5044
## RMSE    259.1354   239.8361   204.5352   102.5502
## VRMSE     0.3222     0.2982     0.2543     0.2543
```

Rys.21 Wartości miar ocen dokładności modeli predykcyjnych.

Podsumowując, model 2 ma najlepsze dopasowanie według współczynnika I2, co oznacza, że może lepiej oddawać ogólną tendencję danych. Model 5 wydaje się jednak być najlepszym wyborem na podstawie analizy błędów prognozy. Wyróżnia się najniższymi wartościami MSE, RMSE i MAPE, co wskazuje na najmniejsze błędy średnie i średnie błędy procentowe.

Wizualizacja pięciu modeli (*Rys.22*) pokazuje różnorodność w przewidywanych wartościach na każdym indeksie. Model5 jest najbardziej spójny, ponieważ wszystkie jego punkty są na tym samym poziomie. Modele 1, 2, 3 i 4 wykazują różne stopnie zmienności w przewidywaniach, co sugeruje, że próbują one uwzględnić pewne zmienne lub wzorce w danych. Nie ma wyraźnego trendu wzrostowego czy spadkowego dla żadnego z modeli w odniesieniu do indeksu.

Należy jednak napomnieć, że żaden z tych modeli nie spełnił założeń wymaganych dla reszt. Więc żaden z nich nie będzie się niestety nadawać do modelowania tych danych.



Rys.22 Graficzne porównanie uzyskanych modeli. Modele 1, 2, 3 i 5 są oznaczone różnymi kolorami, faktyczne dane są przedstawione kółkami.

2.Model szeregu czasowego

Celem tego etapu było stworzenie modelu szeregu czasowego, który w najlepszy możliwy sposób potrafiłby przewidzieć zmiany w liczbie ludzi bezrobotnych. Pierwszym etapem było przekonwertowanie danych (*Rys.23*) do obiektu serii czasowej za pomocą funkcji ts(). Następnie tak jak w modelu ekonometrycznym podzielono dane na zbiór testowy (*Rys.24*) oraz zbiór uczący.

```
##         Jan    Feb    Mar    Apr    May    Jun    Jul    Aug    Sep    Oct
## 2015 1918.6 1918.7 1860.6 1782.2 1702.1 1622.3 1585.7 1563.5 1539.4 1516.9
## 2016 1647.5 1652.7 1600.5 1521.8 1456.9 1392.5 1361.5 1346.9 1324.1 1308.0
## 2017 1397.1 1383.4 1324.2 1252.7 1202.1 1151.6 1140.0 1136.1 1117.1 1069.5
## 2018 1133.7 1126.7 1092.2 1042.5 1002.2  967.9  961.8  958.6  947.4  937.3
## 2019 1023.1 1016.7  984.7  938.3  906.0  877.1  868.4  865.5  851.2  840.5
## 2020  922.2  919.9  909.4  965.8 1011.7 1026.5 1029.5 1028.0 1023.7 1018.4
## 2021 1090.4 1099.5 1078.4 1053.8 1026.7  993.4  974.9  960.8  934.7  910.9
## 2022  927.1  921.8  902.1  878.0  850.2  818.0  810.2  806.9  801.7  796.0
## 2023  857.6  864.8  846.9  821.9  802.3  783.5  782.4  782.5  776.0  770.4
##         Nov    Dec
## 2015 1530.6 1563.3
## 2016 1313.6 1335.2
## 2017 1067.7 1081.7
## 2018  950.5  968.9
## 2019  849.6  866.4
## 2020 1025.7 1046.4
## 2021  898.8  895.2
## 2022  800.2  812.3
## 2023  773.4  788.2
```

Rys.23 Wartości uzyskane po zastosowaniu funkcji ts() w modelu szeregów czasowych.

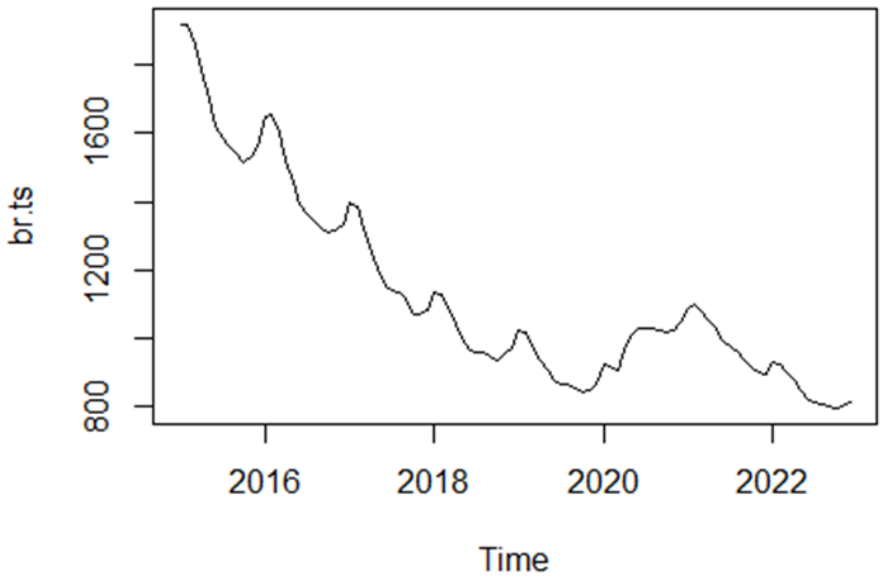
```
##         Jan    Feb    Mar    Apr    May    Jun    Jul    Aug    Sep    Oct
## 2015 1918.6 1918.7 1860.6 1782.2 1702.1 1622.3 1585.7 1563.5 1539.4 1516.9
## 2016 1647.5 1652.7 1600.5 1521.8 1456.9 1392.5 1361.5 1346.9 1324.1 1308.0
## 2017 1397.1 1383.4 1324.2 1252.7 1202.1 1151.6 1140.0 1136.1 1117.1 1069.5
## 2018 1133.7 1126.7 1092.2 1042.5 1002.2  967.9  961.8  958.6  947.4  937.3
## 2019 1023.1 1016.7  984.7  938.3  906.0  877.1  868.4  865.5  851.2  840.5
## 2020  922.2  919.9  909.4  965.8 1011.7 1026.5 1029.5 1028.0 1023.7 1018.4
## 2021 1090.4 1099.5 1078.4 1053.8 1026.7  993.4  974.9  960.8  934.7  910.9
## 2022  927.1  921.8  902.1  878.0  850.2  818.0  810.2  806.9  801.7  796.0
##         Nov    Dec
## 2015 1530.6 1563.3
## 2016 1313.6 1335.2
## 2017 1067.7 1081.7
## 2018  950.5  968.9
## 2019  849.6  866.4
## 2020 1025.7 1046.4
## 2021  898.8  895.2
## 2022  800.2  812.3
```

Rys.24 Podzielenie danych na zbiór testowy w modelu szeregów czasowych.

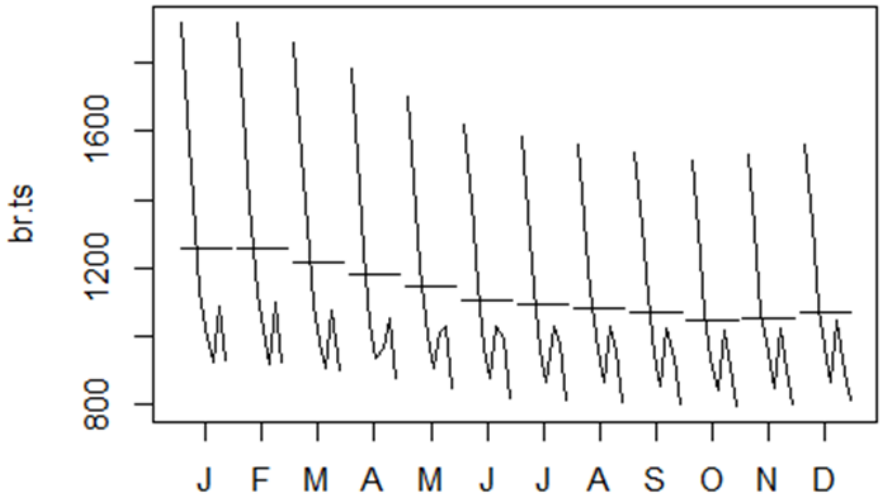
Zaobserwowano trend wskazujący na spadek liczby bezrobotnych w latach 2015-2023 (*Rys.25*), chociaż zmniejszenie to nie jest jednostajne, a liczba bezrobotnych wykazuje pewne wahania. W 2020 roku, widać wyraźne odwrócenie trendu i wzrost liczby bezrobotnych, co może odzwierciedlać wpływ wydarzeń globalnych, takich jak pandemia COVID-19.

Dodatkowo można zauważyć powtarzający się wzorzec w ciągu roku (*Rys.26*), który sugeruje sezonowość w danych o bezrobociu. W każdym roku obserwujemy powtarzające się szczyty i dołki. Szczyty (większa liczba bezrobotnych) wydają się występować w okolicach pierwszych miesięcy roku (stycznia i lutego), a następnie liczba bezrobotnych spada. To może być związane z sezonowymi zwolnieniami po okresie świątecznym lub innymi czynnikami ekonomicznymi typowymi dla tego okresu. Różnica między najwyższym, a najniższym poziomem bezrobocia jest znacząca i regularna, co wskazuje na silne sezonowe zmiany w liczbie bezrobotnych.

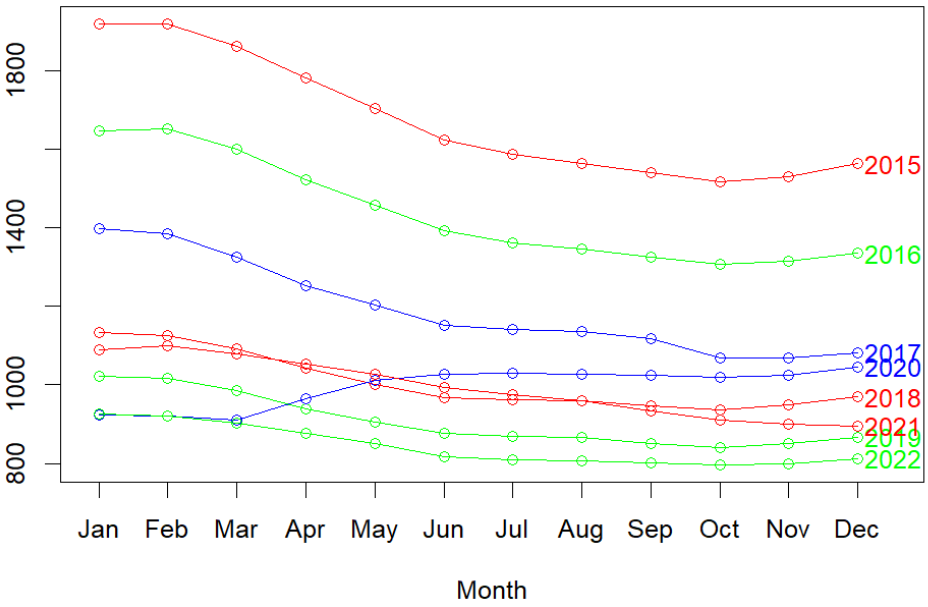
Rok 2015 zaczyna się z najwyższym poziomem bezrobocia (*Rys.27*), a następne lata pokazują niższe wartości na początku roku. To może wskazywać na poprawę sytuacji na rynku pracy w kolejnych latach. Rok 2022 rozpoczyna się z najniższą liczbą bezrobotnych w styczniu w porównaniu do innych lat i utrzymuje tę tendencję w kolejnych miesiącach.



Rys.25 Przedstawienie liczby bezrobotnych w latach 2015-2023 (br.ts czyli liczba bezrobotnych).



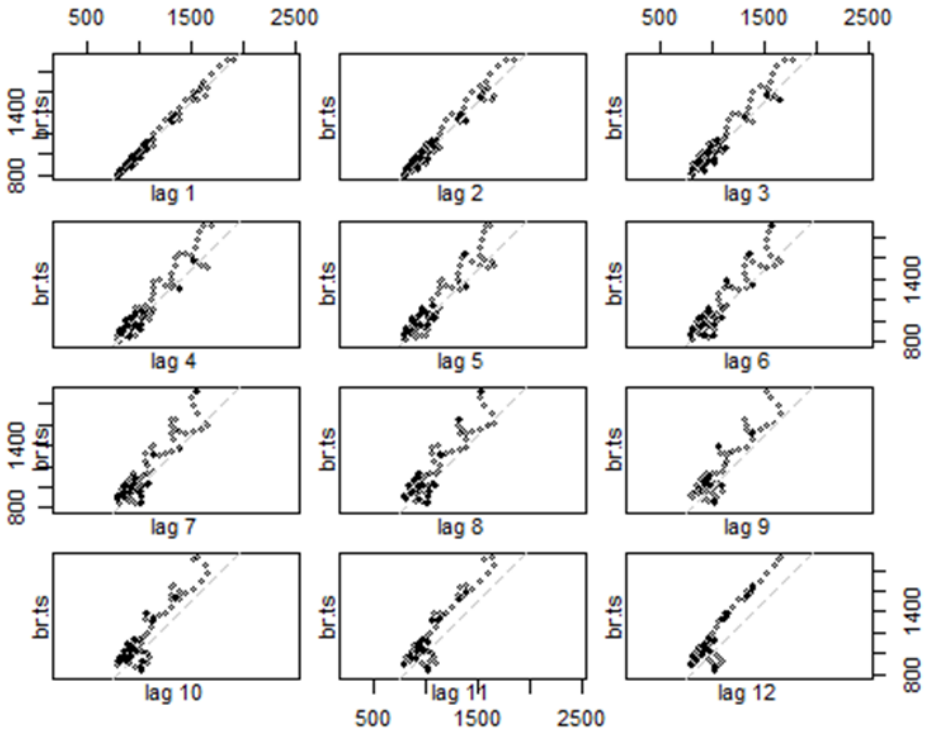
Rys.26 Przedstawienie liczby bezrobotnych w skali miesięcznej (br.ts czyli liczba bezrobotnych).



Rys.27 *Przedstawienie rocznej zależności liczby bezrobotnych w skali miesięcznej (na osi y liczba bezrobotnych).*

Na wykresie lag można zaobserwować na wszystkich wykresach silną dodatnią korelację między wartościami oryginalnymi a ich opóźnieniami (*Rys.28*). Oznacza to, że wysokie wartości w jednym okresie często odpowiadają wysokim wartościom w poprzednich okresach, co jest typową cechą autokorelowanych danych. Chociaż wszystkie wykresy wykazują dodatnią korelację, wydaje się, że siła korelacji zmniejsza się w miarę zwiększania liczby opóźnień (lag). Na przykład, wykres dla opóźnienia 1 wykazuje bardziej skoncentrowane punkty wokół linii

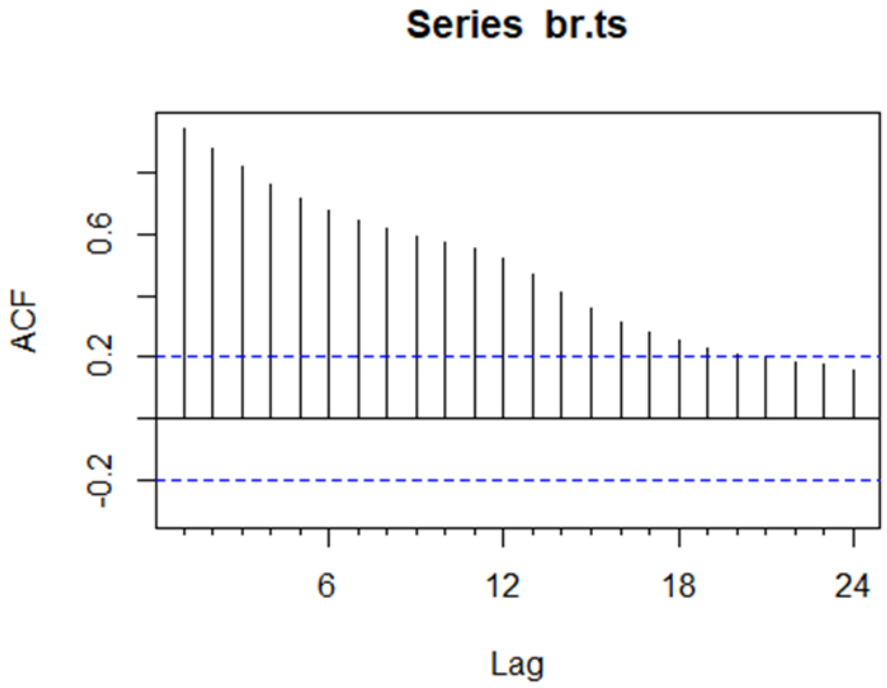
przerywanej w porównaniu do wykresu dla opóźnienia 12, gdzie punkty są bardziej rozproszone.



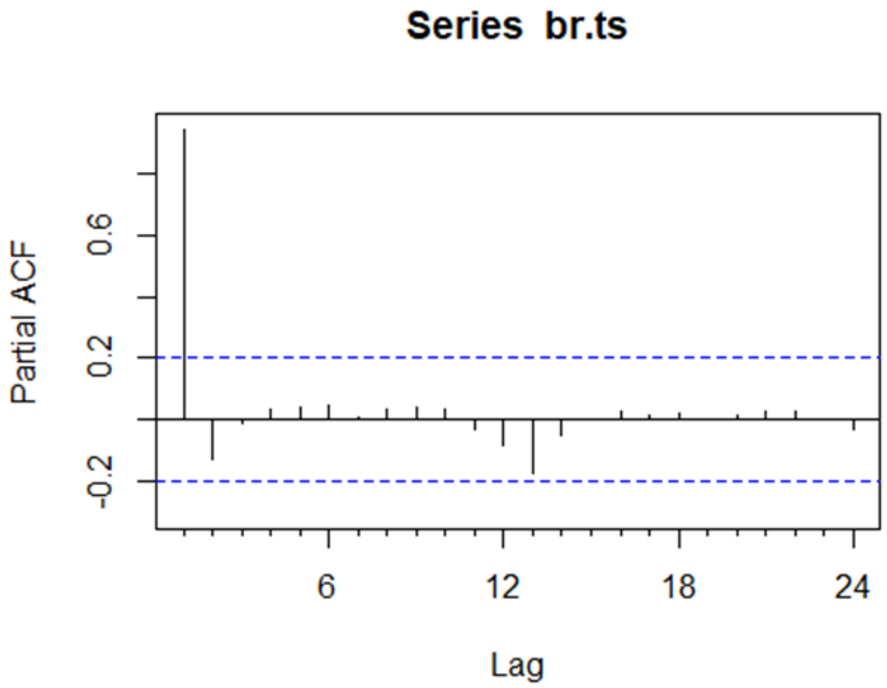
Rys.28 Zestaw wykresów (lag plot), które pokazują zależności między wartościami obserwacji, a ich opóźnieniami. Każdy wykres rozrzutu odpowiada parze wartości: oryginalnych danych i tych samych danych przesuniętych o określoną liczbę okresów (od 1 do 12).

Pierwsze opóźnienia wykazują wyższe wartości ACF (*Rys.29*), co wskazuje na silną dodatnią autokorelację krótkoterminową w danych. Oznacza to, że liczba bezrobotnych w jednym miesiącu ma tendencję do bycia podobną do liczby w poprzednich miesiącach. Autokorelacja jest znacząca przynajmniej do laga 6, po czym słupki są niższe, ale nadal pozostają w granicach istotności statystycznej aż do około laga 14. Stopniowy spadek wartości ACF wraz ze wzrostem lagu może wskazywać na to, że wpływ przeszłych danych maleje z czasem. Znacząca autokorelacja w serii czasowej oznacza, że do modelowania i prognozowania liczby bezrobotnych mogą być przydatne modele ARIMA lub inne modele szeregów czasowych, które biorą pod uwagę autokorelację.

Wartość PACF (*Rys.30*) dla pierwszego opóźnienia wskazuje, że model AR(1) (autoregresyjny pierwszego rzędu) może być odpowiedni dla tej serii czasowej, ponieważ tylko pierwsze opóźnienie ma istotną korelację.



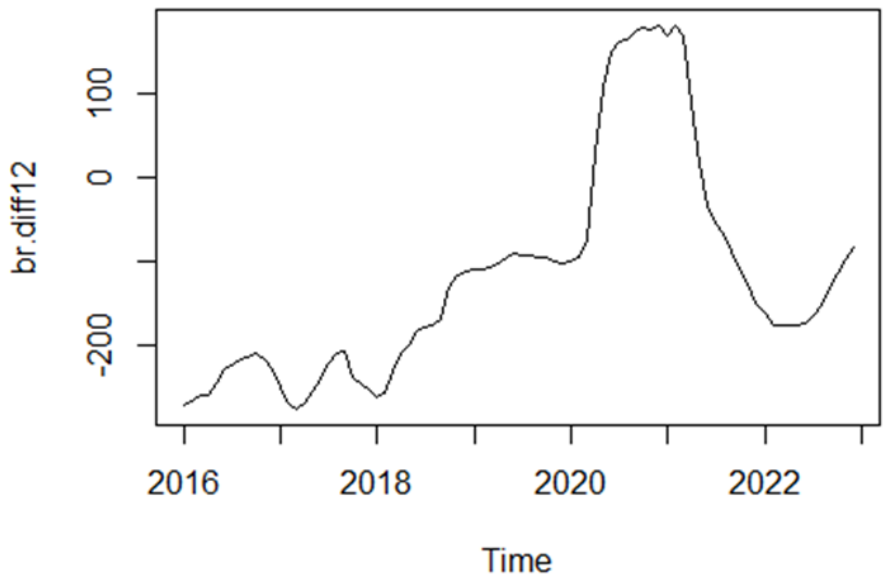
Rys.29 *Wykres funkcji autokorelacji (ACF) dla serii czasowej reprezentującej liczbę bezrobotnych (br.ts).* Poziome przerywane linie reprezentują granice istotności statystycznej.



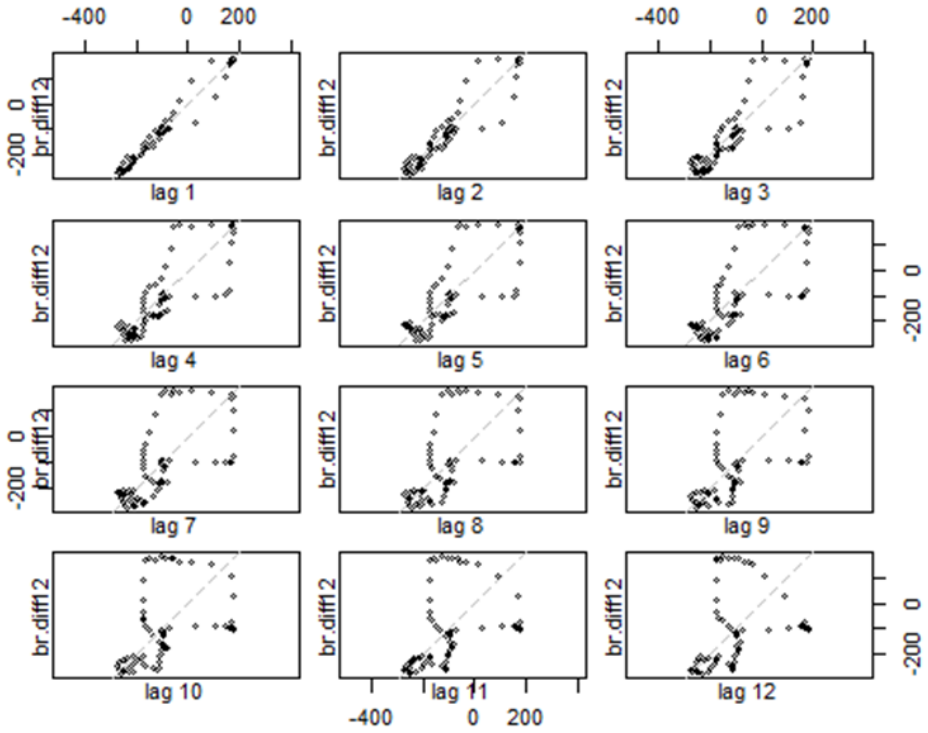
Rys.30 Wykres funkcji autokorelacji cząstkowej (pACF) dla serii czasowej reprezentującej liczbę bezrobotnych (br.ts). Poziome przerywane linie reprezentują granice istotności statystycznej.

2.1 Różnicowanie z opóźnieniem 12

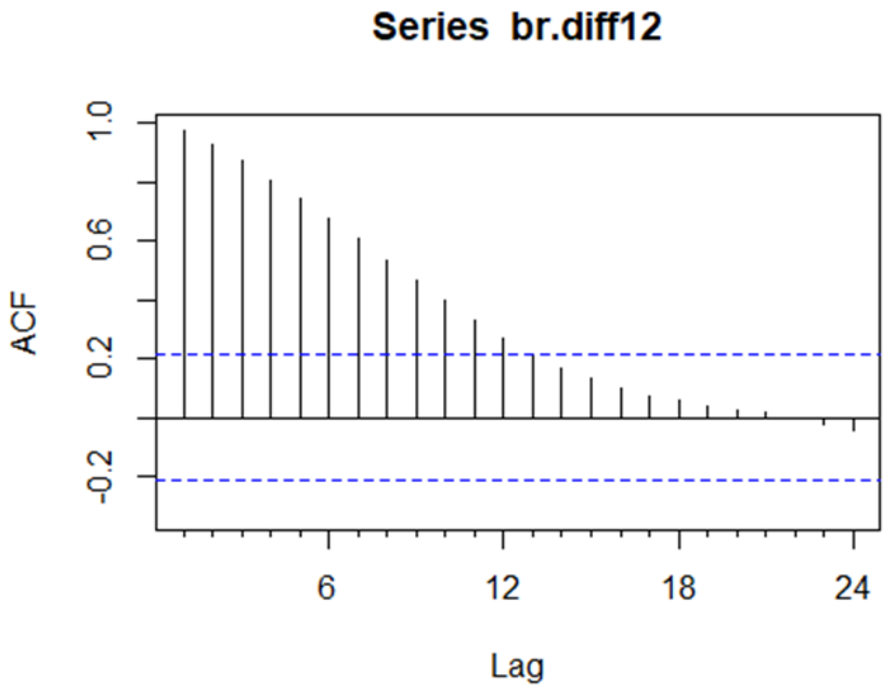
Różnicowanie z opóźnieniem 12 miesięcy jest stosowane do usunięcia sezonowości w danych. Jeśli początkowe dane miały silny składnik sezonowy, proces ten może pomóc w wyizolowaniu innych efektów. Przez większość czasu, wartości oscylują wokół zera (*Rys.31*), co sugeruje, że po usunięciu sezonowości i trendów, pozostałe fluktuacje są nieregularne. Widać znaczący szczyt w okolicy 2020 roku, który wypada poza typowy zakres fluktuacji. Może to wskazywać na nieoczekiwane wydarzenie spowodowane globalna pandemia COVID-19, która miała duży wpływ na rynek pracy.
Punkty na wykresach lag są dość rozproszone (*Rys.32*), co jest dobrym wskaźnikiem, że w danych jest losowość, sugerując, że różnicowanie z powodzeniem usunęło część autokorelacji. Dodatkowo autokorelacje są znaczące przy pierwszych kilku opóźnieniach (*Rys.33*), a następnie szybko zanikają, co sugeruje, że zróżnicowany szereg może nadal zawierać pewną strukturę autokorelacji, którą można by uchwycić za pomocą modelu średniej ruchomej (MA).
Większość słupków PACF znajduje się wewnątrz granic istotności statystycznej (*Rys.34*), co oznacza, że nie ma silnych dowodów na autokorelację cząstkową w danych poza pierwszym lagiem. Słupek dla lagu 1 przekracza górną granicę istotności, co sugeruje, że pierwsze opóźnienie ma statystycznie istotny wpływ na serię. Na podstawie tego wykresu, model AR(1) może być odpowiedni dla różnicowanego szeregu czasowego, ponieważ tylko opóźnienie pierwsze jest istotne.



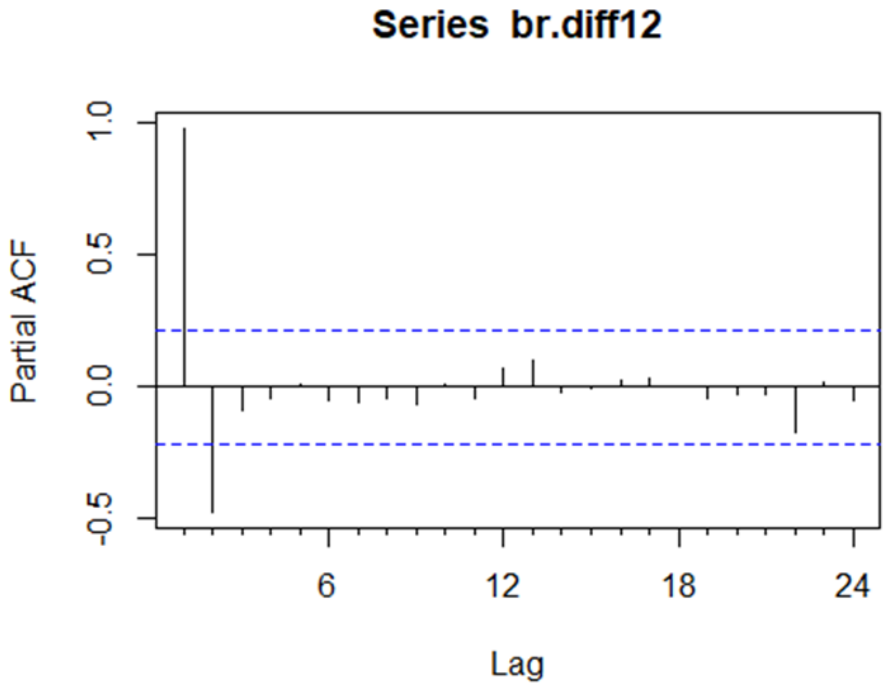
Rys.31 Wykres szeregu czasowego dla danych z opóźnieniem 12.



Rys.32 Zestaw wykresów lag dla liczby bezrobotnych w latach 2015-2023 z opóźnieniem 12.



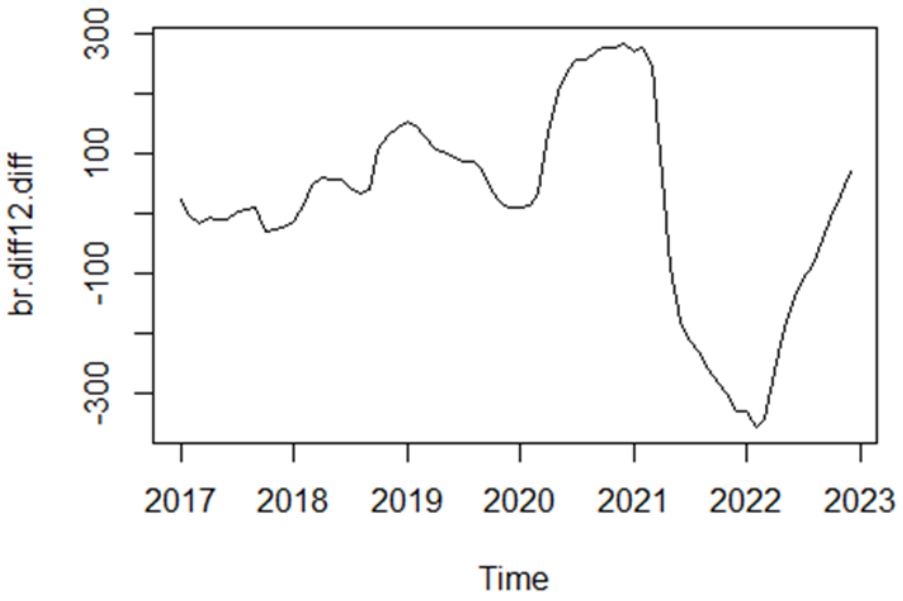
Rys.33 Wykres funkcji autokorelacji (ACF) dla serii czasowej reprezentującej liczbę bezrobotnych (br.ts) z późnieniem 12. Poziome przerywane linie reprezentują granice istotności statystycznej.



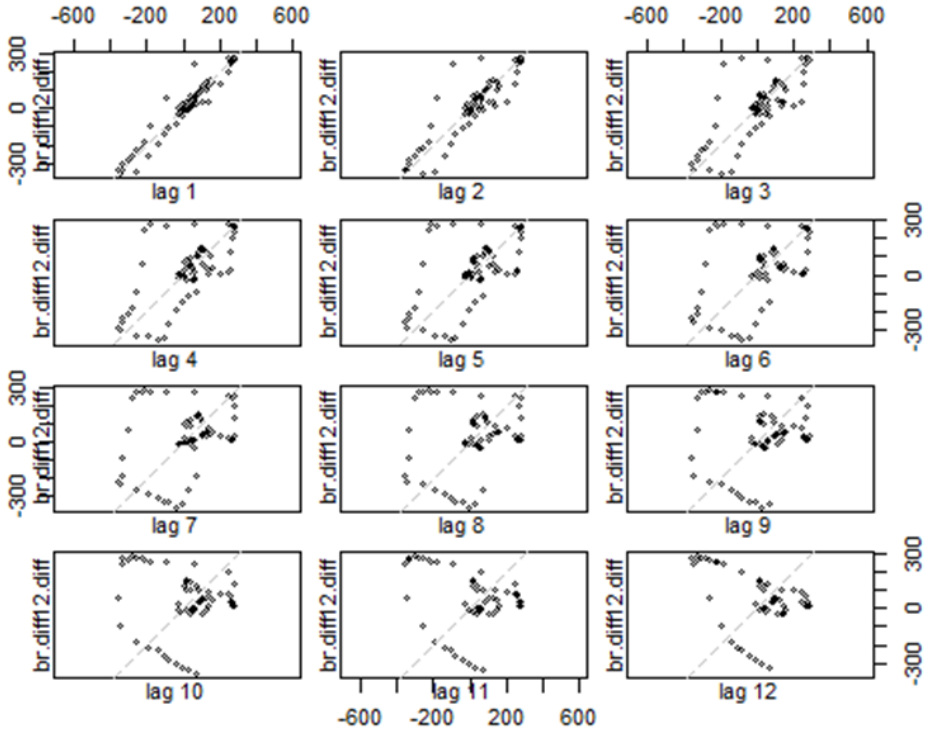
Rys.34 Wykres częściowej autokorelacji pACF z opóźnieniem 12. Większość częściowych autokorelacji mieści się w przedziałach ufności (z wyjątkiem być może pierwszego opóźnienia).

2.2 Różnicowanie z opóźnieniem 1

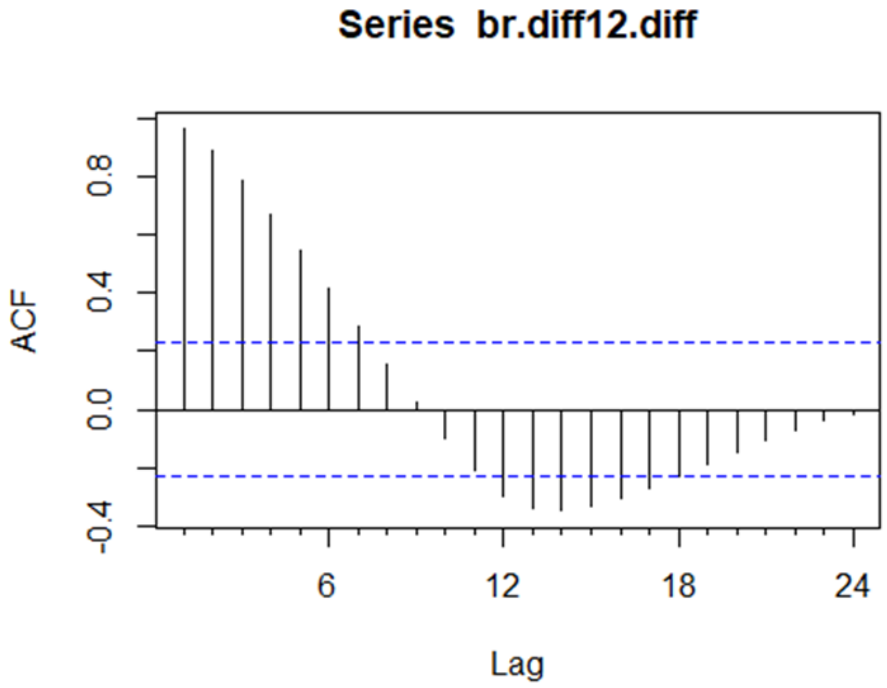
Punkty na wykresach lag są jeszcze lepiej rozproszone (*Rys.36*), co jest dobrym wskaźnikiem, że w danych jest losowość, sugerując, że kolejne różnicowanie z powodzeniem usunęło część autokorelacji. Można stwierdzić że zróżnicowane dane nadal posiadają znaczące autokorelacje (*Rys.37*). Większość częściowych autokorelacji mieści się w przedziałach ufności (*Rys.38*).



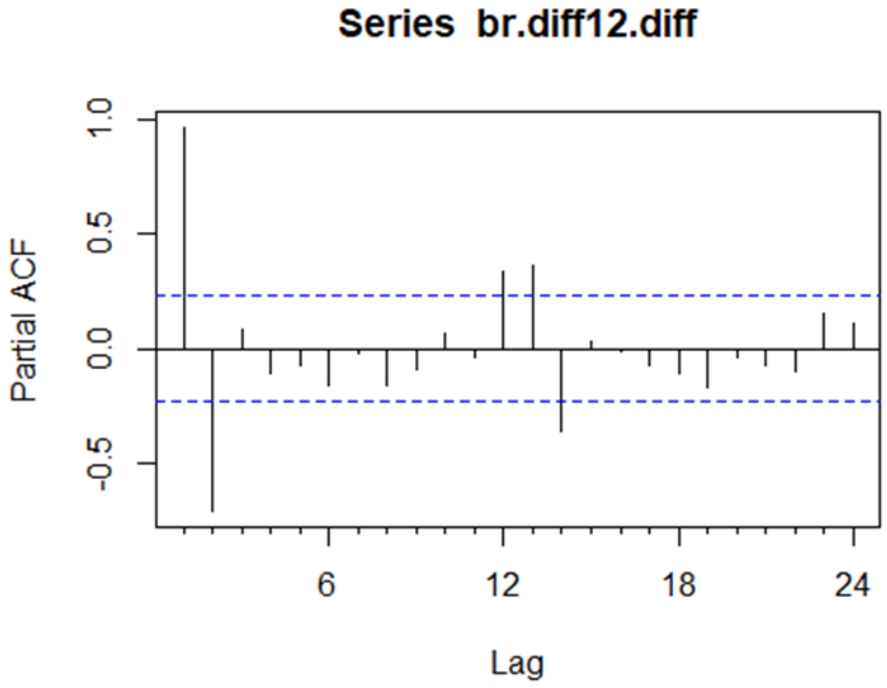
Rys.35 Wykres szeregu czasowego dla danych z opóźnieniem z kolejnym opóźnieniem 1.



Rys.36 Zestaw wykresów z opóźnieniem równym 1.



Rys.37 Wykres funkcji autokorelacji (ACF) dla serii czasowej reprezentującej liczbę bezrobotnych (br.ts) z późnieniem 1. Poziome przerywane linie reprezentują granice istotności statystycznej.

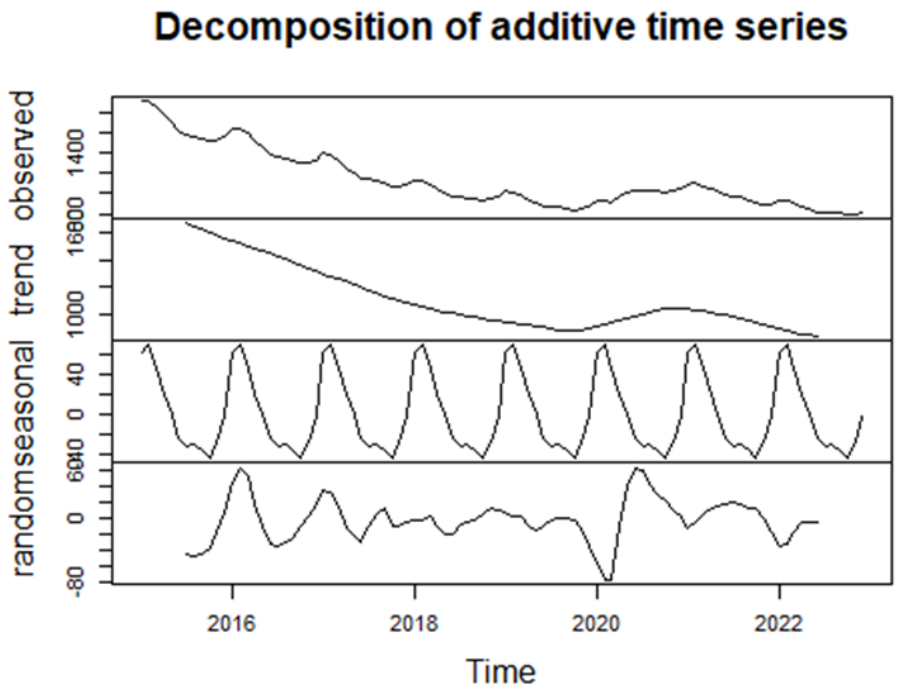


Rys.38 Wykres częściowej autokorelacji pACF z opóźnieniem 1.

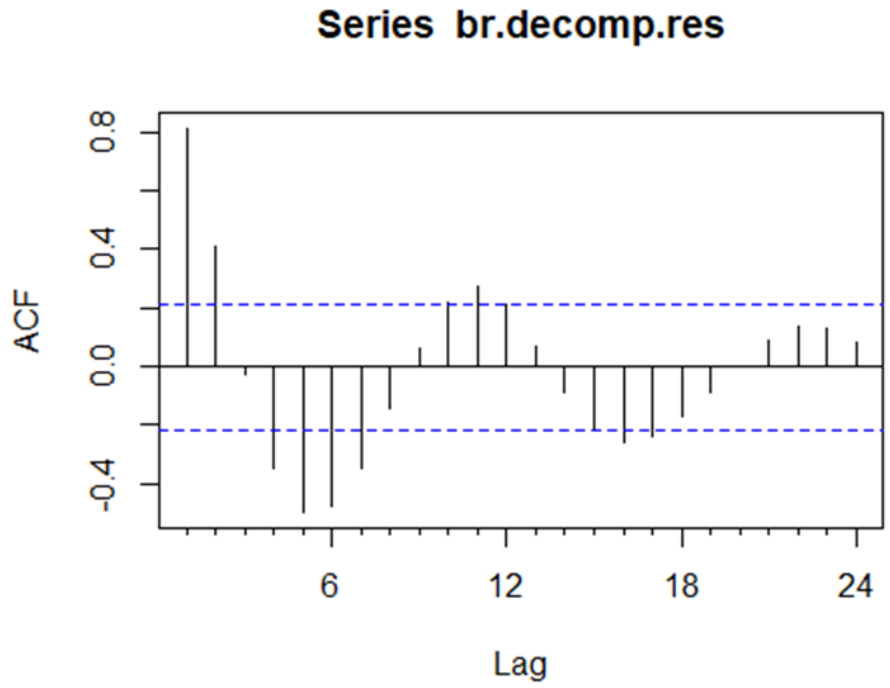
2.3 Dekompozycja klasyczna

Klasyczna dekompozycja addytywnego szeregu czasowego rozkłada dane na trzy główne składniki: trend, sezonowość i reszty. Pierwszy panel od góry (*Rys.39*) pokazuje

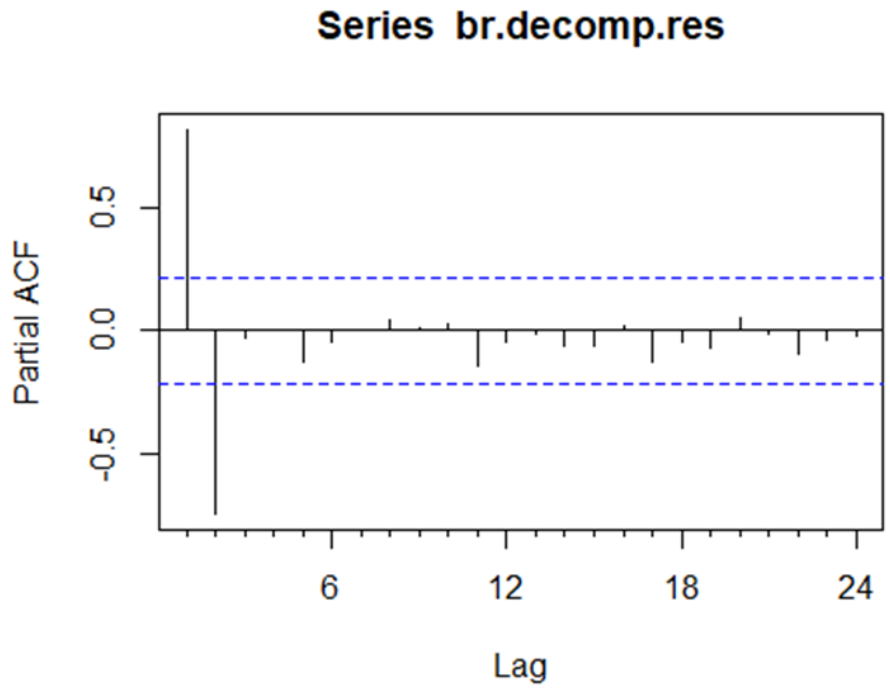
oryginalne obserwacje (observed), czyli rzeczywiste dane. Drugi panel przedstawia wyodrębniony trend (trend), wskazujący na długoterminowy wzrost lub spadek w danych. Trzeci panel ilustruje sezonowość (seasonal), czyli regularne wzorce powtarzające się w cyklicznych okresach. Ostatni panel pokazuje składnik losowy (random), również nazywany resztami (residuals), które są różnicą między danymi oryginalnymi, a sumą trendu i sezonowości. Składnik ten zawiera nieregularności, które nie zostały wyjaśnione przez dwa poprzednie składniki.
Współczynniki są bliskie zeru dla większości opóźnień (*Rys.40*) co sugeruje, że nie ma silnej autokorelacji po dekompozycji serii. Jest to pożądane dla reszt modelu. Wykres (*Rys.41*) pokazuje, że nie ma potrzeby dodatkowego modelowania autoregresji na szeregach, a model AR o wiekszy rzędzie nie jest konieczny.



Rys.39 Wykres przedstawia klasyczną dekompozycję addytywnego szeregu czasowego, która rozkłada dane na trzy główne składniki: trend, sezonowość i reszty.



Rys.40 Wykres przedstawia funkcję autokorelacji ACF dla serii czasowej po dekompozycji.

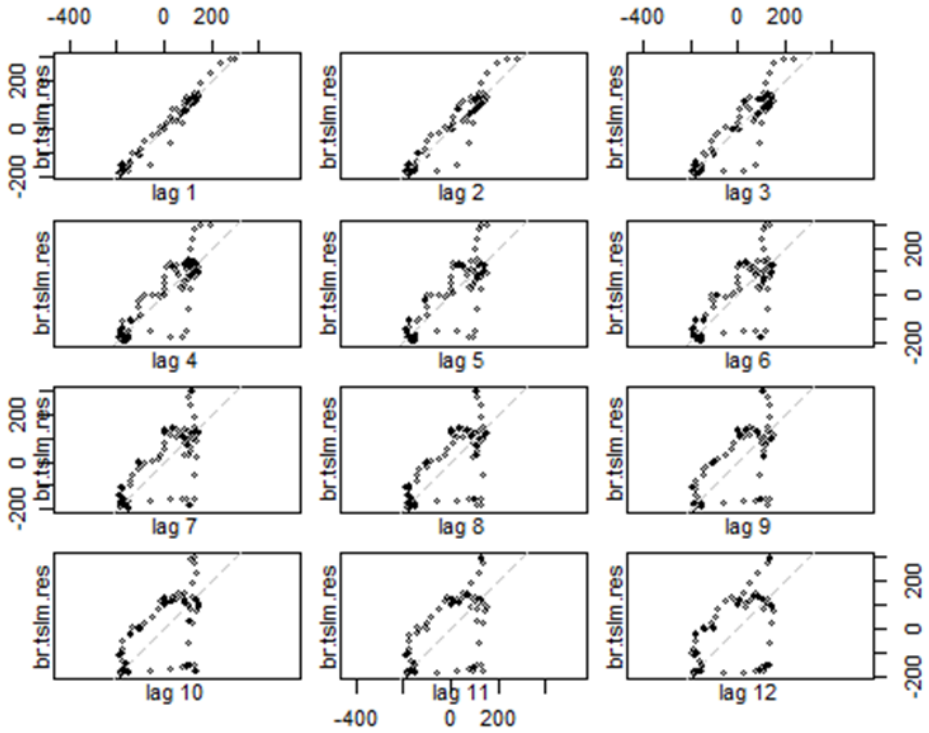


Rys.41 Wykres przedstawia funkcję autokorelacji cząstkowej pACF dla serii czasowej po dekompozycji.

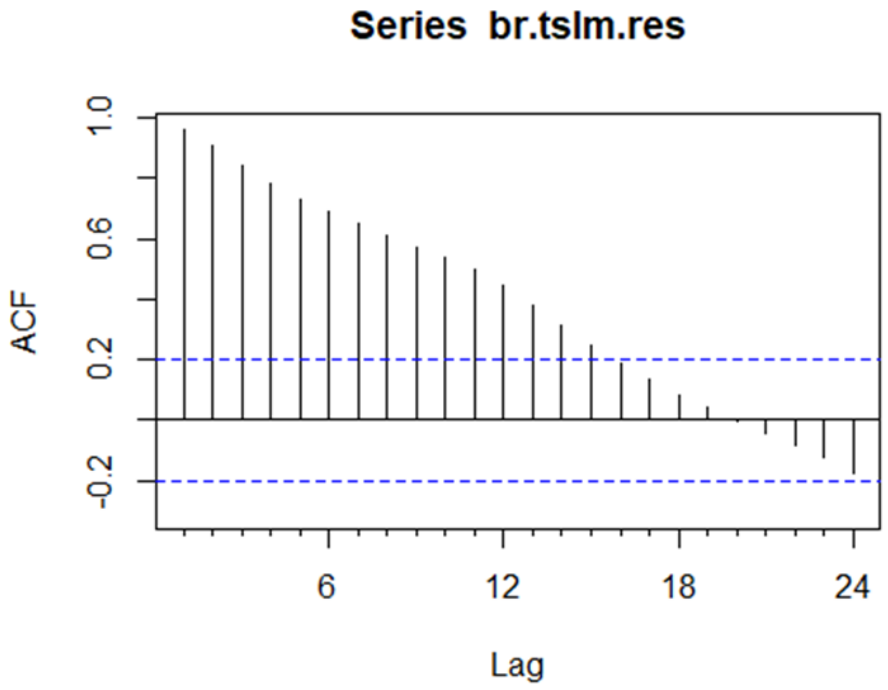
2.4 Dekompozycja na podstawie modelu regresji.

Obserwowalna jest silna dodatnia korelacja między serią czasową, a jej opóźnionymi wartościami (*Rys.42*). Długie ogony współczynników ACF (*Rys.43*) wskazują, że reszty modelu

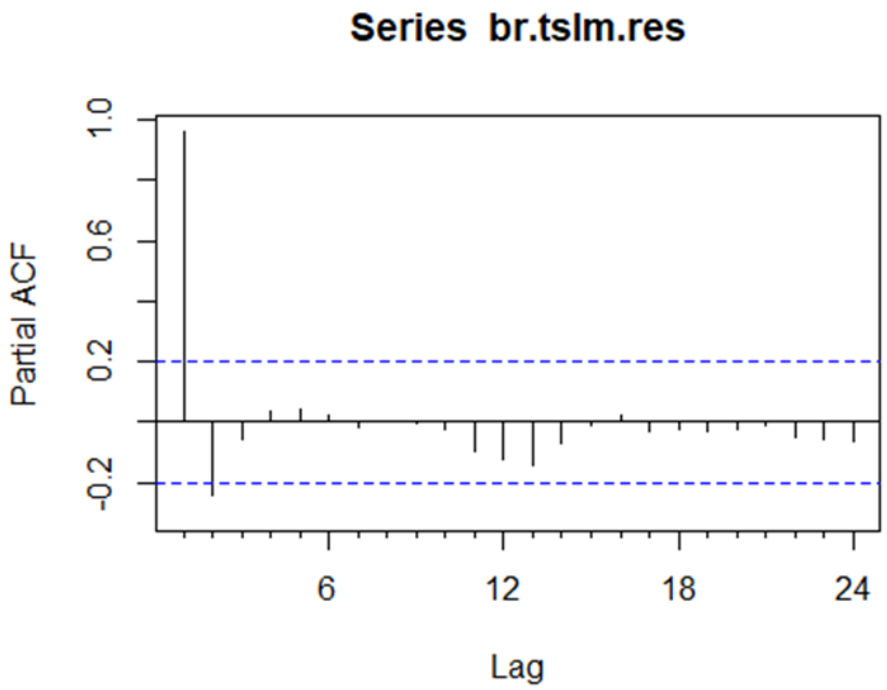
regresji wykazują pewien stopień autokorelacji a więc model regresji nie wyjaśnił całkowicie autokorelacji w danych. Współczynniki pACF szybko spadają do zera (*Rys.44*) i wiekszość z nich jest wewnątrz przedziału ufności co wskazuje na to, że nie ma dodatkowych autkorelacji cząstkowych w resztach.



Rys.42 Zestaw wykresów Lag dekompozycja na podstawie modelu regresji.



Rys.43 Wykres ACF dekompozycja na podstawie modelu regresji.



Rys.44 Wykres pACF dekompozycja na podstawie modelu regresji.

2.5 Konstruowanie modelu MA(1)

```
## Series: br.ts
## ARIMA(0,1,1)(0,1,0)[12]
##
## Coefficients:
##          ma1
##       0.6812
## s.e.  0.0643
##
## sigma^2 = 299.4:  log likelihood = -354.12
## AIC=712.23   AICc=712.38   BIC=717.07
##
## Training set error measures:
##                     ME     RMSE      MAE       MPE     MAPE       MASE      ACF1
## Training set 1.126258 15.99122 9.553194 0.1283627 0.936699 0.05796339 0.2849569
```

Rys.45 Wyniki analizy szeregu czasowego z zastosowaniem modelu ARIMA(0,1,1)(0,1,0) [12].

Sprawdzono następnie istotność współczynników modelu (*Rys.45*) i wykonano korekty (*Rys.46*), które okazały się bezużyteczne. Kolejnym etapem było skonstruowanie prognozy dla tego modelu (*Rys.47, rys.48*).

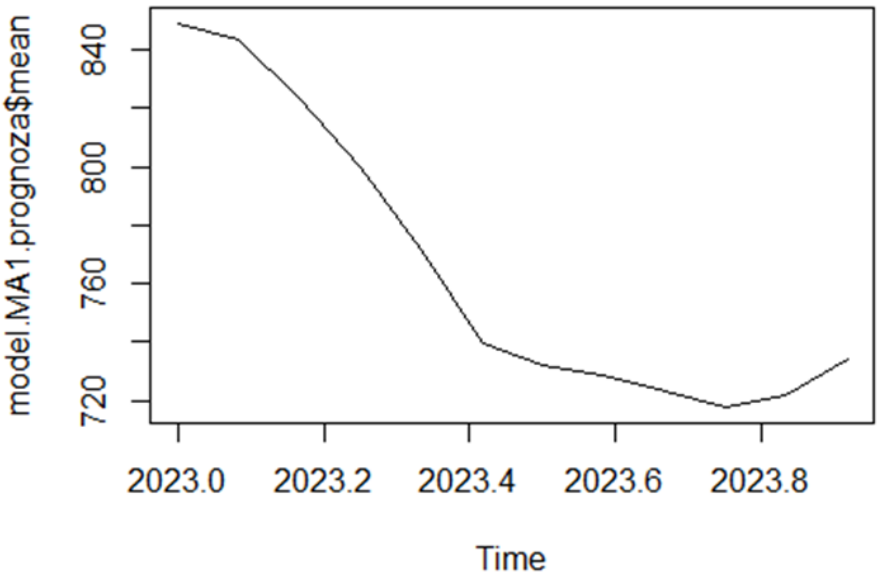
```
## Series: br.ts
## ARIMA(0,1,1)(0,1,0)[12]
##
## Coefficients:
##           ma1
##       0.6812
## s.e.  0.0643
##
## sigma^2 = 299.4:  log likelihood = -354.12
```

```
## AIC=712.23   AICc=712.38   BIC=717.07
##
## Training set error measures:
##                     ME     RMSE      MAE       MPE     MAPE       MASE      ACF1
## Training set 1.126258 15.99122 9.553194 0.1283627 0.936699 0.05796339 0.2849569
```

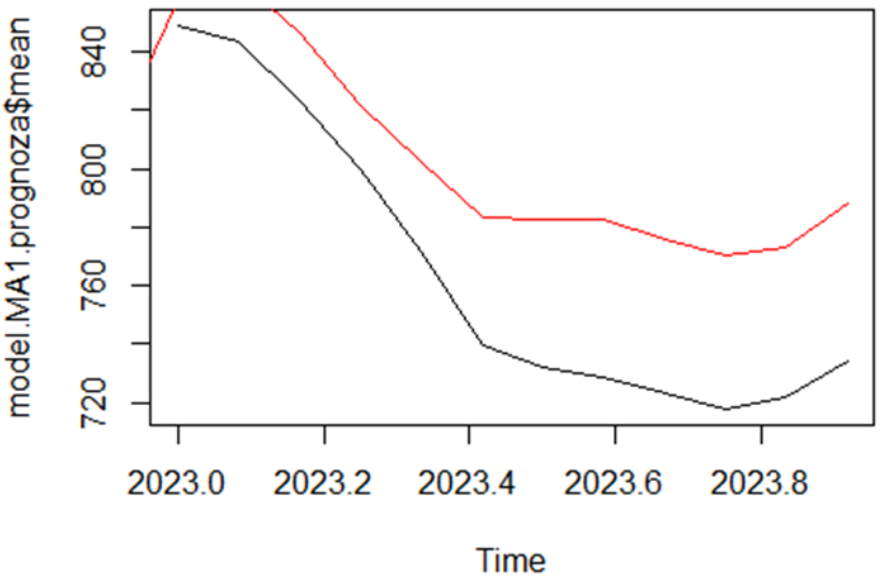
Rys.46 Wyniki analizy szeregu czasowego z zastosowaniem modelu ARIMA(0,1,1)(0,1,0) po korektach [12].

```
##              Jan        Feb        Mar        Apr        May        Jun        Jul        Aug
##  2023   848.9673   843.6673   823.9673   799.8673   772.0673   739.8673   732.0673   728.7673
##                              Sep                Oct                Nov                Dec
## 2023 723.5673 717.8673 722.0673 734.1673
```

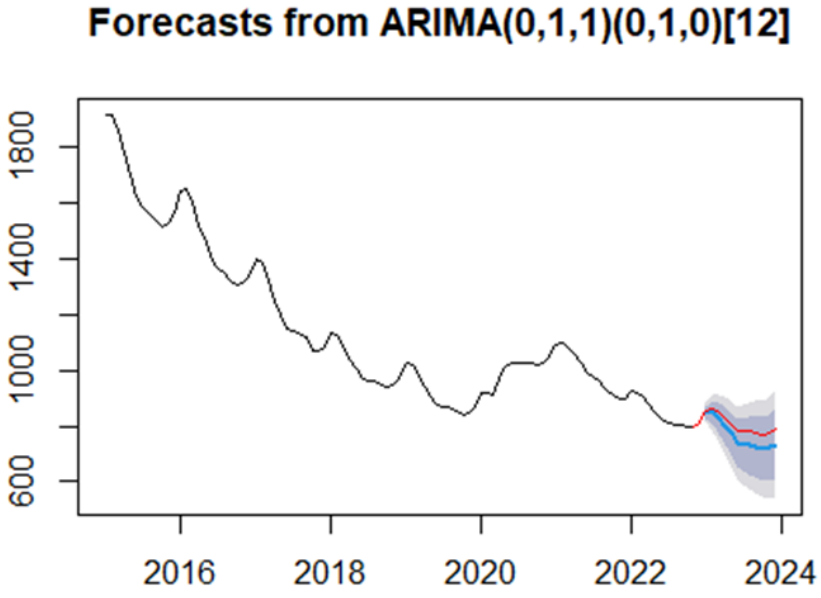
Rys.47 Prognozowane wartości osób bezrobotnych na rok 2023.



Rys.48 Wizualne prognozowane wartości osób bezrobotnych na rok 2023.



Rys.49 Porównanie prognozowanych danych z faktycznymi danymi (kolor czerwony) na rok 2023 dla modelu MA(1).



***Rys.*50** Porównanie prognozowanych danych (kolor niebieski) z faktycznymi danymi (kolor czerwony) na rok 2023.

```
##                              ME          MAE          MPE       MAPE       RMSE  Theil's U
##  Training  set    1.126258    9.553194  0.1283627  0.936699  15.99122                NA
## Test set     38.582742 38.582742 4.8786784 4.878678 41.66770  3.347574
```

Rys.51 Miary błędu dla zestawów treningowego i testowego

Statystyki modelu, takie jak wartość logarytmu wiarygodności i kryteria informacyjne AIC, AICc oraz BIC, są w zakresach akceptowalnych (*Rys.46*). Wartości te pomagają ocenić, jak dobrze model dopasowuje się do danych, biorąc jednocześnie pod uwagę złożoność modelu, aby uniknąć nadmiernego dopasowania.

Pod względem błędów prognozy (*Rys.51)*, model wykazuje pewne niedokładności, co widać po wartościach RMSE oraz Theil's U. Szczególnie wartość Theil's U powyżej 1 dla danych testowych wskazuje, że model nie jest w stanie przewidywać z dużą precyzją. W związku z tym, chociaż model może być przydatny do wstępnego zrozumienia dynamiki szeregu, jego użyteczność jako narzędzia prognostycznego może być ograniczona.

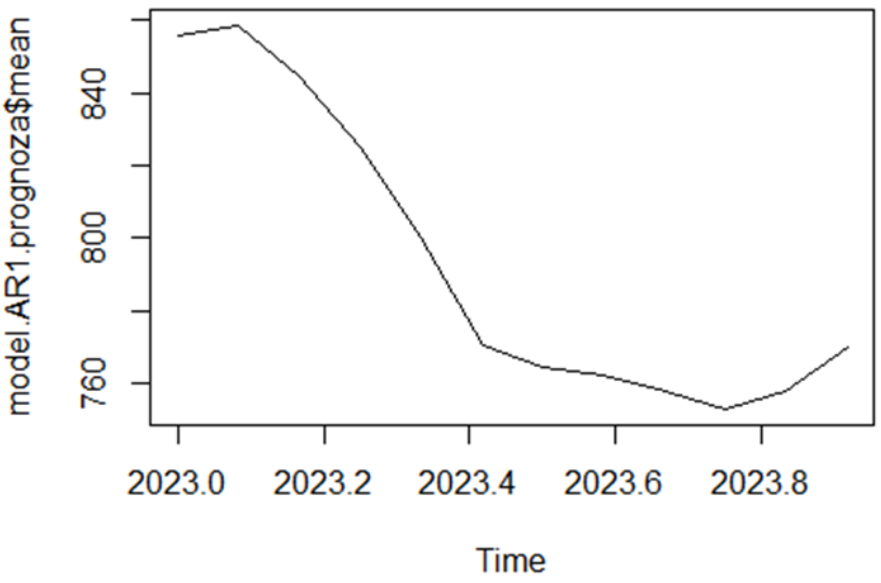
2.6 Konstruowanie modelu AR(1)

```
## Series: br.ts
## ARIMA(1,1,0)(0,1,0)[12]
##
## Coefficients:
##          ar1
##       0.7263
## s.e.  0.0735
##
## sigma^2 = 257.7:  log likelihood = -347.95
## AIC=699.9   AICc=700.05   BIC=704.74
##
## Training set error measures:
##                      ME     RMSE      MAE        MPE     MAPE       MASE
## Training set 0.5734193 14.83729 8.138886 0.07475213 0.787562 0.04938216
##                    ACF1
## Training set 0.1830271
```

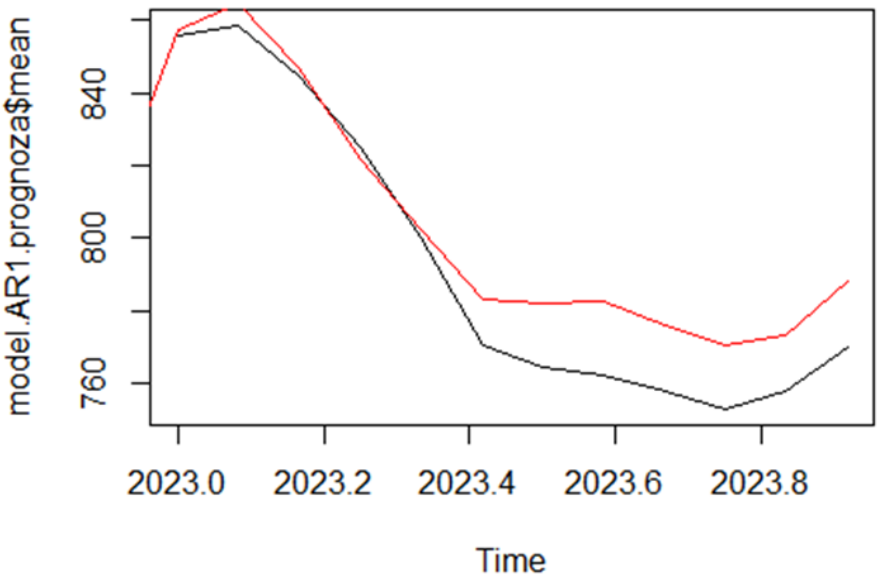
Rys.52 Wyniki analizy szeregu czasowego z zastosowaniem modelu ARIMA(1,1,1)(0,1,0) [12].

```
##             Jan       Feb       Mar       Apr       May       Jun       Jul       Aug
##  2023  855.6029  858.5849  844.9000  825.1688  800.5419  770.6465  764.5204  762.4361
##                            Sep               Oct               Nov               Dec
## 2023 758.1190 753.0603 757.7261 770.1644
```

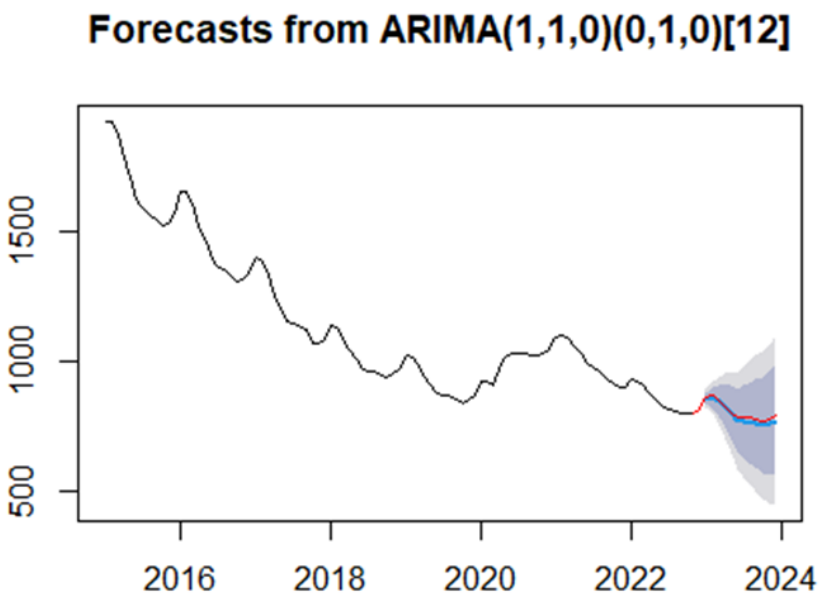
Rys.53 Prognozowane liczby osób bezrobotnych na rok 2023



Rys.54 Wizualna prognoza osób bezrobotnych na rok 2023 za pomocą modelu AR(1).



Rys.55 *Porównanie prognozowanych danych z faktycznymi danymi (kolor czerwony) na rok 2023 dla modelu AR(1).*



Rys.56 Porównanie prognozowanych danych przy użyciu modelu AR(1) (kolor niebieski)z faktycznymi danymi (kolor czerwony) na rok 2023.

```
##                       ME       MAE        MPE     MAPE     RMSE Theil's U
## Training set  0.5734193  8.138886 0.07475213 0.787562 14.83729        NA
## Test set     10.7023824 11.247190 1.36406231 1.430349 13.35163  1.080175
```

Rys.57 Miary błędu dla zestawów treningowego i testowego

Model wydaje się dobrze dopasowywać do historycznych danych, z uwzględnieniem sezonowości i trendów. Jednakże, statystyki błędów prognozy wskazują na pewne ograniczenia. Wysokie wartości RMSE i Theil's U (*Rys.57*) dla zestawu testowego sygnalizują, że model może nie być całkowicie adekwatny do przewidywania przyszłych wartości, przynajmniej w kontekście danych testowych, które mogły zawierać nowe wzorce lub zmienność, nieobecną w danych treningowych.

Biorąc pod uwagę przedział ufności, który rozszerza się w czasie, można zauważyć, że niepewność prognozy wzrasta wraz z długością horyzontu czasowego. Jest to typowe dla modeli prognozowania, gdzie dalsze przewidywania są z natury bardziej niepewne.

2.7 Konstruowanie modelu AR(2)

Model po sprawdzeniu istotności współczynników:

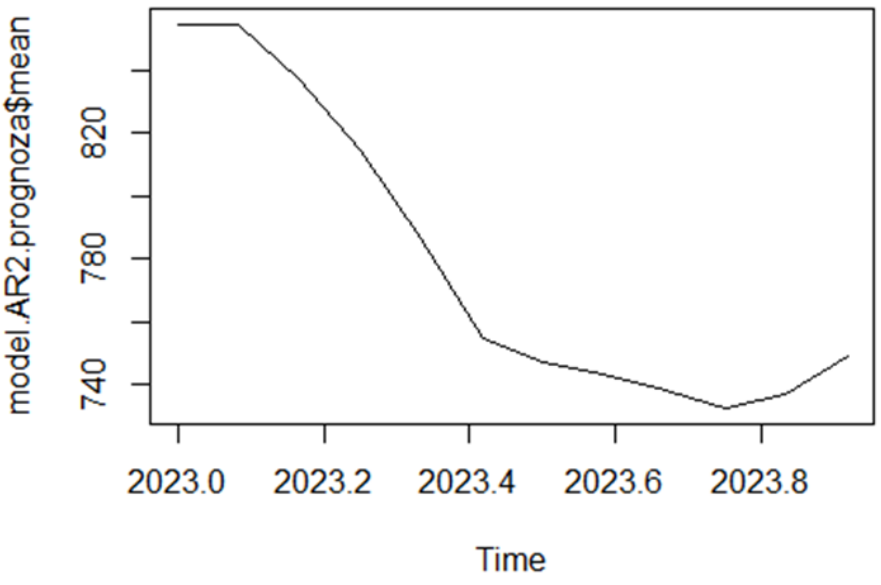
```
## Series: br.ts
## ARIMA(2,1,0)(0,1,0)[12]
##
## Coefficients:
##          ar1      ar2
##       0.9029  -0.2397
## s.e.  0.1055   0.1054
```

```
##
## sigma^2 = 245.3:  log likelihood = -345.45
## AIC=696.9   AICc=697.21   BIC=704.16
##
## Training set error measures:
##                     ME     RMSE      MAE        MPE      MAPE       MASE
## Training set 0.6529121 14.38762 7.787655 0.07976798 0.7567364 0.04725109
##                    ACF1
## Training set 0.02029003
```

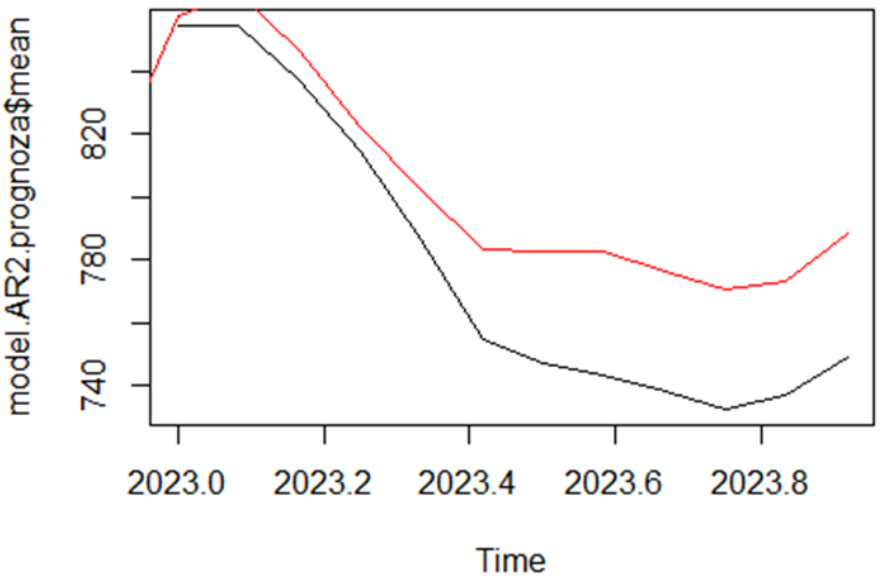
Rys.58 Wyniki analizy szeregu czasowego z zastosowaniem modelu ARIMA(2,1,0)(0,1,0) [12].

```
##           Jan      Feb      Mar      Apr      May      Jun      Jul      Aug
## 2023 854.4698 854.6803 837.4947 814.3445 786.7994 754.6020 746.7433 743.3896
##           Sep      Oct      Nov      Dec
## 2023 738.1552 732.4370 736.6289 748.7258
```

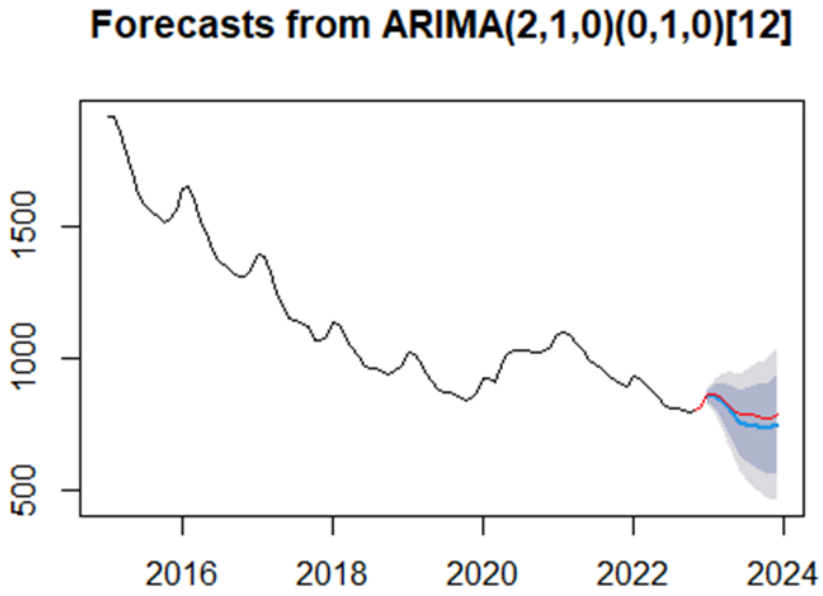
Rys.59 Prognozowane liczby osób bezrobotnych na rok 2023



Rys.60 Wykres przedstawiający liczbę osób bezrobotnych prognozowane przy pomocy modelu AR(2).



Rys.61 Porównanie liczby osób bezrobotnych prognozowanych przez model AR(2)z faktycznymi danymi (kolor czerwony) na rok 2023.



Rys.62 Porównanie prognozy przy użyciu modelu AR(2) (kolor niebieski) z faktycznymi danymi (kolor czerwony) na rok 2023.

```
##                            ME         MAE           MPE        MAPE      RMSE  Theil's U
## Training set    0.6529121    7.787655  0.07976798  0.7567364  14.38762              NA
## Test set     25.1191112 25.119111 3.19233608 3.1923361 28.74873  2.324944
```

Rys.63 Miary błędu dla zestawów treningowego i testowego

Model tak jak poprzednie modele ma problem z dokładnością na zestawie testowym. Wskazuje to na to, że faktyczny model może być bardzo skomplikowany. Wartość ME pokazuje (*Rys.63*), że model świetnie dopasowuje się do danych treningowych, niestety ma problem z danymi testowymi. Wskazywać to może na lekki overfitting modelu. Pozostałe wartości również wskazują, że model ma problem z dokładnością dla danych, których nie widział podczas treningu.

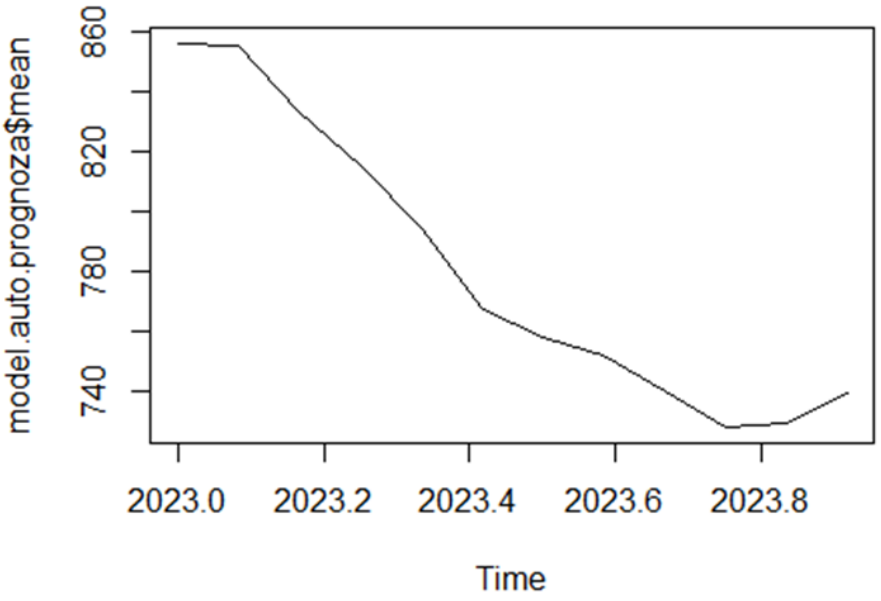
2.8 Konstruowanie modelu za pomocą automatu

```
## Series: br.ts
## ARIMA(2,1,0)(0,1,1)[12]
##
## Coefficients:
##          ar1      ar2     sma1
##       0.9812  -0.3084  -0.5303
## s.e.  0.1072   0.1062   0.1076
##
## sigma^2 = 185.6:  log likelihood = -335.39
## AIC=678.77   AICc=679.29   BIC=688.45
##
## Training set error measures:
##                     ME   RMSE      MAE       MPE      MAPE       MASE
## Training set 1.291477 12.438 7.239287 0.1348045 0.7056829 0.04392391
##                      ACF1
## Training set 0.001830219
```

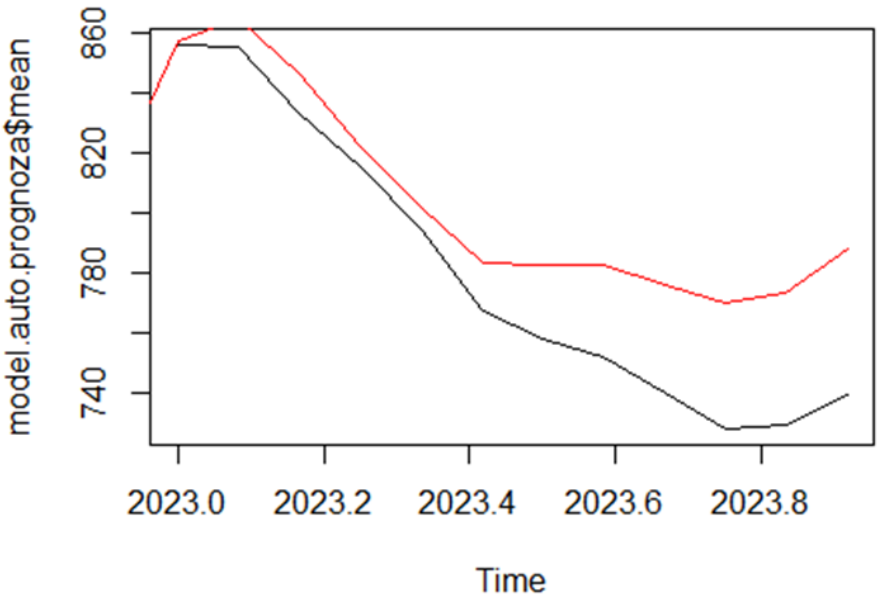
Rys.64 Wyniki analizy szeregu czasowego z zastosowaniem modelu ARIMA(2,1,0)(0,1,1) [12].

```
##             Jan       Feb       Mar       Apr       May       Jun       Jul       Aug
##  2023  856.0121  855.1457  833.4200  815.0916  794.9244  767.5511  757.8583  751.8073
##                           Sep               Oct               Nov               Dec
## 2023 739.9965 728.2042 729.4275 739.6731
```

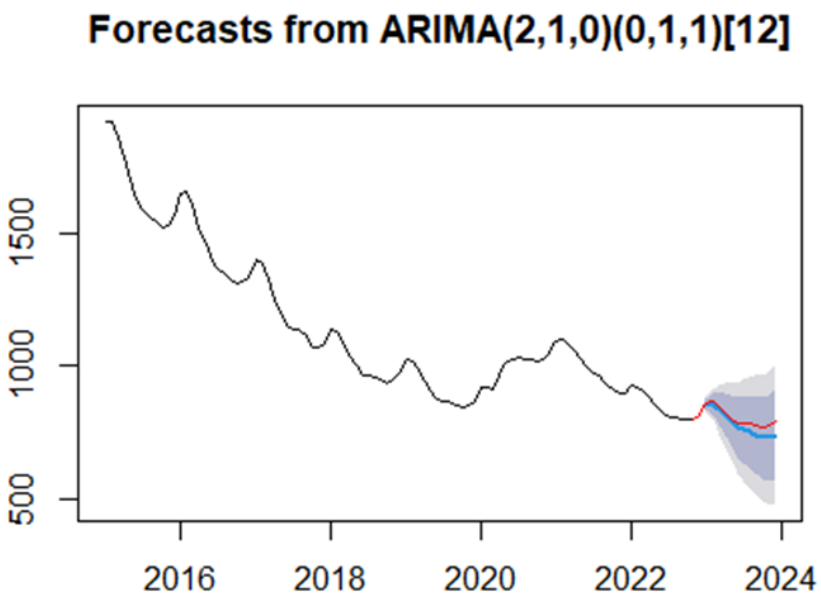
Rys.65 Prognozowana liczba osób bezrobotnych na rok 2023.



Rys.66 Prognoza liczby osób bezrobotnych na rok 2023 za pomocą modelu stworzonego przez auto.arima.



Rys.67 Porównanie prognozowanych liczby osób bezrobotnych za pomocą modelu stworzonego przez auto.arima z faktycznymi danymi na rok 2023.



Rys.68 Porównanie prognozowanych danych za pomocą modelu stworzonego przez auto.arima (kolor niebieski) z faktycznymi danymi na rok 2023(kolor czerwony).

```
##                      ME       MAE       MPE      MAPE     RMSE Theil's U
## Training set  1.291477  7.239287 0.1348045 0.7056829 12.43800        NA
## Test set     23.399024 23.399024 2.9745535 2.9745535 28.19882  2.290407
```

Rys.69 Miary błędu dla zestawów treningowego i testowego.

Miary błędu na zbiorze treningowym (ME, RMSE, MAE, MPE, MAPE, MASE) wydają się być stosunkowo niskie (*Rys.69*), co jest dobrym znakiem. Na zbiorze testowym błędy (ME, MAE, MPE, MAPE, RMSE) są większe niż na zbiorze treningowym, co jest typowe, gdy model jest stosowany do danych, których nie widział podczas treningu. Model ARIMA został skutecznie dopasowany do serii czasowej i użyty do wykonania prognozy. Miary błędów wskazują na to, że model ma pewną zdolność do generalizacji, chociaż błędy na zestawie testowym są wyższe, co sugeruje, że model może być przeuczony lub że dane testowe zawierają nowe wzorce nieobecne w danych treningowych.

2.9 Automatyczny dobór modelu dla reszt po tslm()

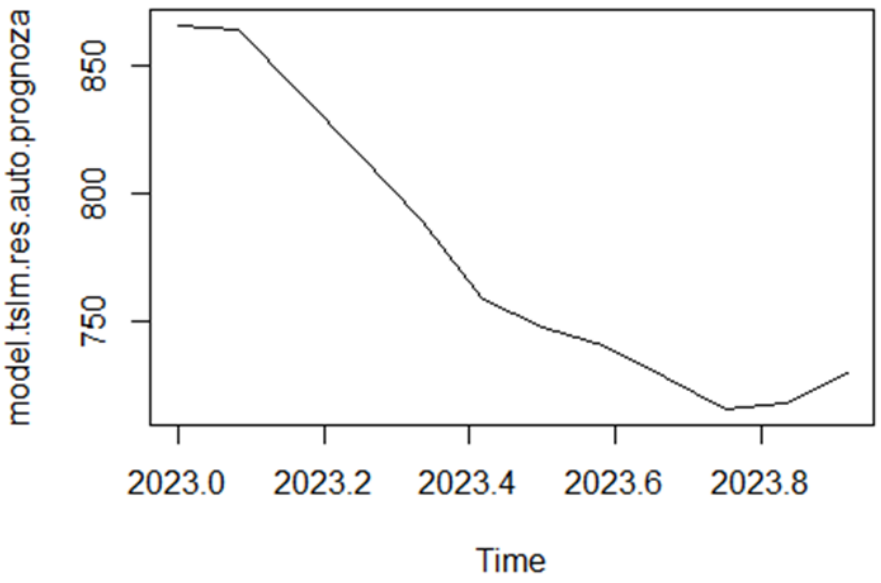
```
## Series: br.tslm.res
## ARIMA(2,2,1)(1,0,1)[12]
##
## Coefficients:
##          ar1     ar2      ma1    sar1     sma1
##       0.9144  -0.245  -0.9899  0.8186  -0.4024
## s.e.  0.1066   0.105   0.0388  0.1120   0.1713
```

```
##
## sigma^2 = 177.7:  log likelihood = -377.65
## AIC=767.3   AICc=768.26   BIC=782.55
##
## Training set error measures:
##                     ME     RMSE      MAE      MPE     MAPE       MASE
## Training set 1.072625 12.83538 8.153878 2.871972 22.36465 0.07948626
##                      ACF1
## Training set -0.009766402
```

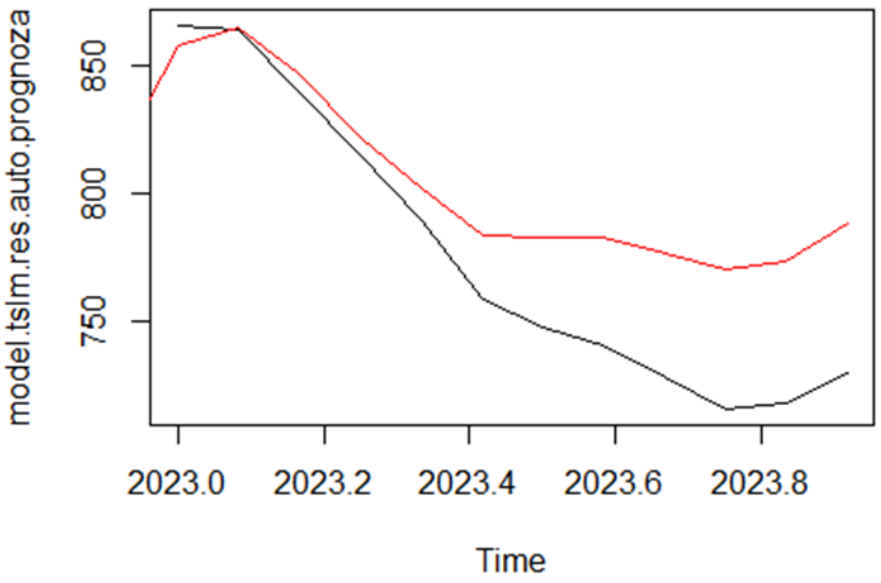
Rys.70 Wyniki analizy szeregu czasowego z zastosowaniem modelu ARIMA(2,2,1)(1,0,1) [12].

```
##             Jan       Feb       Mar       Apr       May       Jun       Jul       Aug
##  2023  865.6924  864.3678  839.5166  814.7797  789.4374  758.0917  746.9831  740.4147
##                           Sep               Oct               Nov               Dec
## 2023 728.0555 715.2446 717.3030 728.9090
```

Rys.71 Prognozowana liczba osób bezrobotnych na rok 2023



Rys.72 Prognoza liczby osób bezrobotnych na rok 2023 za pomocą modelu stworzonego przez automatyczny dobór modelu dla reszt po tslm().



Rys.73 Prognoza danych na rok 2023 za pomocą modelu stworzonego przez automatyczny dobór modelu dla reszt po tslm(). Na czerwono nałożono faktyczne dane na rok 2023.

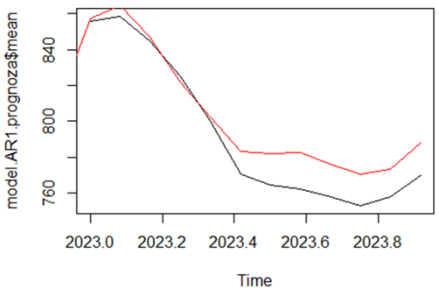
```
##          ME        MAE        MPE       MAPE       RMSE Theil's U
## 28.425384 29.774121  3.642419  3.799688 36.526864  2.965330
```

Rys.74 Miary błędu modelu

Wartość ME wynosi około 28.42 (*Rys.74*) co oznacza, że prognozy są średnio o 28.42 jednostki różne od rzeczywistych wartości. Wartość Theil's U powyżej 1 wskazuje, że model prognozy ma stosunkowo niską dokładność co oznacza że model nie jest zbyt dobry.

2.10 Ocena modeli szeregu czasowego

AR(1):

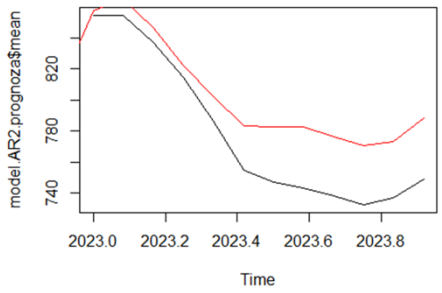


```
##                       ME       MAE         MPE     MAPE     RMSE Theil's U
## Training set  0.5734193  8.138886 0.07475213 0.787562 14.83729        NA
## Test set     10.7023824 11.247190 1.36406231 1.430349 13.35163  1.080175
```

Rys.75,76 Wykres porównujący dane wygenerowane przez model AR(1) z faktycznymi danymi oraz miary błędów tego modelu.

Wykres (*Rys.75,76*) przedstawia porównanie rzeczywistych danych czasowych (czerwona linia) z prognozami wygenerowanymi przez model autoregresyjny AR(1) (czarna linia). Linie pokazują, że model dość dobrze naśladuje kierunek zmian w danych rzeczywistych, ale istnieją widoczne rozbieżności, szczególnie w zestawie testowym, co wskazuje na możliwość przetrenowania lub niedostatecznej generalizacji modelu. Wskaźniki błędów dla zestawu treningowego są generalnie niższe niż dla zestawu testowego, co sugeruje, że model lepiej radzi sobie z danymi, na których był trenowany, niż z nowymi, nieznajomymi danymi. Wysoki błąd RMSE oznacza, że prognozy modelu mają ograniczoną dokładność.

AR(2):

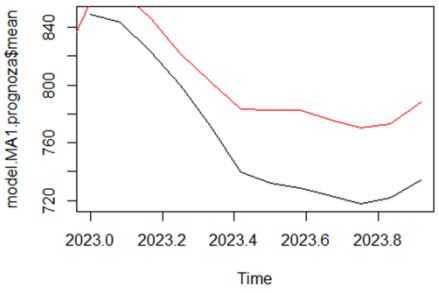



```
##                              ME         MAE           MPE        MAPE       RMSE  Theil's U
##  Training  set    0.6529121    7.787655  0.07976798  0.7567364  14.38762               NA
## Test set     25.1191112 25.119111 3.19233608 3.1923361 28.74873  2.324944
```

Rys.77,78 Wykres porównujący dane wygenerowane przez model AR(2) z faktycznymi danymi oraz miary błędów tego modelu.

Na wykresie (*Rys.77,78*) przedstawiono wyniki modelowania autoregresyjnego drugiego rzędu AR2. Analiza wskaźników błędów jak i wykresu wskazuje na wyraźną różnicę między jakością dopasowania modelu do danych treningowych a jego zdolnością do generalizacji na danych testowych. Błędy dla zestawu treningowego są stosunkowo niskie, co sugeruje dobre dopasowanie do danych, na których model był uczony. Jednak dla zestawu testowego błędy są znacząco wyższe, szczególnie RMSE oraz wartość Theil's U, która jest dużo większa niż 1, co świadczy o słabej wydajności modelu.

MA(1):

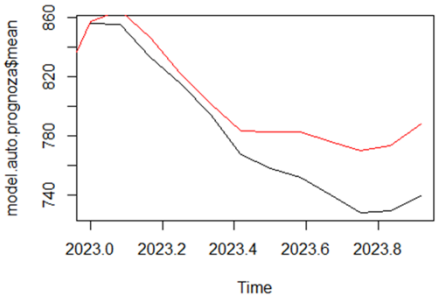



```
##                              ME          MAE          MPE       MAPE       RMSE  Theil's U
##  Training  set    1.126258    9.553194  0.1283627  0.936699  15.99122                 NA
## Test set     38.582742 38.582742 4.8786784 4.878678 41.66770  3.347574
```

Rys.79,80 Wykres porównujący dane wygenerowane przez model MA(1) z faktycznymi danymi oraz miary błędów tego modelu.

Prognozy MA(1) pokazują estymacje modelu oparte na błędach z poprzedniego okresu. Wysokie błędy (*Rys.79,80*) w zestawie testowym, szczególnie RMSE i Theil's U znacznie przekraczające 1, wskazują, że model MA(1) ma ograniczoną zdolność przewidywania nowych danych.

Model otrzymany przy użyciu automatu:



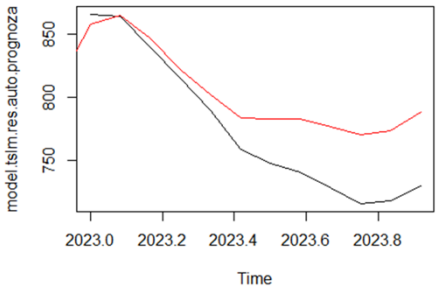

```
##                     ME       MAE       MPE      MAPE     RMSE Theil's U
## Training set  1.291477  7.239287 0.1348045 0.7056829 12.43800        NA
## Test set     23.399024 23.399024 2.9745535 2.9745535 28.19882  2.290407
```

Rys.81,82 Wykres porównujący dane wygenerowane przez model otrzymany dzięki auto.arima z faktycznymi danymi oraz miary błedów tego modelu.

Wykres (*Rys.81,82*) przedstawia prognozę uzyskaną za pomocą automatycznego modelowania ARIMA (auto.arima). Model ARIMA jest powszechnie stosowany w analizie szeregów czasowych do prognozowania przyszłych punktów danych na podstawie wcześniejszych obserwacji, łącząc autoregresję (AR), różnicowanie (I) i średnią ruchomą (MA). Wskaźniki błędów dla zestawu treningowego są akceptowalne, ale dla zestawu testowego RMSE oraz

Theil's U są wysokie, co wskazuje na istotne błędy prognozy. Wartość Theil's U powyżej 1 sugeruje, że model nie prognozuje najlepiej. To może oznaczać, że dane te są bardziej złożone i zmienne.

Automatyczny dobór modelu dla reszt po tslm():




```
##         ME        MAE        MPE      MAPE      RMSE Theil's U
## 28.425384 29.774121  3.642419  3.799688 36.526864  2.965330
```

Rys.83,84 Wykres porównujący dane wygenerowane przez dobór automatyczny modelu dal reszt po tslm() z faktycznymi danymi oraz miary błedów tego modelu.

Wykres (*Rys.83,84*) prezentuje wyniki automatycznego doboru modelu ARIMA dla reszt uzyskanych z modelu regresji sezonowej (tslm). Biorąc pod uwagę wskaźniki błędów wysokie wartości RMSE i Theil's U wskazują, że jakość prognoz modelu jest daleka od idealnej, zwłaszcza na danych testowych. Można wnioskować, że dane testowe mogą zawierać wzorce, które nie zostały uchwycone przez model.

2.11 Wnioski

Na podstawie wartości zawartych w podrozdziale 2.10 możemy wnioskować, że model AR(1) wydaje się najlepszym modelem do prognozowania. Ma on najniższe wartości MAE (Mean Absolute Error), RMSE (Root Mean Square Error) oraz Theil’s U co wskazuje na jego większą dokładność w stosunku do reszty modeli.