Trabalho Prático 1

Séries Temporais - 1/2023

Carolina Musso 18/0047850

Gabriela Carneiro de Almeida 18/0120816

Renan Menezes de Araujo

Introdução

```
pacman::p_load(Mcomp, tidyverse, forecast, flextable)
```

O pacote Mcomp disponibiliza milhares de séries de competições de previsão de séries temporais. A série apresentada nesse trabalho é a de id 2342, que se refe a uma pesquisa sobre "Manufacturers' shipments, paper and allied products", da comepetição M3. Ela fornece dados estatísticos mensais sobre as condições econômicas no setor de manufatura doméstica (empresas pequenas). A pesquisa mensura a atividade industrial atual e fornece uma indicação das tendências futuras desses tipos de negócios. Essa série mensal apresenta dados observados de Janeiro de 1983 a Agosto de 1992, e um horizonte de previsão de 18, projetando os resultados até Fevereiro de 1994.

Abaixo podemos observar os dados fornecidos nessa série, bem como seu gráfico, onde o horizonte de previsão aparece em vermelho.

```
data(M3)
id1 <- 2342
#id2 <- 1965

serie <- M3[[id1]] # serie
serie
```

```
Series: M941
Type of series: MACRO
Period of series: MONTHLY
Series description: Manufacturers' shipments, paper and allied products

HISTORICAL data
        Jan    Feb    Mar    Apr    May    Jun    Jul    Aug    Sep    Oct
1983 3281.0 3397.0 3498.5 3538.0 3449.5 3673.0 3350.5 3604.0 3673.5 3747.0
1984 3710.0 3994.5 4091.0 3954.5 4004.0 4287.0 3831.0 4046.5 4079.5 4029.5
1985 3841.5 4123.5 4133.0 3958.5 4003.0 4151.5 3723.0 3957.0 3965.5 3861.5
1986 3950.0 4140.5 4090.0 4162.0 4066.0 4358.5 4022.5 4285.5 4373.5 4284.5
1987 4181.5 4535.5 4497.0 4420.5 4370.0 4712.0 4475.0 4578.5 4751.5 4746.0
1988 4751.0 4952.5 4996.5 4998.0 4986.5 5348.0 4933.0 5263.0 5330.5 5301.0
1989 5411.5 5536.5 5613.0 5505.5 5476.0 5782.5 5283.0 5451.5 5578.0 5548.5
1990 5316.5 5505.5 5620.5 5383.5 5461.5 5658.5 5357.5 5622.0 5608.0 5604.5
1991 5221.0 5379.5 5333.0 5214.0 5206.5 5630.0 5285.5 5512.5 5592.5 5554.5
1992 5241.5 5455.0 5548.5 5375.0 5346.0 5730.5 5457.0 5603.0
        Nov    Dec
1983 3616.0 3580.5
1984 3880.0 3855.0
1985 3917.5 3704.0
1986 4077.5 4122.0
1987 4581.5 4645.5
1988 5159.0 5258.5
1989 5379.5 5117.5
1990 5399.0 5185.0
1991 5284.5 5198.5
1992

FUTURE data
        Jan    Feb    Mar    Apr    May    Jun    Jul    Aug    Sep    Oct
1992                                                        5628.5 5515.0
1993 5274.0 5382.5 5503.5 5146.5 5364.0 5608.0 5280.5 5432.5 5636.5 5442.5
1994 5114.0 5429.5
        Nov    Dec
1992 5378.0 5375.5
1993 5321.0 5195.5
1994
```

```r
plot(serie)
```

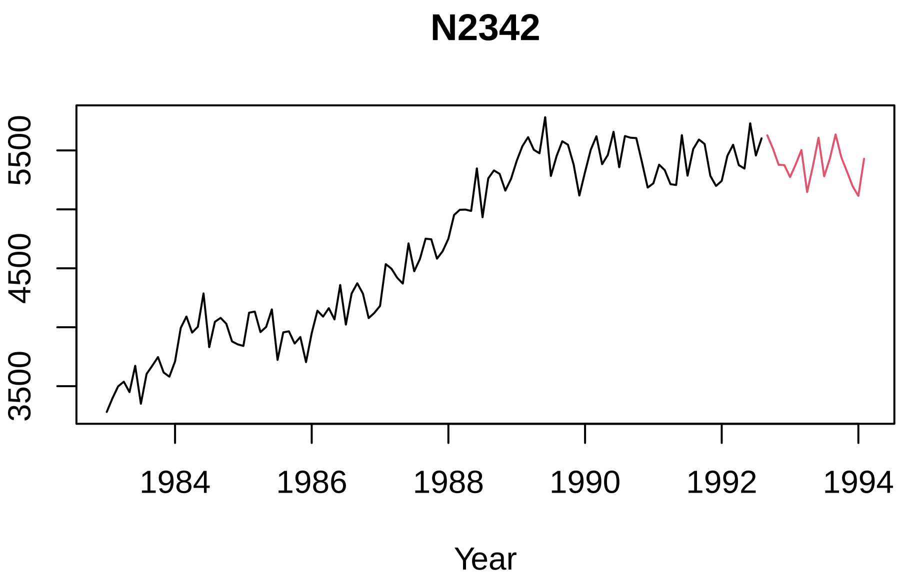
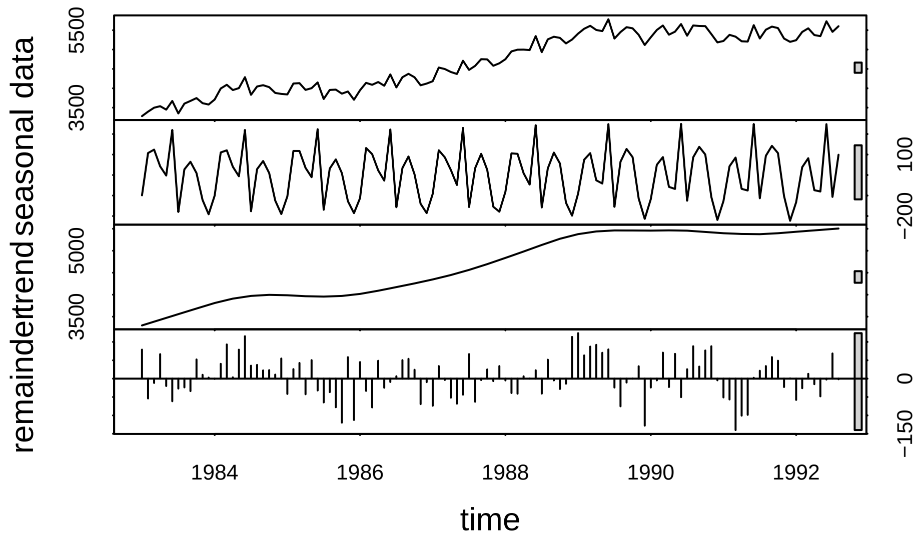


Figure 1: Série 2342

a. Decomposição da série temporal via STL (ou MSTL).

```
# obtendo os dados observados

dados <- serie$x

dados %>% stl(s.window = 7) %>%
  plot()
```

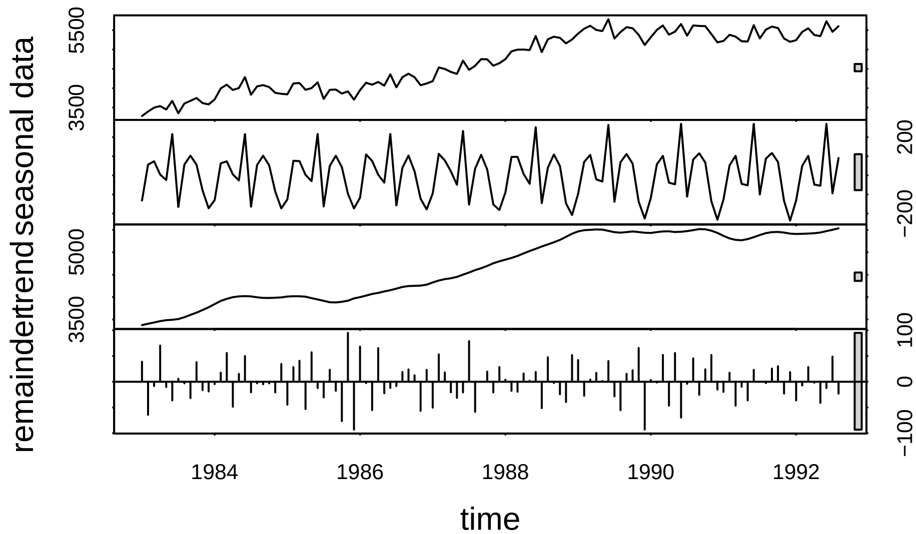


Primeiramente, a função stl() - "Seasonal an trending using Loess"- foi utilizada para decompor a série analisada com o parâmetro s.window (janela sazonal) configurado para 7, como é recomendado por Cleveland et al. (1990).

No caso específico dessa série, ela já não aparentava ter comportamentos sazonais dinâmicos muito expressivos. Assim, ao usar a função stl() com o parâmetro de janela "periodic" (para utilizar a média) ou simplesmente utilizando uma decomposição aditiva clássica da função `decompose()`, a tendência e parâmetro sazonais já estavam sendo capturados de forma clara.

Entretanto, apesar de ter sido observado que a partir dessa janela sazonal o comportamento dos resíduos já melhorava (em contraste com janelas de 3, ou 5), ainda consideramos que havia um pouco de comportamento sazonal restante nos erros. Por esse motivo foram experimentados outros tamanhos de janela de tendência, chegando-se a um valor ideal de t.window=7.

```
dados %>% stl(s.window = 7, t.window =7) %>%
  plot()
```



Agora, consideramos que os resíduos estão mais bem comportados em relação à decompsição anterior. Agora notamos uma pequena dinamicidade nos termos sazonais, sendo que os termos mais recentes tem uma amplitude levemente maior.

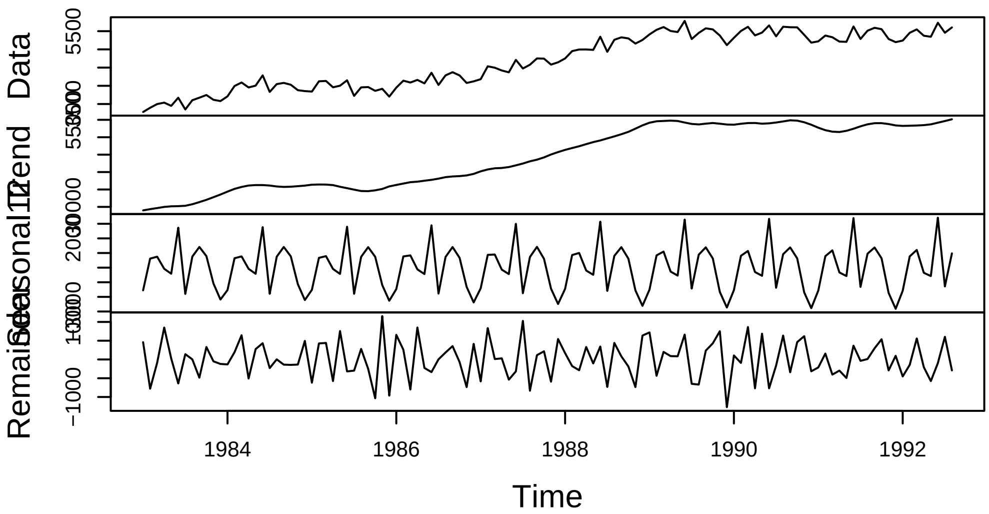
Como pode ser observado no gráfico, há uma tendência crescente clara na série, seguido de um platô a partir de 1990, já observável mesmo antes da decomposição. Uma vez que vamos considerar esses resíduos da última decompsição como compatíveis a um ruido branco, vemos que o componente de tendência foi bem capturado pela decomposição.

Há também um padrão sazonal anual na série, aque apresenta três picos a cada ano e, como dito anteriormente, parece apresentar um leve aumento na amplitude para os anos mais recentes.

Uma alternativa à função stl() seria a função mstl(), que aplica o mesmo método de forma automatizada, ou seja, sem a necessidade de setar previamente o tamanho da janela sazonal. Apesar de ele também automaticamente sugerir um tamanho para t.window(), consideramos que o melhor resultado surgiu quando ele também foi configurado manualmente para `t.window= 7`, como discutimos anteriormente. Outra vantagem da função mstl() é que ela seria capaz de identificar multiplas sazonalidades, caso ocorressem.

```
dados %>% mstl(lambda = "auto", t.window=7 ) %>% plot
```

.



O gráfico da decomposição MSTL indicou que não há multiplas sazonalidades, estando compatível com as interpretações já apresentadas acima.

b. Escolha um modelo ARIMA adequado de forma manual.

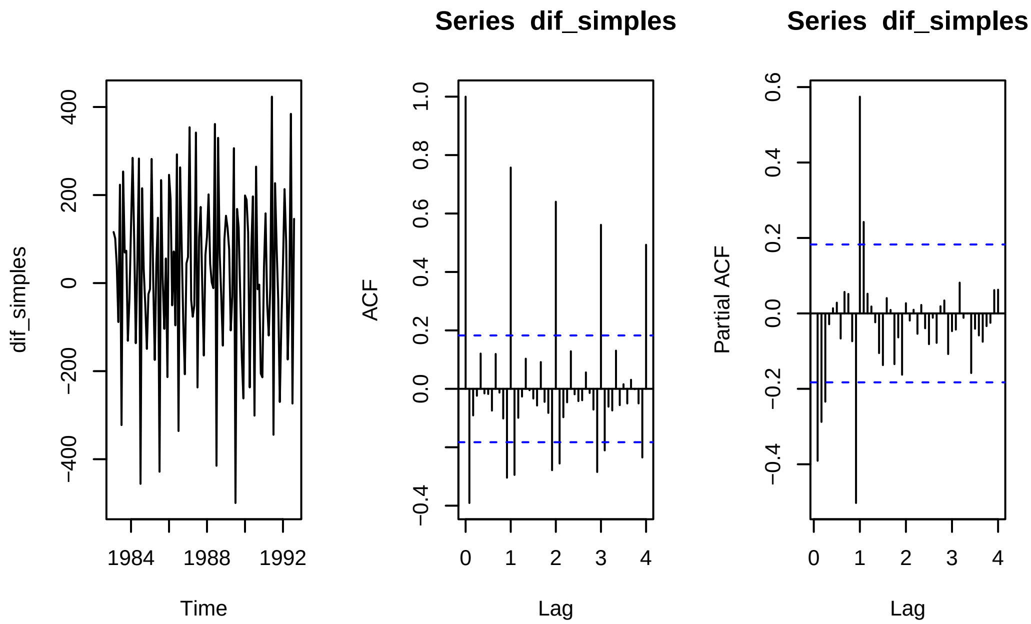
Primeiramente, vamos verificar se há necessidade de difenciações simples ou sazonais. Aplicando a função ndiffs() encontramos a necessidade de:

```
dados %>% ndiffs()
```

```
[1] 1
```

Observando o gráfico após essa diferenciação simples:

```
dif_simples <- dados %>% diff()

par(mfrow=c(1,3))
plot(dif_simples )
acf(dif_simples , lag.max = 12*4)
pacf(dif_simples , lag.max = 12*4)
```



Notamos que após essaa diferenciação simples ainda há evidência de muita autocorrelação nos termos sazonais.

Ou seja, o modelo ainda não é estacionário, e não podemos analisar esses gráficos para encontrar a ordem do modelo. Assim, procede-se com a busca pelo número de diferenciações necessárias.

```
dif_sazonal <- dif_simples %>% nsdiffs()
dif_sazonal
```

```
[1] 1
```

Resumindo então, conforme observado na tabela abaixo, a série se torna estacionária com um diferenciação simples e necessita, também, de uma diferenciação sazonal.

```
tabela <- tibble(
  var1 = dados %>% ndiffs(),
  var2 = dados %>% diff() %>% nsdiffs()

)

tabela %>%
```

```r
knitr::kable(
  format = "latex",
  align = "c",
  booktabs = TRUE,
  longtable = TRUE,
  linesep = "",
  col.names = c("Número de diferenciações simples (d)", "Número de diferenciações sazona
  ) %>%
kableExtra::kable_styling(
    position = "center",
    latex_options = c("striped", "repeat_header"),
    stripe_color = "gray!15")
```

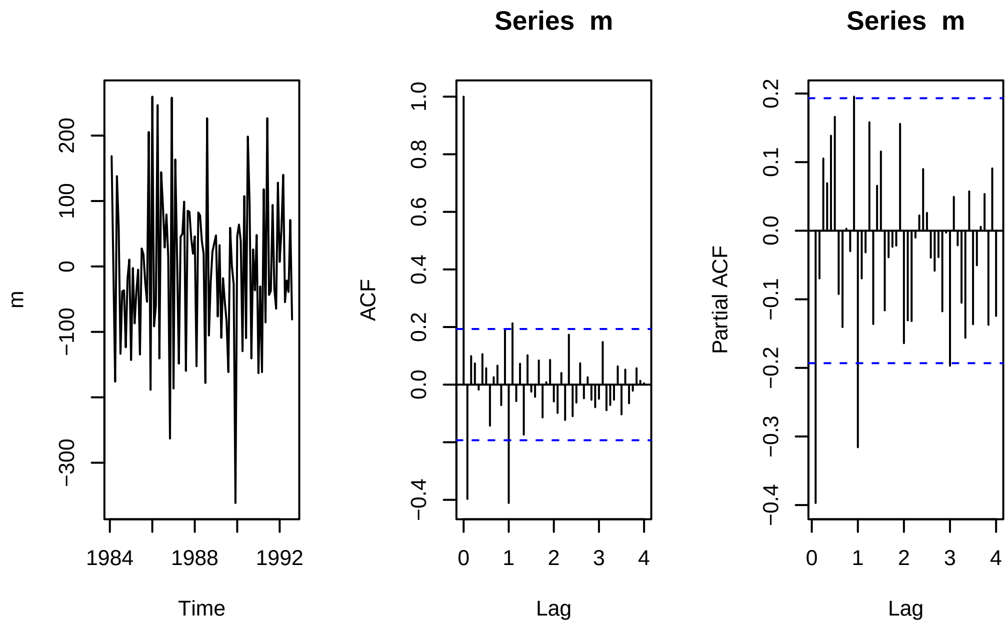
Número de diferenciações simples (d)	Número de diferenciações sazonais (D)
1	1

Assim, a série seguiria um modelo:

$$SARIMA(p, 1, q)X(P, 1, Q)_{12}$$

Aplicando então as diferenciações necessárias, podemos agora prosseguir com a análise dos gráficos ACF e PACF.

```r
m <- dados %>% diff() %>% diff(lag = 12)
par(mfrow=c(1,3))
plot(m)
acf(m, lag.max = 12*4)
pacf(m, lag.max = 12*4)
```



Olhando primeiramente os termos sazonais, eles parecem ter uma quebra na ordem 1 no ACF e um decaimento mais amortizado no PACF, o que caracterizaria um padrão de médias móveis para o modelo sazonal, ou seja com P=0, 1, Q=1.

Já a parte simples parece também seguir um padrão mais póximo ao MA (em comparação a um AR ou ARMA). Entretanto o padrão das quebras e decaimentos não está muito claro. Por esse motivo, para a busca do melhor modelo vamos fixar, d=1, D=1, P=0, p=0, e Q=1 mas vamos testar valores para q.

Testamos então combinações de p e q variando de 0 a 3.

```
melhor_AICc = Inf
for(p in 0:3){
  for(q in 0:3){
    fit = Arima(dados,order=c(p,1,q),seasonal=c(0,1,1))
    if(fit$aicc < melhor_AICc){
      melhor_AICc = fit$aicc
      cat("p =",p,", q =",q,", AICc =", fit$aicc, "\n")
    }
  }
}
```

```
p = 0 , q = 0 , AICc = 1231.266
```

```
p = 0 , q = 1 , AICc = 1225.773
p = 0 , q = 2 , AICc = 1224.901
p = 0 , q = 3 , AICc = 1224.69
p = 1 , q = 0 , AICc = 1224.411
```

Melhor menor valor do critério de informação de Akaike encontrado foi apresentado para o modelo com p=0 e q=3 (sem parâmetros autoregressivoss), ou com o modelo p=1 e q=0 (sem parâmetro de média móvel na parte simples). Assim, os modelos possíveis seriam:

$$SARIMA(0,1,3)X(0,1,1)_{12}$$

ou

$$SARIMA(1,1,0)X(0,1,1)_{12}$$

Procedemos então com o ajuste do modelo selecionado para obtenção dos parâmetros.

```r
fit1 = Arima(dados, order = c(0,1,3), seasonal = c(0,1,1))
fit1
```

```
Series: dados
ARIMA(0,1,3)(0,1,1)[12]

Coefficients:
          ma1     ma2     ma3     sma1
      -0.3485  0.1070  0.1906  -0.7767
s.e.   0.1031  0.0992  0.1180   0.1222

sigma^2 = 7182:  log likelihood = -607.04
AIC=1224.07   AICc=1224.69   BIC=1237.24
```

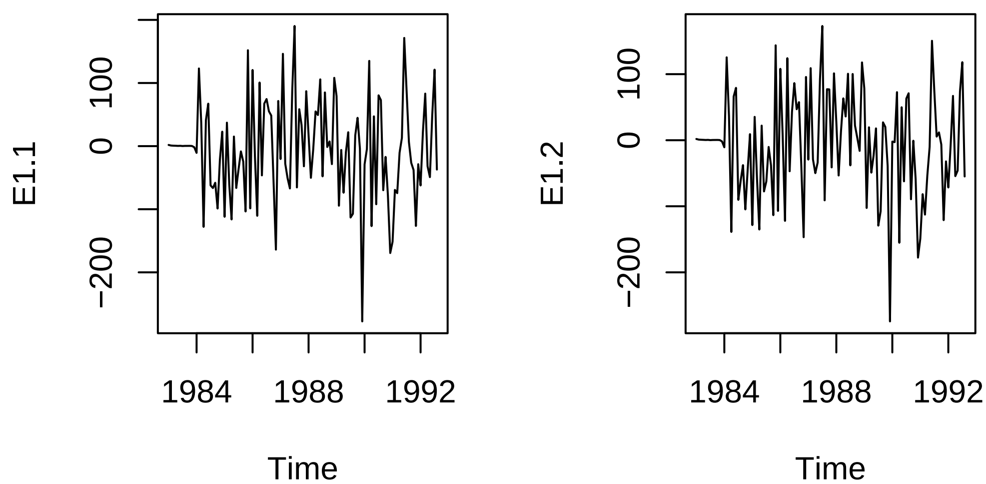
```r
fit2 = Arima(dados, order = c(1,1,0), seasonal = c(0,1,1))
fit2
```

```
Series: dados
ARIMA(1,1,0)(0,1,1)[12]

Coefficients:
          ar1     sma1
      -0.3035  -0.7976
```

```
s.e.   0.0983   0.1302

sigma^2 = 7271:  log likelihood = -609.08
AIC=1224.17   AICc=1224.41   BIC=1232.07
```

c. Análise de resíduos do modelo selecionado.

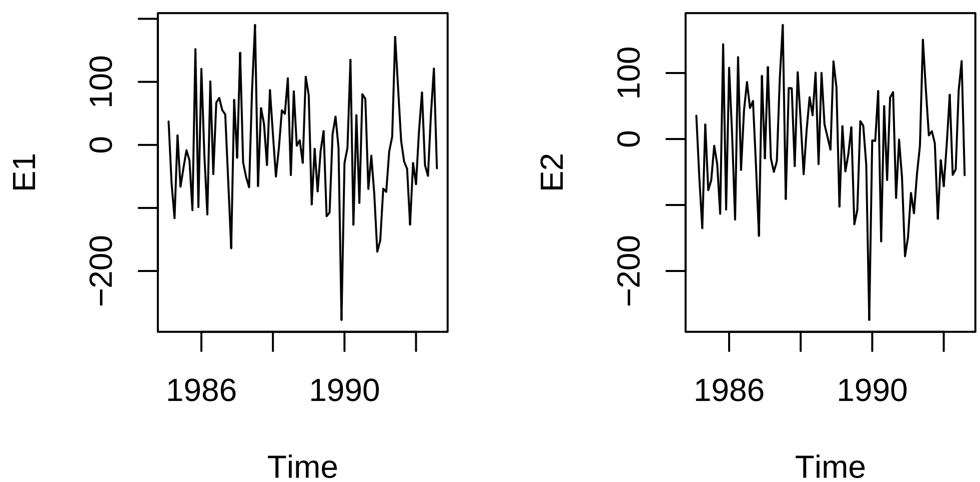
```
par(mfrow=c(1,2))
E1.1 <- fit1$residuals
plot(E1.1);# resíduos com zeros na inicialização
E1.2 <- fit2$residuals
plot(E1.2);# resíduos com zeros na inicialização
```



Por ser um modelo que requer diferenciação, nota-se que perdemos as primeiras observações dos resíduos para a inicialização. Assim, vamos analizar os resíduos após a remoção dessa primeira parte de zeros.

```
par(mfrow=c(1,2))
E1 <- fit1$residuals %>% window(start=c(1985,2))
plot(E1);# resíduos sem a inicializaçã
```

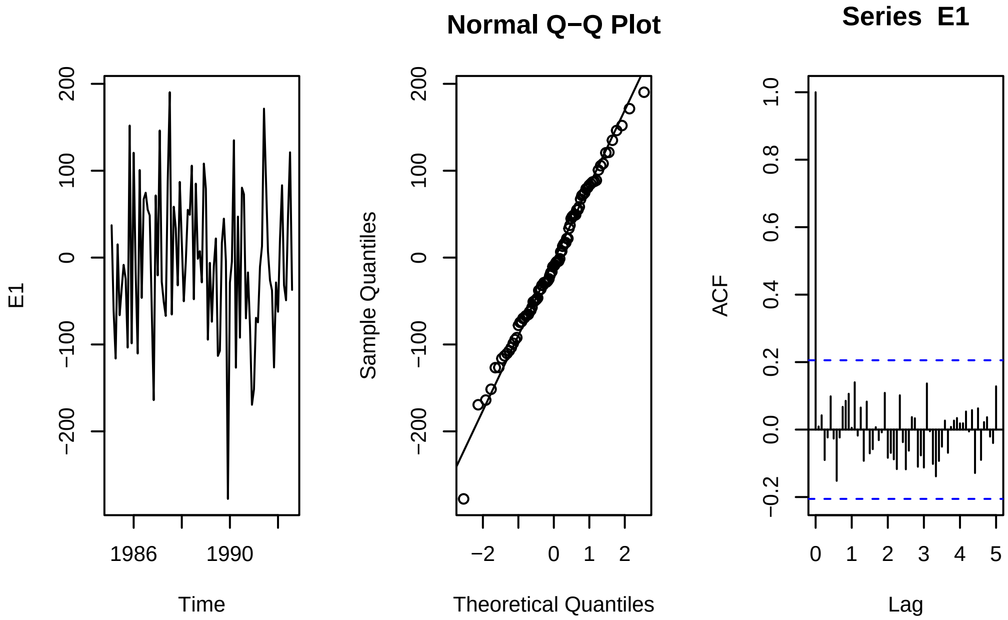
```
E2 <- fit2$residuals %>% window(start=c(1985,2))
plot(E2);# resíduos sem a inicializaçã
```



Agora observemos a análise visual dos resíduos.

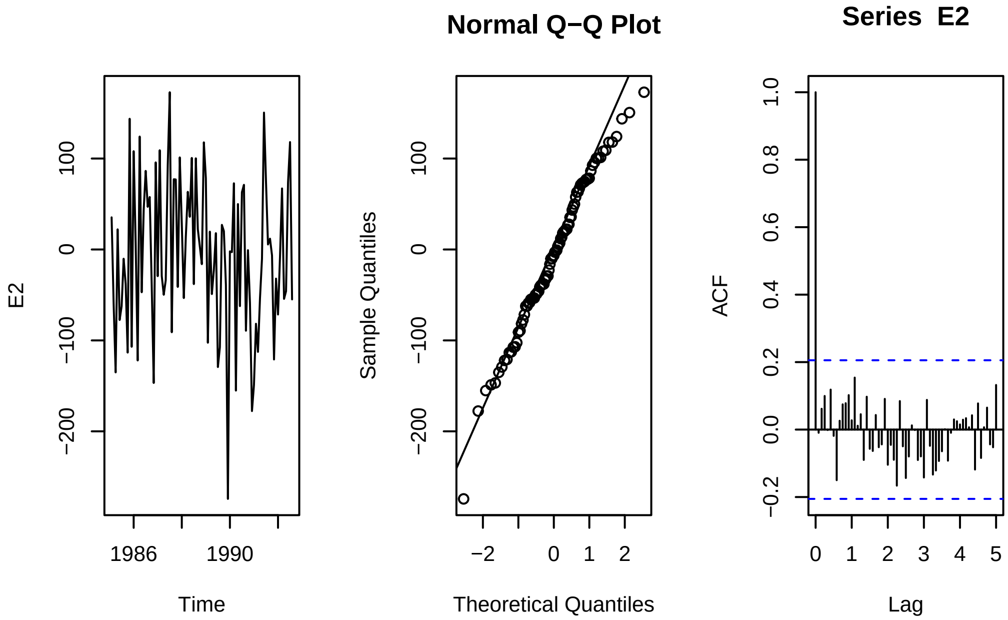
Modelo 1:

```
par(mfrow=c(1,3))
plot(E1)
qqnorm(E1); qqline(E1)
acf(E1, lag.max=12*5)
```



Modelo 2:

```
par(mfrow=c(1,3))
plot(E2)
qqnorm(E2); qqline(E2)
acf(E2, lag.max=12*5)
```



Em ambos os modelos, os resíduos em si parecem estar estacionários (primeiro gráfico), já que nenhuma tendência clara é observada, com os valores oscilando em torno do zero. O gráfico QQplot, que compara os quantis esperados da normal e os observados mostra que a distrubuição se comporta como uma distribuição normal para a maior parte dos dados, mostrando apenas alguns desvios nas caudas, principalmente à direita para o segundo modelo. Finalmente, o gráfico ACF mostra a autocorrelação apenas para o primeiro valor, como se espera, e não apresenta mais valores significativos de autocorrelação a partir daí. Isso também indica que erros estão de fato se comportando como um ruído branco. Entretanto, os resíduos parecem ter se comportado de forma mais aleatória para o primeiro modelo.

Para completar essa análise, vamos proceder com alguns testes de hipótese com significância de 5%. O Teste de KPSS para verificar estacionariedade , o teste de Ljung-Box para verificar se há autocorrelação e finalmente o teste de Shapiro-WIlk para verificar se a distribuição é compatível com a normal. Para todos esses testes rejeitaríamos a Hipótese nula se o p-valor encontrado for <0.05.

A tabela abaixo mostra os resultados dos testes:

```
tabela2 <- tibble(
  estac = c(tseries::kpss.test(E1)$p.value,
            tseries::kpss.test(E2)$p.value),
  indep = c(Box.test(E1, lag=15, type = "Ljung-Bo")$p.value,
            Box.test(E2, lag=15, type = "Ljung-Bo")$p.value),
```

```
  normlt = c(shapiro.test(E1)$p.value,shapiro.test(E2)$p.value)

)
tabela2 %>%
  knitr::kable(
    format = "latex",
    align = "c",
    booktabs = TRUE,
    longtable = TRUE,
    linesep = "",
    col.names = c("Teste KPSS - estacionariedade",
                  "Teste Box-Ljung - independência",
                  "Teste Shapiro-Wilk - normalidade"),
    ) %>%
  kableExtra::kable_styling(
      position = "center",
      latex_options = c("striped", "repeat_header"),
      stripe_color = "gray!15")
```

Teste KPSS - estacionariedade	Teste Box-Ljung - independência	Teste Shapiro-Wilk - normalidade
0.1	0.8506581	0.7572759
0.1	0.8002241	0.5495539

Ambos os modelo ajustado cumpre os pré-requisitos de estacionariedade, independência e normalidade, indicando que são moledo que pode explicar a série. Entretanto, visualmente consideramos que o primeiro modelo (que tem um AICc ligeiramente maior) apresentou resíduos mais bem comportados e está em maior conssnância com o absevado nos gráficos ACF e PACF da série diferenciada.

d. Apresente a equação do modelo selecionado.

Modelo SARIMA $(0, 1, 3)(0, 1, 1)_{12}$ definido como:

```
fit1$coef
```

```
       ma1        ma2        ma3       sma1 
-0.3484640  0.1069792  0.1905532 -0.7767125 
```

Inicialmente vamos escrever o processo w_t, após a diferenciação:

$w_t = \nabla_s \nabla x_t = \nabla_s (x_t - x_{t-1}) = x_t - x_{t-12} - x_{t-1} - x_{t-11}, t > 13$

Agora o processo w_t é um modelo ARIMA (0,0,3) x (0,0,1)

$w_t = (1 - 0.78B^{12})(1 - 0.35B + 0.11B^2 + 0.19B^3)\epsilon_t$

Modelo SARIMA (1,1,0)(0,1,1) definido como

```
fit2$coef
```

```
       ar1       sma1 
-0.3035277 -0.7976118 
```

Inicialmente vamos escrever o processo w_t, após a diferenciação:

$w_t = \nabla_s \nabla x_t = \nabla_s (x_t - x_{t-1}) = x_t - x_{t-12} - x_{t-1} - x_{t-11}, t > 13$

Agora o processo w_t é um modelo ARIMA (0,0,2) x (0,0,1)

$(1 + 0.303B)w_t = (1 - 0.797B^{12})\epsilon_t$

e. Extra

Apenas para efeito de comparação, ajustamos o modelo com a função auto-arima. Essa solução selecionou o segundo melhor modelo selecionado por esse algoritimo. O modelo com menor AICc no nosso algoritmo de fato apresenta o menor número de parâmetros e era mais parcimonioso, entretanto, o modelo equivalemnte ao selecionado pelo autoArima aparenta apresentar melhor cmportamento dos resíduos.

```
fit2 = auto.arima(dados)
fit2
```

```
Series: dados 
ARIMA(0,1,3)(0,1,1)[12] 

Coefficients:
          ma1     ma2     ma3     sma1
      -0.3485  0.1070  0.1906  -0.7767
s.e.   0.1031  0.0992  0.1180   0.1222

sigma^2 = 7182:  log likelihood = -607.04
AIC=1224.07   AICc=1224.69   BIC=1237.24
```

e. No final do relatório, inclua como um apêndice o código do R que foi utilizado.