UNIVERSIDADE DO MINHO

Séries Temporais

Docente: Raquel Menezes

2021/2022

# Projeto de Análise de Dados


Carina Ribeiro a92311

Daniela Brasileiro a92314

# Índice

# Introdução

Este projeto foi proposto pela docente Raquel Menezes no âmbito da Unidade Curricular de Séries Temporais do curso de Estatística Aplicada. Este consiste na análise de dados temporalmente correlacionados, aplicando o conhecimento adquirido ao longo desta UC e com o auxílio do software R, sendo este o principal objetivo. Iremos assim realizar a modelação e previsão de uma série temporal univariada. Os principais conceitos aplicados neste trabalho são referentes ao capítulo 4, modelação ARIMA de séries temporais.

Neste projeto tivemos a oportunidade de explorar a base de dados escolhida e tirar conclusões no contexto da mesma. Além disso, realizamos representações gráficas que consideramos oportunas e adequadas. Tudo contribuiu para uma melhor assimilação da matéria lecionada na UC em questão, bem como para o aperfeiçoamento do conhecimento sobre o programa utilizado R e também do Word, que foi o formato em que o trabalho foi realizado inicialmente. Além disso melhoramos a nossa capacidade de trabalho em grupo, criatividade e organização.

Conseguimos abordar todos os conteúdos pedidos e cumprir todas as tarefas propostas no tempo estipulado.

## Série Temporal escolhida

Para a realização deste projeto, o nosso grupo escolheu a base de dados denominada "accdeaths" da biblioteca MASS localizada no R.

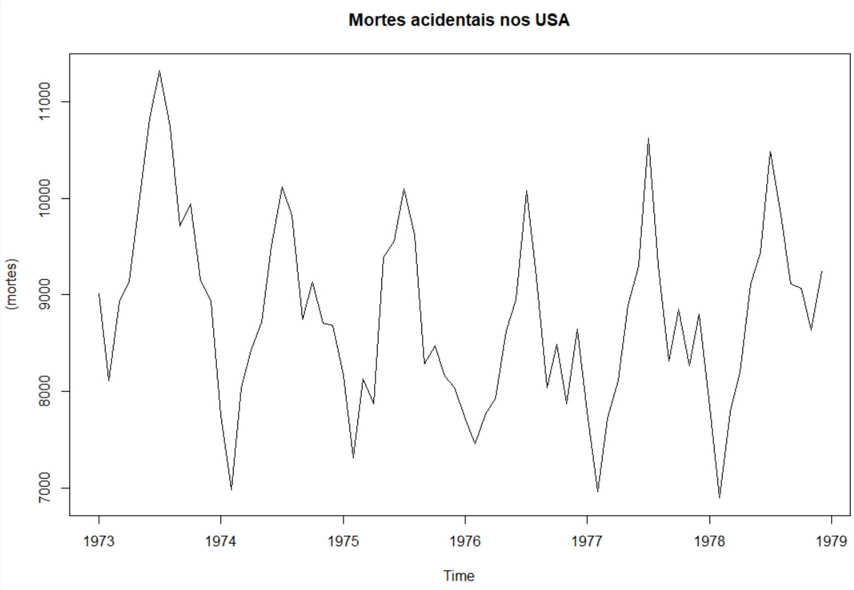
```
> accdeaths
       Jan   Feb   Mar   Apr   May   Jun   Jul   Aug   Sep   Oct   Nov   Dec
1973  9007  8106  8928  9137 10017 10826 11317 10744  9713  9938  9161  8927
1974  7750  6981  8038  8422  8714  9512 10120  9823  8743  9129  8710  8680
1975  8162  7306  8124  7870  9387  9556 10093  9620  8285  8466  8160  8034
1976  7717  7461  7767  7925  8623  8945 10078  9179  8037  8488  7874  8647
1977  7792  6957  7726  8106  8890  9299 10625  9302  8314  8850  8265  8796
1978  7836  6892  7791  8192  9115  9434 10484  9827  9110  9070  8633  9240
```

Trata-se de uma série temporal com sazonalidade e 72 observações.

A variável de interesse é uma variável discreta e representa o número de mortes acidentais nos Estados Unidos da América, sendo estes dados recolhidos mensalmente, desde 1973 até 1978.

Note-se que uma **morte acidental** é uma morte não natural causada por um acidente, um evento imprevisível como escorregões e quedas, acidentes de trânsito, envenenamento acidental, asfixia acidental ou afogamento. As mortes acidentais são diferenciadas da morte por causas naturais (doença) e de homicídios e suicídios intencionais. Uma morte acidental ainda pode ser considerada um homicídio ou suicídio se uma pessoa foi a causa não intencional.

## Gráfico da Série Temporal




- Através do gráfico vemos que à primeira vista a variância parece constante, mas a média não, pelo que a série temporal não é estacionária

-Também vemos que esta parece ter sazonalidade

# Análise de dados

- Começamos por utilizar os seguintes comandos, com o objetivo de chamar a série temporal e obter algumas informações gerais sobre a mesma (para uma melhor perceção desta):

```
> library(MASS)
> mortes<-accdeaths
> class(mortes)
[1] "ts"
> str(mortes)
 Time-Series [1:72] from 1973 to 1979: 9007 8106 8928 9137 10017 ...
> summary(mortes)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   6892    8089    8728    8789    9323   11317
```

Como visto acima, a série temporal em estudo passará a denominar-se "mortes".

# Modelo SARIMA(p,d,q)x(P,D,Q)s

Nesta fase pretende-se encontrar um modelo SARIMA (p,d,q)x(P,D,Q)s mais adequado para que posteriormente se possam fazer previsões sobre os dados em estudo.

Para encontrar este modelo iremos utilizar a Metodologia Box-Jenkins que consiste em:

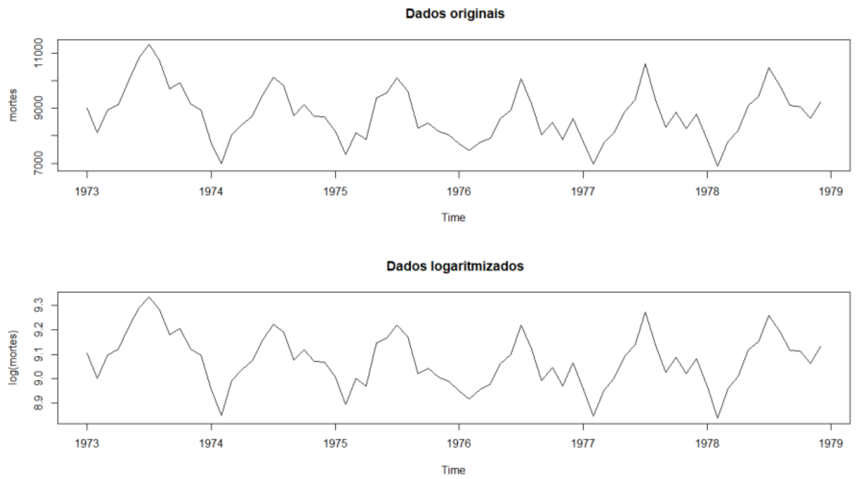
- PASSO 1: Estabilizar variância
- PASSO 2: Ordem da diferenciação d=? (estabilizar tendência)
- PASSO 3: Período s=? (detetar sazonalidade)
- PASSO 4: se s diferente de 0 então é necessário identificar as ordens sazonais D=?, P=? e Q=? (Para P=1, D=Q=0, se coeficiente AR próximo de 1, então D=1)
- PASSO 5: Ordens da parte regular p=? e q=?

## PASSO 1: Estabilizar a variância

Utilizamos o seguinte código para a visualização do gráfico dos dados originais e dos dados logaritmizados.

```
> par(mfrow=c(2,1))
> plot(mortes,main=" Dados originais")
> plot(log(mortes),main=" Dados logaritmizados")
```

Obtendo:



Pela observação dos gráficos vemos que não há grande diferença entre os dados originais e os logaritmizados pelo que podemos concluir que não há evidência para consideramos que a variância não é constante e, portanto, iremos continuar a utilizar os dados originais.

## PASSO 2: Ordem da diferenciação

Neste passo iremos descobrir qual a ordem da diferenciação (d=?) de modo a estabilizar a tendência e assim atingir a estacionariedade. A tendência dos dados pode ser visualizada no gráfico seguinte:

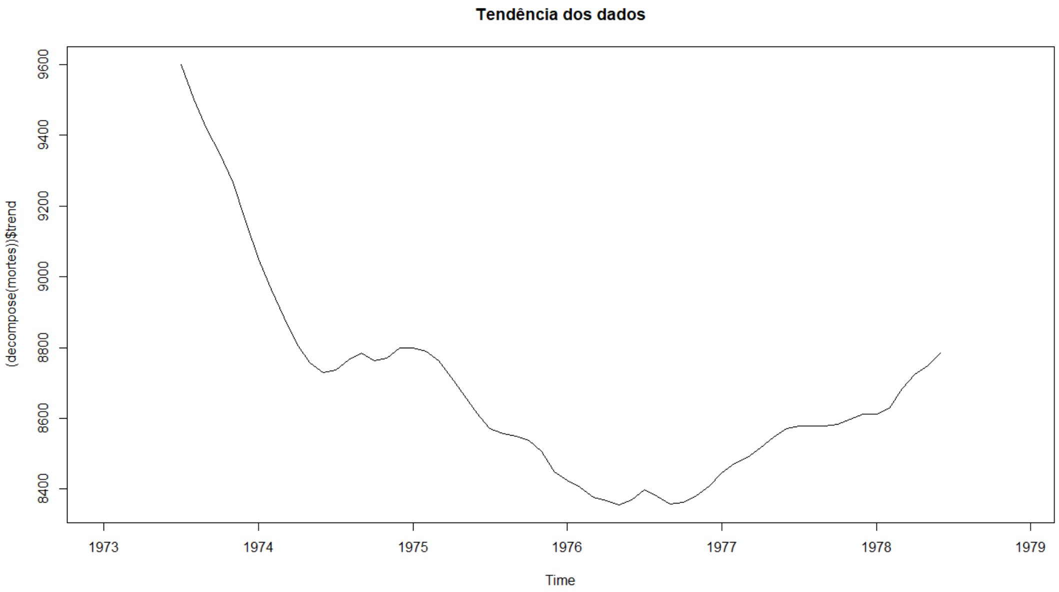


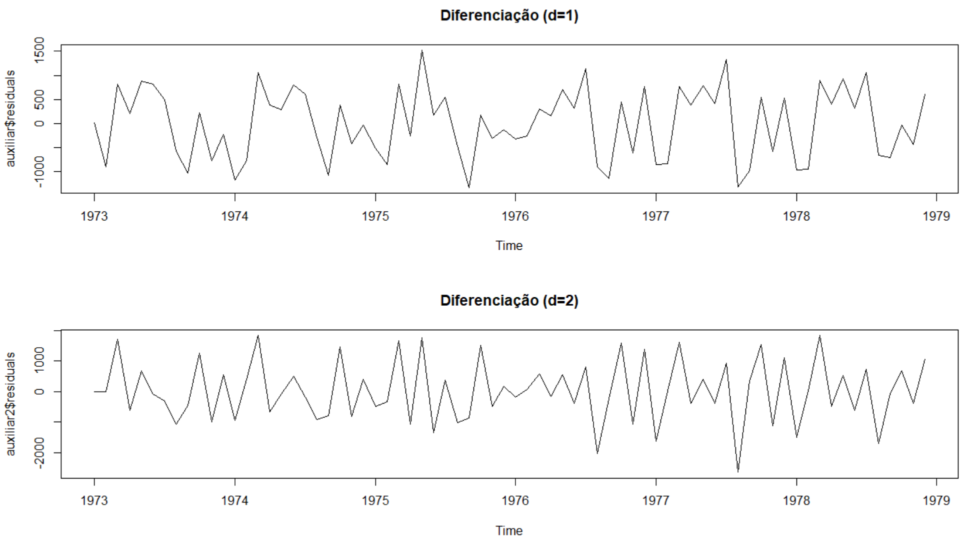

Gráfico este obtido através de:

```
> par(mfrow=c(1,1))
> plot((decompose(mortes))$trend, main="Tendência dos dados")
```

Para estabilizar esta tendência e torná-la assim constante utilizamos os seguintes comandos, que são os modelos com d=1 e d=2 e a sua comparação gráfica.

```
> par(mfrow=c(2,1))
> auxiliar = arima(mortes, order=c(0,1,0), seasonal=list(order=c(0,0,0), period=0))
> plot(auxiliar$residuals, main="Diferenciação (d=1)")
> auxiliar2 = arima(mortes, order=c(0,2,0), seasonal=list(order=c(0,0,0), period=0))
> plot(auxiliar2$residuals, main="Diferenciação (d=2)")
```

Obtendo:



Através das representações gráficas observamos que para ambas as ordens de diferenciação a tendência foi de facto estabilizada. Vemos também que existem algumas diferenças entre d=1 e para d=2 pelo que decidimos utilizar as duas e no final concluir qual o modelo mais adequado.

## PASSO 3: Cálculo do período (detetar a sazonalidade)

Para detetar a sazonalidade aplicamos as seguintes instruções no R:

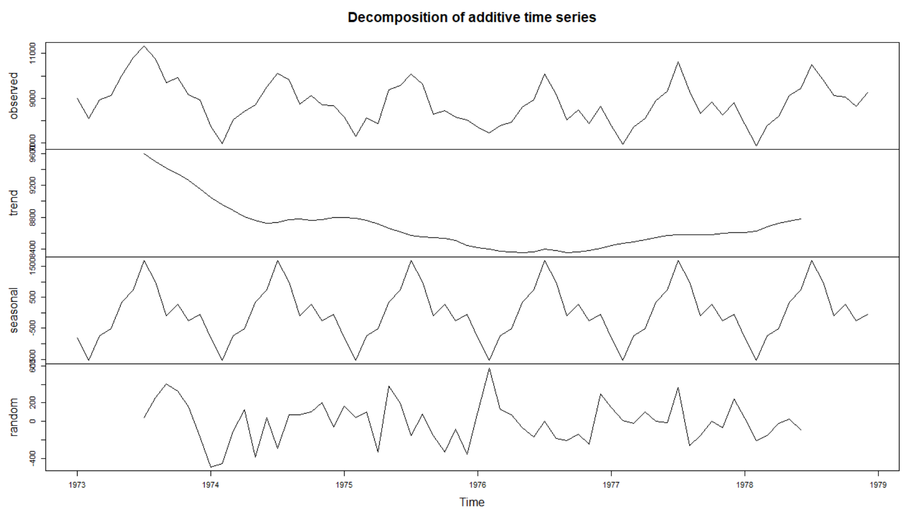
```
> periodogram=spectrum(as.vector(mortes), plot=F)
> imax=which.max(periodogram$spec)
> periodo=1/periodogram$freq[imax]
> periodo
[1] 12
```

Concluímos que de facto a série temporal tem sazonalidade e o seu período é igual a 12, ou seja, esta tem sazonalidade com período igual a 12 meses.

Podemos também verificar o período através do comando decompose()

Obtendo:

```
> plot(decompose(mortes))
```




No gráfico da parte sazonal (terceiro gráfico) podemos ver que ocorre sazonalidade com período de 1 ano (12 meses).

Visto que o período é 12 meses, achamos relevante ver em que mês do ano ocorrem os picos de mortes acidentais. Para isso utilizamos os códigos:

```
> #Ver em que mês ocorrem os picos
> which(mortes[1:12]==max(mortes[1:12]))
[1] 7
> which(mortes[12:24]==max(mortes[12:24]))
[1] 8
> which(mortes[24:36]==max(mortes[24:36]))
[1] 8
> which(mortes[36:48]==max(mortes[36:48]))
[1] 8
> which(mortes[48:60]==max(mortes[48:60]))
[1] 8
> which(mortes[60:72]==max(mortes[60:72]))
[1] 8
```

Concluímos que na maioria dos anos os picos ocorrem no mês de agosto (mês 8) com exceção do ano de 1973 em que o pico de mortes acidentais ocorreu no mês de julho (mês 7). Pensamos que estes valores se devem provavelmente à maior propensão para afogamentos no mês de agosto, aumentando assim o número de mortes acidentais.

## Primeiramente vamos considerar a diferenciação com d=1.

### PASSO 4: Calcular as ordens da parte sazonal

Primeiramente iremos descobrir o valor de D.

Para P=1 e D=Q=0, realizamos a seguinte instrução do R:

Através do output sabemos que o coeficiente tem um valor de 0.8579, valor este próximo de 1, pelo que concluímos que D=1, útil para efetuar os modelos vizinhos.

*** (explicação detalhada da primeira linha da tabela)**

Para as restantes linhas, iremos selecionar o melhor modelo utilizando o critério de informação de Akaike (AIC = Akaike information Criteria).

| Modelo | Coeficiente | Erro(S.E) | Significância | AIC |
|---|---|---|---|---|
| * $\mathbf{ARIMA(0,1,0)\times(1,0,0)_{s=12}}$ | Sar1: 0.8579 | 0.0481 | Coeficiente próximo de 1 | ______ |
| $\mathbf{ARIMA(0,1,0)\times(1,1,0)_{s=12}}$ | Sar1: -0.3491 | 0.1207 | Significativo | 868.3 |
| $\mathbf{ARIMA(0,1,0)\times(0,1,1)_{s=12}}$ | Sma1:-0.5665 | 0.1650 | Significativo | 864.13 |
| $\mathbf{ARIMA(0,1,0)\times(1,1,1)_{s=12}}$ | Sar1: 0.2850<br>Sma1:-0.9996 | 0.1607<br>0.7138 | Não significativo, aceita-se $H_0: \alpha = 0$ e $H_0: \beta = 0$ | ______ |
| $\mathbf{ARIMA(0,1,0)\times(2,1,0)_{s=12}}$ | Sar1: -0.4617<br>Sar2: -0.2928 | 0.1320<br>0.1415 | Significativo | 866.48 |
| $\mathbf{ARIMA(0,1,0)\times(0,1,2)_{s=12}}$ | Sma1: -0.5935<br>Sma2: -0.1738 | 0.2755<br>0.2080 | Não significativo, aceita-se $H_0: \beta_2 = 0$ | ______ |

Escolhemos assim o modelo com menor AIC :

$$\boldsymbol{ARIMA(p,1,q)\times(0,1,1)_{s=12}}$$

## PASSO 5: Calcular as ordens da parte regular

Novamente, iremos selecionar o melhor modelo utilizando o critério de informação de Akaike (AIC = Akaike information Criteria).

| Modelo | Coeficiente | Erro(S.E) | Significância | AIC |
|---|---|---|---|---|
| $\mathbf{ARIMA(0,1,0)\times(0,1,1)_{s=12}}$ | Sma1:-0.5665 | 0.1650 | Significativo | 864.13 |
| $\mathbf{ARIMA(1,1,0)\times(0,1,1)_{s=12}}$ | Ar1: -0.3304<br>Sma1: -0.5893 | 0.1225<br>0.1807 | Significativo | 859.28 |
| $\mathbf{ARIMA(0,1,1)\times(0,1,1)_{s=12}}$ | Ma1: -0.4303<br>Sma1: -0.5528 | 0.1228<br>0.1784 | Significativo | 856.88 |
| $\mathbf{ARIMA(1,1,1)\times(0,1,1)_{s=12}}$ | Ar1:0.0979<br>Ma1: -0.5109<br>Sma1: -0.5437 | 0.3111<br>0.2736<br>0.1784 | Não significativo, aceita-se $H_0: \phi = 0$ | ______ |
| $\mathbf{ARIMA(2,1,0)\times(0,1,1)_{s=12}}$ | Ar1: -0.4016<br>Ar2: -0.2117<br>Sma1: -0.5498 | 0.1272<br>0.1268<br>0.1768 | Não significativo, aceita-se $H_0: \phi_2 = 0$ | ______ |
| $\mathbf{ARIMA(0,1,2)\times(0,1,1)_{s=12}}$ | Ma1: -0.4081<br>Ma2: -0.0518<br>Sma1: -0.5417 | 0.1341<br>0.1478<br>0.1782 | Não significativo, aceita-se $H_0: \Theta_2 = 0$ | ______ |

Escolhemos assim o modelo com menor AIC :

$$\boldsymbol{ARIMA(0,1,1)\times(0,1,1)_{s=12}}$$

## Qualidade do ajustamento - Análise dos resíduos

Neste passo vamos verificar se os resíduos são ruídos brancos, ou seja, se cumprem os seguintes pressupostos:

- $\epsilon_t$ são independentes (não correlacionados);
- $\epsilon_t$ seguem a distribuição normal;
- A média de $\epsilon_t$ é zero ($\underline{E[\epsilon_t] = 0}$);
- A variância de $\epsilon_t$ é constante, ou seja, homocedástica.

Começamos por denominar o modelo escolhido por modelo1.

```
> modelo1 <- arima(mortes, order=c(0,1,1), seasonal=list(order=c(0,1,1), period=12))
> names(modelo1)
 [1] "coef"      "sigma2"    "var.coef"  "mask"      "loglik"    "aic"
 [7] "arma"      "residuals" "call"      "series"    "code"      "n.cond"
[13] "nobs"      "model"
>
> #variância estimada dos erros
> modelo1$sigma2
[1] 99346.89
```

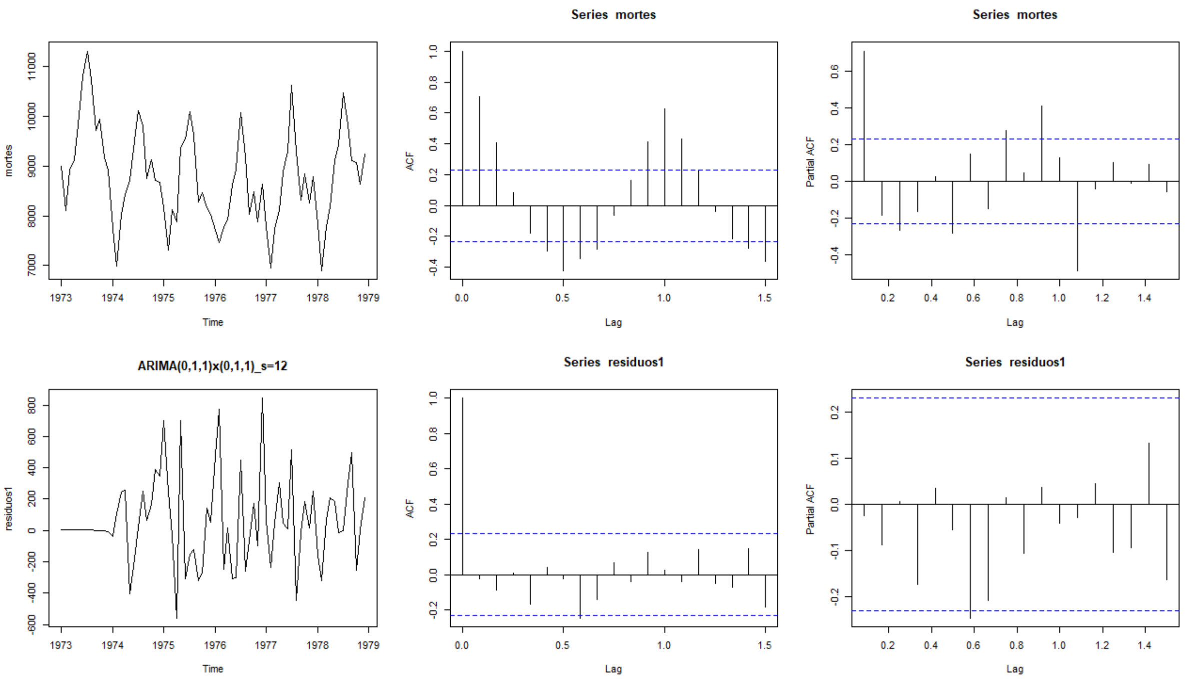
Através do output observamos que a variância estimada dos erros do modelo escolhido é 99346.8.

❖ Independência dos resíduos

Começamos por denominar os resíduos do modelo de residuos1 e de seguida realizamos os gráficos destes, e das respetivas funções de autocorrelação total e parcial. Para tal, realizamos os comandos:

```
> residuos1=modelo1$residuals
> par(mfrow=c(2,3))
> plot(mortes)
> acf(mortes)
> pacf(mortes)
> # testar independência dos resíduos
> plot(residuos1, main="ARIMA(0,1,1)x(0,1,1)_s=12")
> acf(residuos1)
> pacf(residuos1)
```

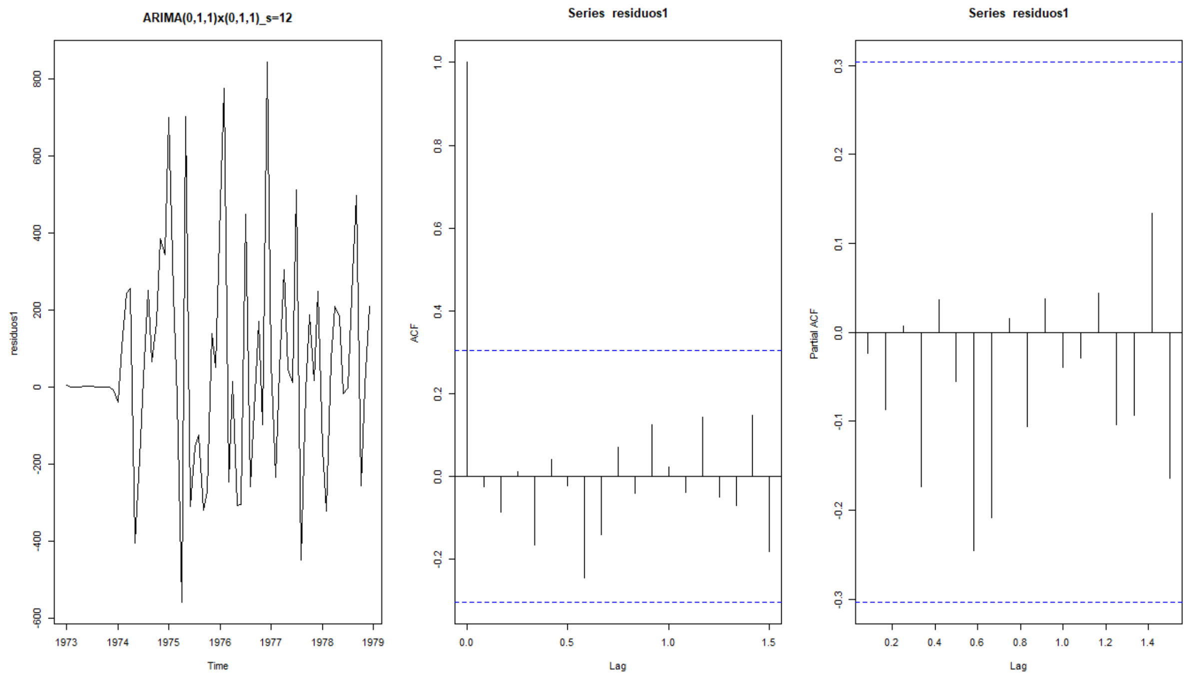
Obtendo:




No quinto gráfico está representada a FAC dos resíduos e no sexto a FACP dos mesmos. Destes gráficos podemos concluir que não há evidencias para rejeitar a independência dos resíduos, logo este pressuposto está verificado.

De forma a poder observar melhor que os resíduos são independentes vamos considerar o grau de significância de 99%.

```
> par(mfrow=c(1,3))
> plot(residuos1, main="ARIMA(0,1,1)x(0,1,1)_s=12")
> acf(residuos1,ci=0.99)
> pacf(residuos1,ci=0.99)
```




Também podemos verificar a independência dos resíduos recorrendo seguinte teste:

```
> Box.test(residuos1)

        Box-Pierce test

data:  residuos1
X-squared = 0.041252, df = 1, p-value = 0.8391
```

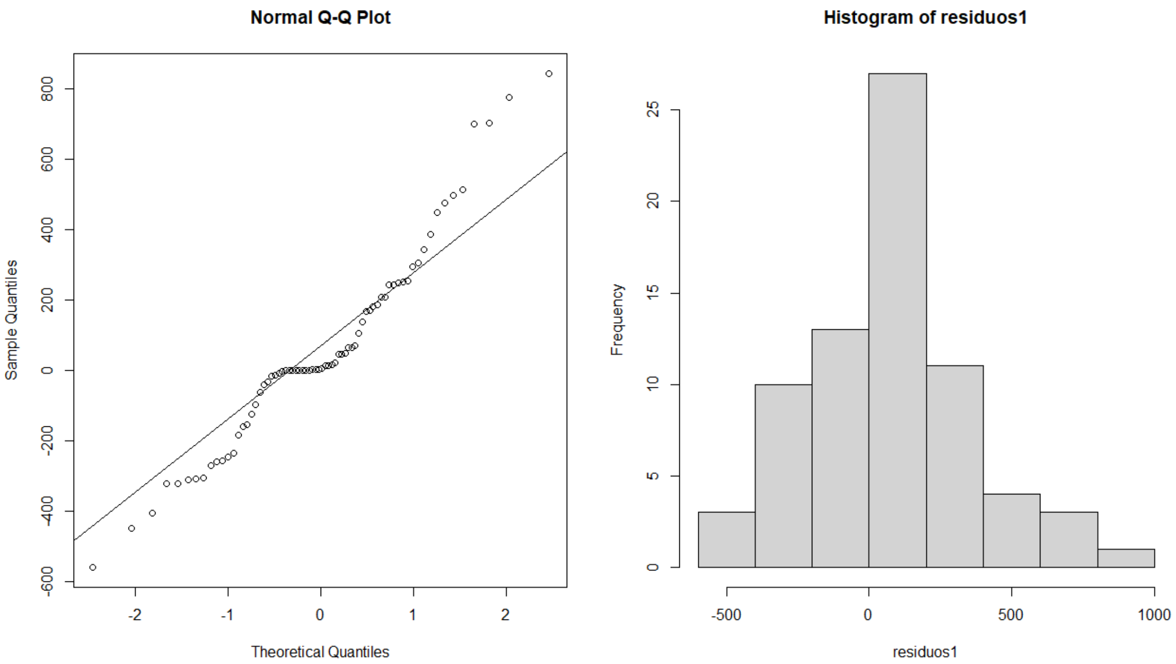
Sendo p-value=0.8391 aceita-se a hipótese nula $(H_0: \varepsilon_t\ são\ independentes)$ para os níveis de significância usuais (5% e 10%), logo concluímos que os resíduos são independentes.

- Verificar se os resíduos seguem a distribuição normal

```
> library(nortest)
> lillie.test(residuos1)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  residuos1
D = 0.1339, p-value = 0.002693

> qqnorm(residuos1)
> qqline(residuos1)
> hist(residuos1)
```

Nos comandos anteriores recorremos à biblioteca nortest para podermos utilizar o teste Lilliefors (Kolmogorov-Smirnov), sendo o teste utilizado para testar a normalidade em amostras grandes (mais de 30 observações). Observar que o valor p é igual a 0.002693 pelo que rejeitamos a hipótese nula $(H_0: \varepsilon_t\ segue\ um\ distribuição\ normal)$ para os níveis de significância usuais, ou seja, os resíduos não seguem uma distribuição normal.

Visualizamos também os seguintes gráficos:




Como é possível ver nas representações gráficas os resíduos não seguem uma distribuição normal, obtendo assim mesma conclusão que anteriormente.

Os resíduos não verificam este pressuposto.

- Média dos resíduos

Para testar se a média dos resíduos tem o valor de 0 utilizamos o teste de T de Student, uma vez que é um teste paramétrico para testar a média. Utilizamos os comandos:

```
> t.test(residuos1)

        One Sample t-test

data:  residuos1
t = 1.7754, df = 71, p-value = 0.08011
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
  -7.240938 124.911706
sample estimates:
mean of x
 58.83538
```

Obtém-se o valor p 0.08011 não rejeitando a hipótese nula, $H_0: E[\varepsilon_t] = 0$, concluindo assim que a média dos resíduos é igual a zero para os níveis de significância usuais.

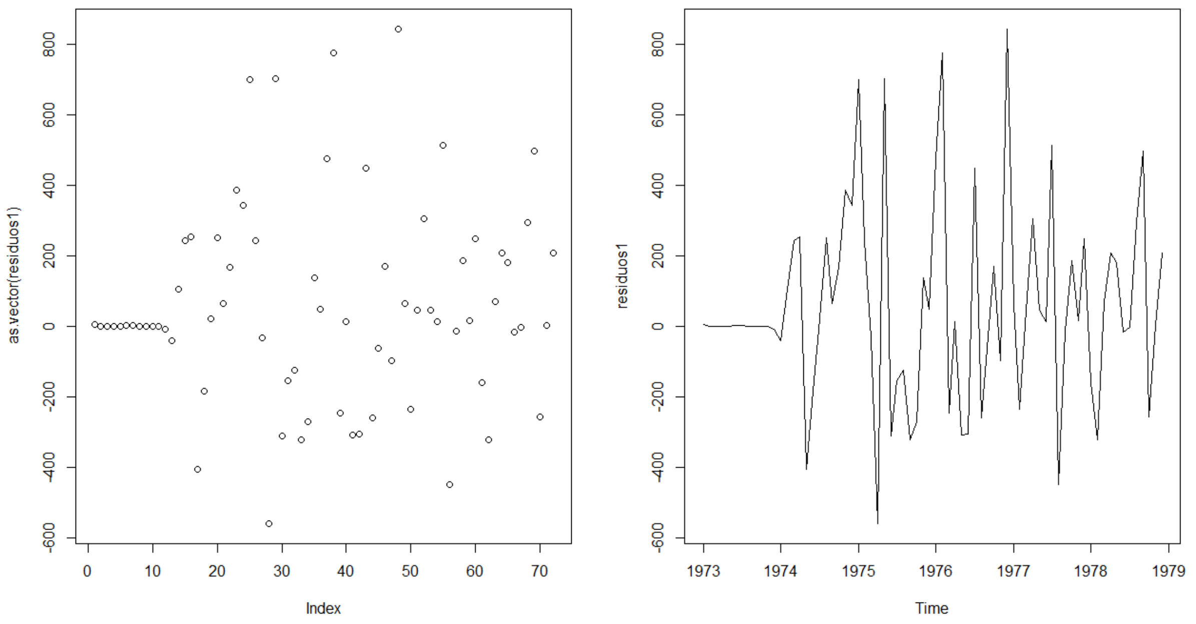
Os resíduos verificam este pressuposto.

❖ Variância dos resíduos

Para verificar se a variância dos resíduos é constante obtemos a representação gráfica dos resíduos através dos comandos:

```
> par(mfrow=c(1,2))
> plot(as.vector(residuos1))
> plot(residuos1)
```

Obtendo:

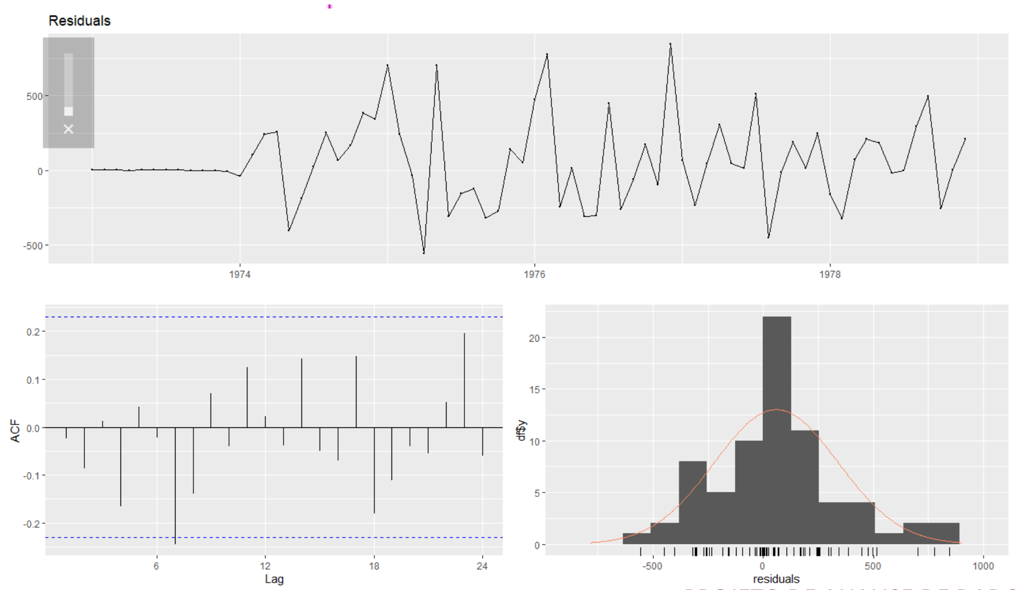



Concluímos que não há evidências para considerar que a variância não é constante uma vez que não se regista qualquer padrão nos gráficos anteriores. Existindo assim homocedasticidade.

Os resíduos verificam este pressuposto.

Além do suprarreferido também podemos utilizar os seguintes comandos para ter uma ideia geral do comportamento dos resíduos e confirmar o que já tínhamos concluído.

```
> library(forecast)
> checkresiduals(residuos1)
```



# Previsões

Apesar dos resíduos do modelo não validarem um dos pressupostos (normalidade) iremos fazer previsões utilizando o mesmo.

- **De modo a validar ou não modelo, iremos ignorar as últimas 24 observações (2 anos) e fazer previsões para esse período ignorado comparar os valores estimados com os observados. Ou seja, estimar $x_{49}, \dots, x_{71}, x_{72}$ .**

Começamos por denominar o modelo escolhido de m1 e a atribuir um nome às previsões.

```
> m1 = arima(mortes[1:48], order=c(0,1,1), seasonal=list(order=c(0,1,1), period=12))
> previsoes <- predict(m1, n.ahead = 24)
> mortes[49:72]
> previsoes$pred
> previsoes$se
```

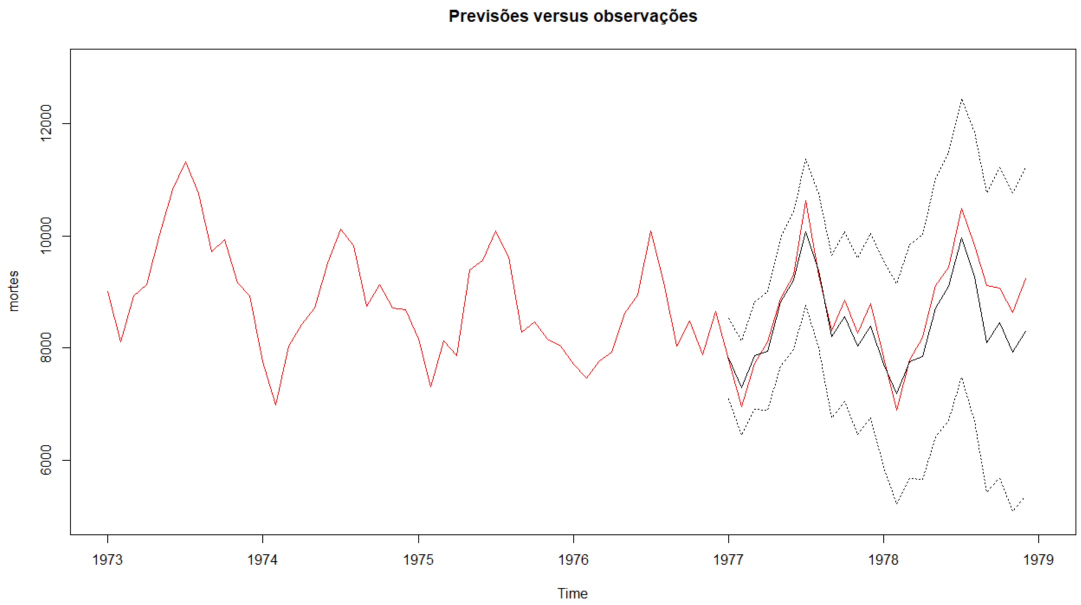
Obtemos para os valores previstos e respetivos erros:

| Mês | Valores observados | Valores previstos | Erro |
|---|---|---|---|
| **Jan 1977** | 7792 | 7814.650 | 365.5615 |
| **Fev 1977** | 6957 | 7290.022 | 430.2666 |
| **Mar 1977** | 7726 | 7862.383 | 486.4395 |
| **Abril 1977** | 8106 | 7948.237 | 536.7657 |
| **Maio 1977** | 8890 | 8817.406 | 582.7620 |
| **Jun 1977** | 9299 | 9199.222 | 625.3844 |
| **Jul 1977** | 10625 | 10063.322 | 665.2817 |
| **Agos 1977** | 9302 | 9374.924 | 702.9180 |
| **Set 1977** | 8314 | 8200.390 | 738.6392 |
| **Out 1977** | 8850 | 8556.643 | 772.7107 |
| **Nov 1977** | 8265 | 8033.055 | 805.3421 |
| **Dez 1977** | 8796 | 8396.007 | 836.7019 |
| **Jan 1978** | 7836 | 7710.393 | 938.4360 |
| **Fev 1978** | 6892 | 7185.764 | 1001.1110 |
| **Mar 1978** | 7791 | 7758.125 | 1060.0869 |
| **Abril 1978** | 8192 | 7843.979 | 1115.9504 |
| **Maio 1978** | 9115 | 8713.149 | 1169.1477 |
| **Jun 1978** | 9434 | 9094.964 | 1220.0277 |
| **Jul 1978** | 10484 | 9959.064 | 1268.8690 |
| **Agos 1978** | 9827 | 9270.667 | 1315.8988 |
| **Set 1978** | 9110 | 8096.132 | 1361.3048 |
| **Out 1978** | 9070 | 8452.385 | 1405.2444 |
| **Nov 1978** | 8633 | 7928.797 | 1447.8511 |
| **Dez 1978** | 9240 | 8291.749 | 1489.2394 |

Para obter o gráfico da série original com as previsões e os intervalos de confiança a 95% utilizamos o comando:

```
> ts.plot(mortes, xlim=c(1973,1979), ylim=c(5000,13000), col="red", main="Previsões versus observações")
> lines(ts(previsoes$pred, start=1977, deltat=1/12))
> low = previsoes$pred - 1.96*previsoes$se
> upper = previsoes$pred + 1.96*previsoes$se
> lines(ts(low, start=1977, deltat=1/12), lty=3)
> lines(ts(upper, start=1977, deltat=1/12), lty=3)
```

O gráfico obtido foi:



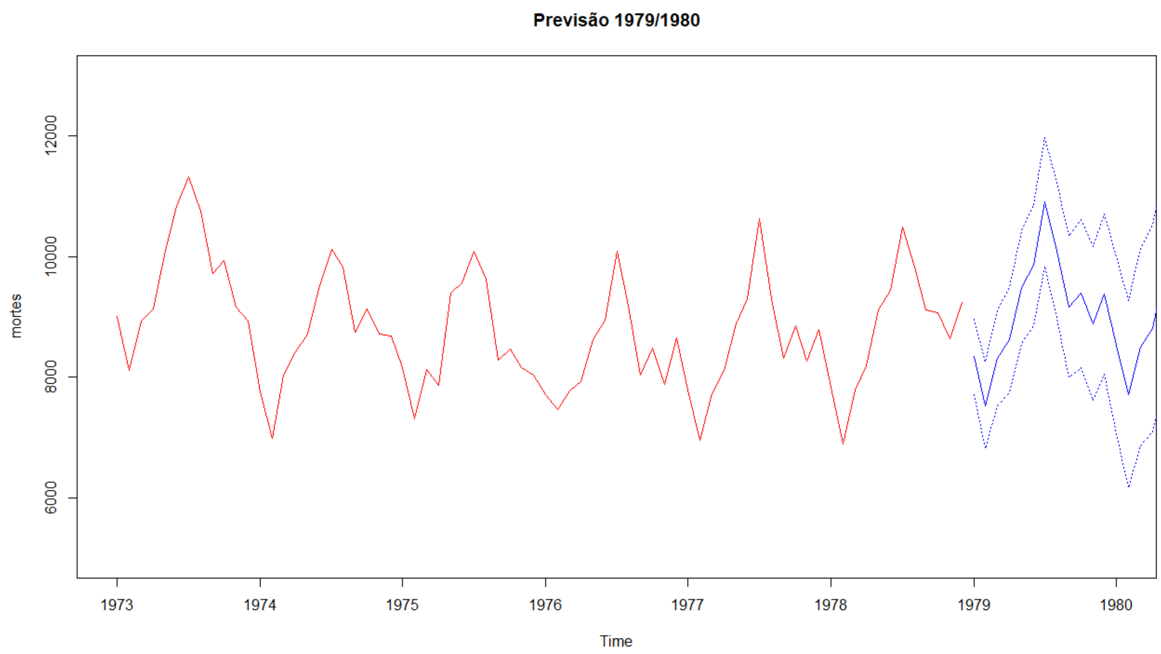

Deduzimos assim que as previsões apresentam valores próximos das observações, apesar de observamos algum erro de previsão. Isto é possível observar a partir da tabela, mas também do gráfico onde a linha de cor preta (previsões) está bastante próxima da linha de cor vermelha (série original). Concluímos assim que este modelo é válido e adequado.

- **Considerando agora o total das observações vamos produzir previsões para o futuro. Escolhemos prever os dados para 2 anos (1979 e 1980).**

Recorremos aos comandos:

```
> ts.plot(mortes, xlim=c(1973,1980), ylim=c(5000,13000), col="red", main="Previsão 1979/1980")
> m1 = arima(mortes, order=c(0,1,1), seasonal=list(order=c(0,1,1), period=12))
> previsoes79_80 <- predict(m1, n.ahead = 24)
> lines(ts(previsoes79_80$pred, start=1979, deltat=1/12), col="blue")
> low = previsoes79_80$pred - 1.96*previsoes79_80$se
> upper = previsoes79_80$pred + 1.96*previsoes79_80$se
> lines(ts(low, start=1979, deltat=1/12), lty=3, col="blue")
> lines(ts(upper, start=1979, deltat=1/12), lty=3, col="blue")
```

Obtemos assim o seguinte gráfico, em que a linha de cor azul representa as previsões dos anos 1979 e 1980, e a linha de cor vermelha a série original.




Através de : `> previsoes79_80$pred` `> previsoes79_80$se`

Obtemos para os valores previstos e respetivos erros:

| Mês | Valores previstos | Erro |
|---|---|---|
| **Jan 1979** | 8336.061 | 315.4481 |
| **Fev 1979** | 7531.829 | 363.0056 |
| **Mar 1979** | 8314.644 | 405.0168 |
| **Abril 1979** | 8616.869 | 443.0623 |
| **Maio 1979** | 9488.913 | 478.0897 |
| **Jun 1979** | 9859.757 | 510.7204 |
| **Jul 1979** | 10907.470 | 541.3879 |
| **Agos 1979** | 10086.508 | 540.4090 |
| **Set 1979** | 9164.959 | 598.0234 |
| **Out 1979** | 9384.259 | 624.4178 |
| **Nov 1979** | 8884.974 | 649.7408 |
| **Dez 1979** | 9376.574 | 674.1133 |
| **Jan 1980** | 8522.584 | 746.7030 |
| **Fev 1980** | 7718.352 | 790.6702 |
| **Mar 1980** | 8501.167 | 832.3180 |
| **Abril 1980** | 8803.392 | 871.9798 |
| **Maio 1980** | 9675.435 | 909.9127 |
| **Jun 1980** | 10046.280 | 946.3272 |

| | | |
|---|---|---|
| **Jul 1980** | 11093.993 | 981.3914 |
| **Agos 1980** | 10273.031 | 1015.2453 |
| **Set 1980** | 9351.482 | 1048.0062 |
| **Out 1980** | 9570.782 | 1079.7735 |
| **Nov 1980** | 9071.497 | 1110.6326 |
| **Dez 1980** | 9563.097 | 1140.6572 |

## Iremos agora considerar a diferenciação de ordem 2 (d=2)

Iremos proceder de forma análoga à diferenciação de ordem 1, isto é, a mesma metodologia utilizada anteriormente, pelo que não iremos explicar tão detalhadamente os passos percorridos dando mais relevância aos resultados pertinentes. Além disso, é importante referir também que omitimos o passo 3 da deteção da sazonalidade uma vez que se mantém inalterado, ou seja o período permanece s=12.

### PASSO 4: Calcular as ordens da parte sazonal

| Modelo | Coeficiente | Erro(S.E) | Significância | AIC |
|---|---|---|---|---|
| * $\mathbf{ARIMA}(0,2,0)\times(1,0,0)_{s=12}$ | Sar1: 0.7890 | 0.0621 | Coeficiente próximo de 1 | |
| $\mathbf{ARIMA}(0,2,0)\times(1,1,0)_{s=12}$ | Sar1: -0.3467 | 0.1180 | Significativo | 917.4 |
| $\mathbf{ARIMA}(0,2,0)\times(0,1,1)_{s=12}$ | Sma1:-0.5665 | 0.1624 | Significativo | 912.82 |
| $\mathbf{ARIMA}(0,2,0)\times(1,1,1)_{s=12}$ | Sar1: 0.2292<br>Sma1:-0.8051 | 0.2790<br>0.4450 | Não significativo, aceita-se $H_0: \alpha = 0$ e $H_0: \beta = 0$ | |
| $\mathbf{ARIMA}(0,2,0)\times(2,1,0)_{s=12}$ | Sar1: -0.4670<br>Sar2: -0.3392 | 0.1258<br>0.1385 | Significativo | 914.26 |
| $\mathbf{ARIMA}(0,2,0)\times(0,1,2)_{s=12}$ | Sma1: -0.5550<br>Sma2: -0.1692 | 0.2155<br>0.1914 | Não significativo, aceita-se $H_0: \beta_2 = 0$ | |

Escolhemos assim o modelo com menor AIC :

$$\boldsymbol{ARIMA(p,2,q)\times(0,1,1)_{s=12}}$$

### PASSO 5: Calcular as ordens da parte regular

| Modelo | Coeficiente | Erro(S.E) | Significância | AIC |
|---|---|---|---|---|
| $\mathbf{ARIMA}(0,2,0)\times(0,1,1)_{s=12}$ | Sma1:-0.5665 | 0.1624 | Significativo | 912.82 |
| $\mathbf{ARIMA}(1,2,0)\times(0,1,1)_{s=12}$ | Ar1: -0.5926<br>Sma1: -0.6633 | 0.1045<br>0.2097 | Significativo | 889.56 |
| $\mathbf{ARIMA}(0,2,1)\times(0,1,1)_{s=12}$ | Ma1: -0.9985 | 0.0556 | Significativo | 862.56 |

| | | | | |
|---|---|---|---|---|
| | Sma1: -0.5598 | 0.1698 | | |
| $\mathbf{ARIMA(1,2,1)\times(0,1,1)_{s=12}}$ | Ar1:-0.3329<br>Ma1: -0.9983<br>Sma1: -0.5972 | 0.1235<br>0.0599<br>0.1909 | Significativo | 857.78 |
| $\mathbf{ARIMA(2,2,0)\times(0,1,1)_{s=12}}$ | Ar1: -0.8988<br>Ar2: -0.5042<br>Sma1: -0.6337 | 0.1135<br>0.1123<br>0.2013 | Significativo | 874.62 |
| $\mathbf{ARIMA(0,2,2)\times(0,1,1)_{s=12}}$ | Ma1: -1.4659<br>Ma2: 0.4671<br>Sma1: -0.5799 | 0.1514<br>0.1390<br>0.1911 | Significativo | 854.96 |

Escolhemos assim o modelo com menor AIC :

$$\boldsymbol{ARIMA(0,2,2)\times(0,1,1)_{s=12}}$$

## Qualidade do ajustamento - Análise dos resíduos

O modelo escolhido passará a ser denominado de modelo2.

```
> modelo2 <- arima(mortes, order=c(0,2,2), seasonal=list(order=c(0,1,1), period=12))
> names(modelo2)
 [1] "coef"      "sigma2"    "var.coef"  "mask"      "loglik"    "aic"
 [7] "arma"      "residuals" "call"      "series"    "code"      "n.cond"
[13] "nobs"      "model"
> #variancia estimadas dos erros
> modelo2$sigma2
[1] 96190.12
```

Através do resultado do comando anterior observamos que a variância estimada dos erros do modelo escolhido é igual a 96190.12.

Primeiramente denominamos os resíduos de residuos2 para em seguida verificar se são ruído branco.

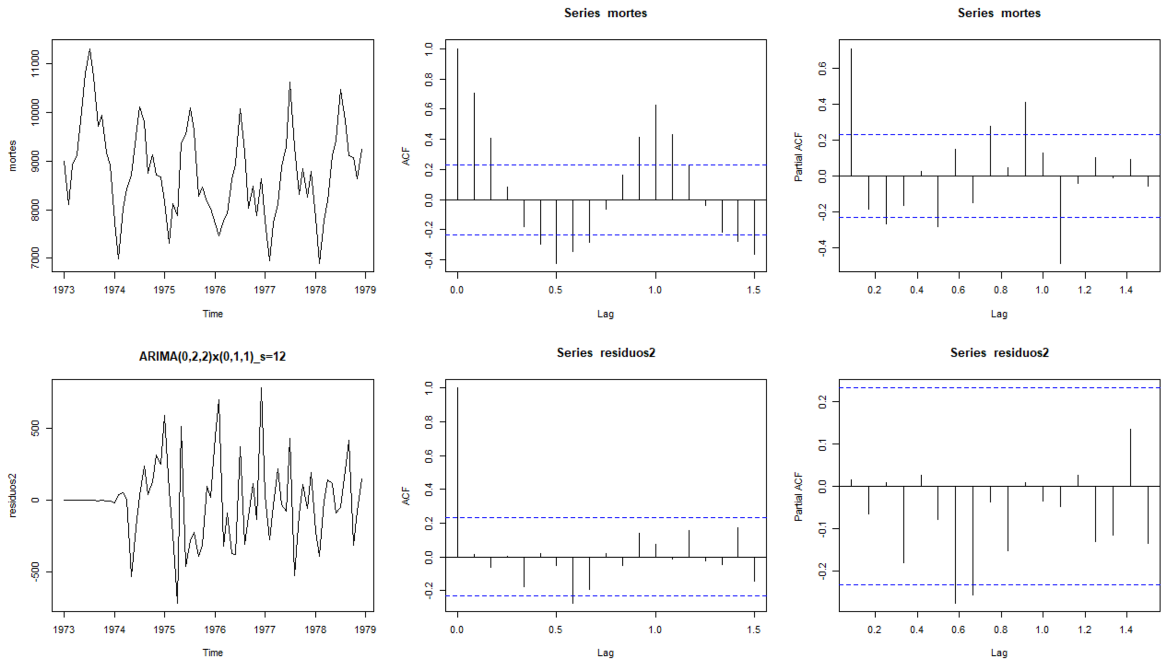
```
> residuos2 <- modelo2$residuals
```

❖ **Independência dos resíduos**

Realizamos os gráficos dos resíduos2 e das respetivas funções de autocorrelação total e parcial. Para tal, realizamos os comandos:

```
> residuos2 <- modelo2$residuals
> par(mfrow=c(2,3))
> plot(mortes)
> acf(mortes)
> pacf(mortes)
> # testar independência dos resíduos
> plot(residuos2, main="ARIMA(0,2,2)x(0,1,1)_s=12")
> acf(residuos2)
> pacf(residuos2)
```

Obtendo:

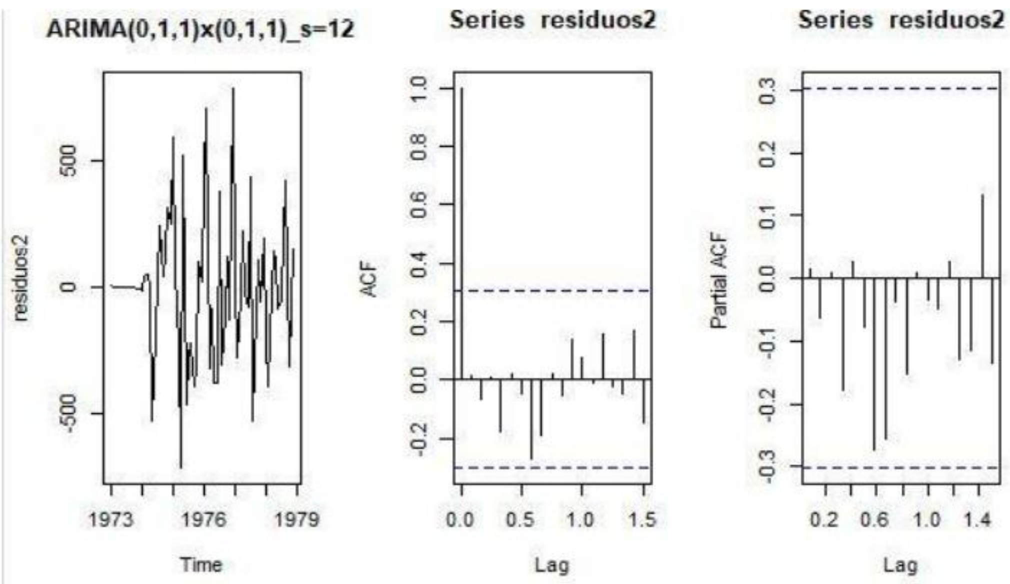



Através dos gráficos da função de autocorrelação e da função de autocorrelação parcial (quinto e sexto gráfico) podemos deduzir que não há fortes evidencias para considerar que os resíduos não são independentes.

Para retirarmos esta conclusão com maior certeza iremos considerar o nível de significância a 99%.

```
> par(mfrow=c(1,3))
> plot(residuos2, main="ARIMA(0,1,1)x(0,1,1)_s=12")
> acf(residuos2,ci=0.99)
> pacf(residuos2,ci=0.99)
```

Obtendo:



Podemos tirar a mesma conclusão através do seguinte teste de hipótese Box- Pierce ou Ljung – Box.

```
> Box.test(residuos2)

        Box-Pierce test

data:  residuos2
X-squared = 0.013953, df = 1, p-value = 0.906
```

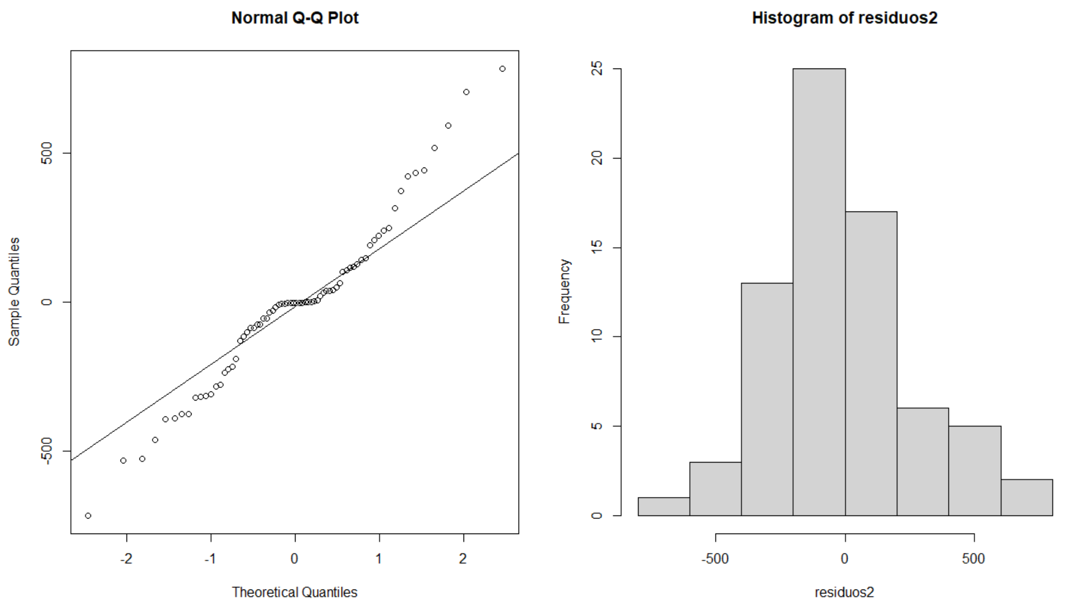
Como o valor p deste teste para os resíduos é 0.906 não há evidências para rejeitar a hipótese nula ($H_0$: $\varepsilon_t$ *são independentes*) para os níveis de significância usuais. Concluindo que os resíduos são independentes.

- Verificar se os resíduos seguem a distribuição normal

```
> library(nortest)
> lillie.test(residuos2)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  residuos2
D = 0.11283, p-value = 0.02384
```

Através do teste Lilliefors (Kolmogorov-Smirnov) da biblioteca nortest (comandos ilustrados acima), obtivemos o valor p igual 0.02384 sendo que não se rejeita a hipótese nula, $H_0$: $\varepsilon_t$ *segue um distribuição normal*, para o nível de significância de 1%. Concluindo que os resíduos seguem uma distribuição normal, podemos também fazer a mesma análise a partir dos seguintes gráficos.

```
> par(mfrow=c(1,2))
> qqnorm(residuos2)
> qqline(residuos2)
> hist(residuos2)
```




- Média dos resíduos

De modo a verificar se a média dos resíduos é igual a zero utilizamos o teste T de student.

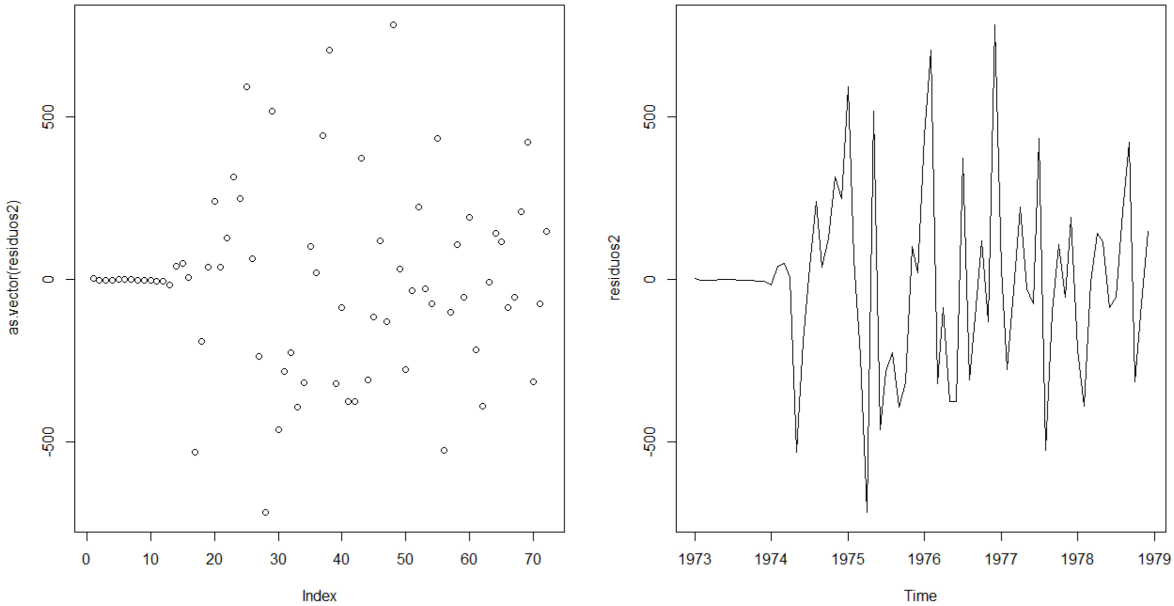
```
> t.test(residuos2)

        One Sample t-test

data:  residuos2
t = -0.23318, df = 71, p-value = 0.8163
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -73.54928  58.14805
sample estimates:
mean of x
-7.700618
```

O valor p resultante do comando anterior é 0.8163 portante não rejeitamos a hipótese nula, $H_0: E[\varepsilon_t] = 0$, deduzindo que os resíduos têm valor médio igual a 0 para os valores de significância usuais.

❖ Variância dos resíduos

Com a finalidade de analisar a variância dos resíduos iremos recorrer à representação gráfica deste.

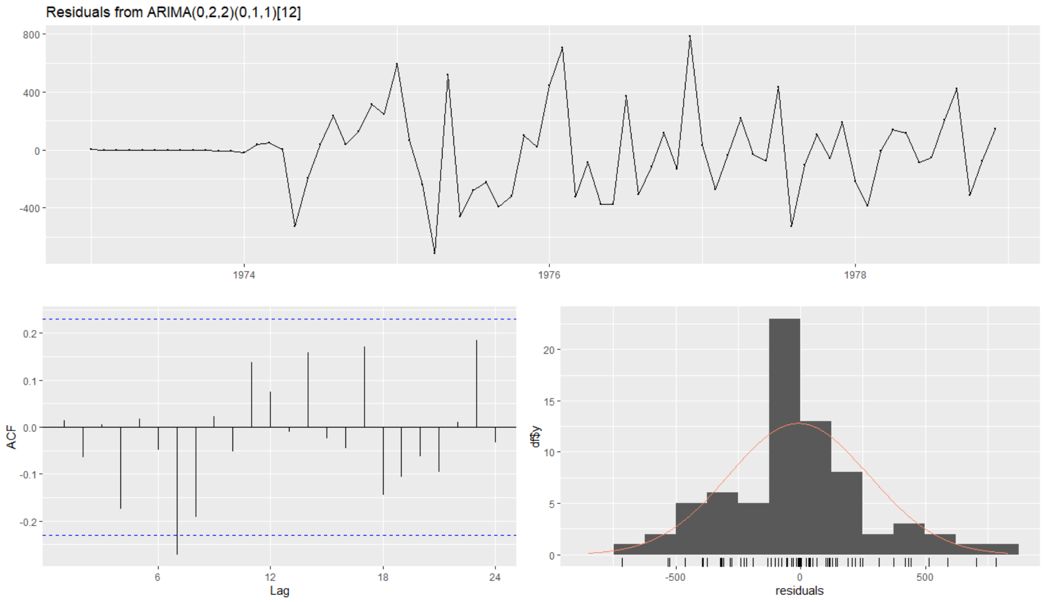
```
> plot(as.vector(residuos2))
> plot(residuos2)
```




Tendo em conta que os gráficos não apresentam nenhum padrão, não há evidências para não considerar a variância dos resíduos constante. Existindo assim homocedasticidade.

Também podemos fazer uma análise extra dos resíduos de forma a reforçar as conclusões anteriores.

```
> library(forecast)
> checkresiduals(modelo2)
```




## Previsões

- De modo a validar ou não modelo, iremos ignorar as últimas 24 observações (2 anos) e fazer previsões para esse período ignorado comparar os valores estimados com os observados. Ou seja, estimar $x_{49}, \dots, x_{71}, x_{72}$ .

Começamos por denominar o modelo escolhido anteriormente por m2 e as previsões por previsoes2.

```
> m2 = arima(mortes[1:48], order=c(0,2,2), seasonal=list(order=c(0,1,1), period=12))
> previsoes2 <- predict(m2, n.ahead = 24)
> mortes[49:72]
> previsoes2$pred
> previsoes2$se
```

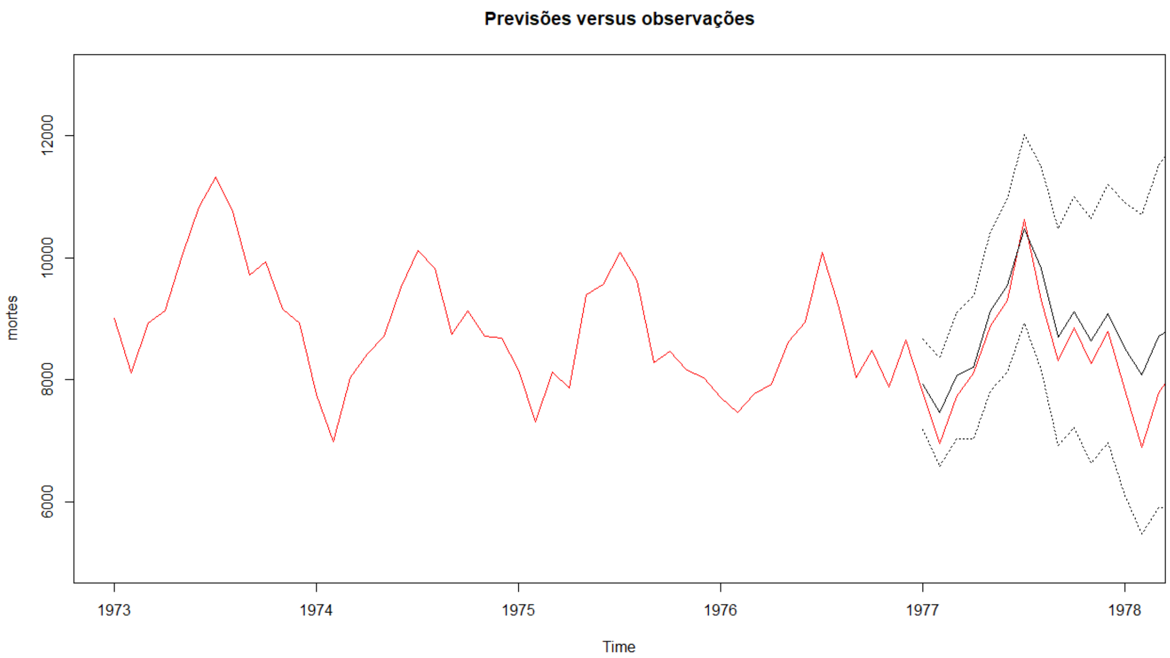
Resultando os seguintes valores previstos e os respetivos erros.

| Mês | Valores observados | Valores previstos | Erro |
|---|---|---|---|
| **Jan 1977** | 7792 | 7922.572 | 376.0704 |
| **Fev 1977** | 6957 | 7470.875 | 455.2374 |
| **Mar 1977** | 7726 | 8070.219 | 527.7833 |
| **Abril 1977** | 8106 | 8205.005 | 596.1352 |
| **Maio 1977** | 8890 | 9122.553 | 661.5987 |
| **Jun 1977** | 9299 | 9540.645 | 724.9608 |
| **Jul 1977** | 10625 | 10477.790 | 786.7330 |
| **Agos 1977** | 9302 | 9822.336 | 847.2666 |
| **Set 1977** | 8314 | 8696.856 | 906.8129 |
| **Out 1977** | 8850 | 9110.071 | 965.5576 |
| **Nov 1977** | 8265 | 8634.433 | 1023.6415 |

| | | | |
|---|---|---|---|
| **Dez 1977** | 8796 | 9082.688 | 1081.1740 |
| **Jan 1978** | 7836 | 8505.544 | 1226.2916 |
| **Fev 1978** | 6892 | 8084.293 | 1331.9023 |
| **Mar 1978** | 7791 | 8714.083 | 1434.7054 |
| **Abril 1978** | 8192 | 8879.314 | 1535.2695 |
| **Maio 1978** | 9115 | 9827.307 | 1634.0126 |
| **Jun 1978** | 9434 | 10275.844 | 1731.2506 |
| **Jul 1978** | 10484 | 11243.434 | 1827.2276 |
| **Agos 1978** | 9827 | 10618.425 | 1922.1363 |
| **Set 1978** | 9110 | 9523.390 | 2016.1313 |
| **Out 1978** | 9070 | 9967.050 | 2109.3381 |
| **Nov 1978** | 8633 | 9521.858 | 2201.8600 |
| **Dez 1978** | 9240 | 10000.557 | 2293.7833 |

Criamos um gráfico com a série temporal original (mortes), as previsões e os intervalos de previsão a 95%.

```
> ts.plot(mortes, xlim=c(1973,1978), ylim=c(5000,13000), col="red", main="Previsões versus observações")
> lines(ts(previsoes2$pred, start=1977, deltat=1/12))
> low = previsoes2$pred - 1.96*previsoes2$se
> upper = previsoes2$pred + 1.96*previsoes2$se
> lines(ts(low, start=1977, deltat=1/12), lty=3)
> lines(ts(upper, start=1977, deltat=1/12), lty=3)
```



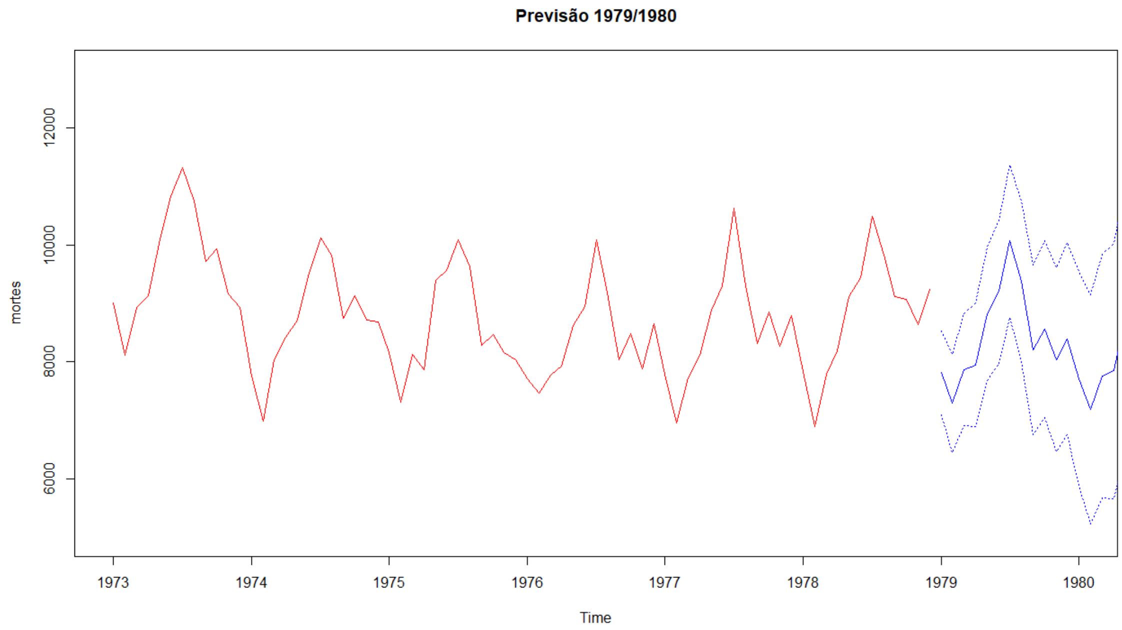

Esta representação gráfica permite-nos uma melhor interpretação da previsão, uma vez que é possível observar que os valores previstos se aproximam dos valores realmente observados. Recorrendo à tabela podemos visualizar o mesmo, concluindo assim que o modelo é adequado.

- **Iremos agora fazer previsão para os anos 1979 e 1980, ou seja, previsões para o futuro uma vez que estes anos não foram observados.**

Recorrendo ao software R, utilizamos a seguinte instrução:

```
> ts.plot(mortes, xlim=c(1973,1980), ylim=c(5000,13000), col="red", main="Previsão 1979/1980")
> m2 = arima(mortes, order=c(0,2,2), seasonal=list(order=c(0,1,1), period=12))
> previsoes79_80 <- predict(m1, n.ahead = 24)
> lines(ts(previsoes79_80$pred, start=1979, deltat=1/12), col="blue")
> low = previsoes79_80$pred - 1.96*previsoes79_80$se
> upper = previsoes79_80$pred + 1.96*previsoes79_80$se
> lines(ts(low, start=1979, deltat=1/12), lty=3, col="blue")
> lines(ts(upper, start=1979, deltat=1/12), lty=3, col="blue")
```

Resultando a seguinte representação gráfica da previsão e com os respetivos intervalos de previsão (valores previstos estão representados a azul e a vermelho a serie original).



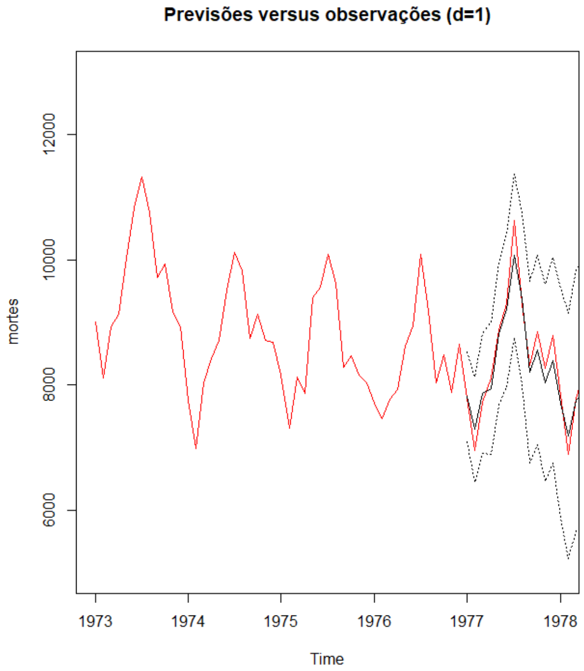

Utilizando `> previsoes79_80$pred` `> previsoes79_80$se` :

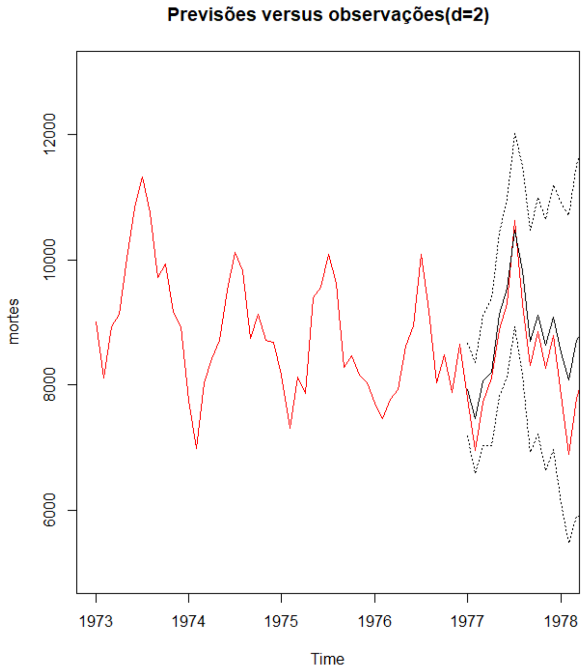
Obtemos:

| Mês | Valores previstos | Erro |
|---|---|---|
| **Jan 1979** | 8409.849 | 314.5031 |
| **Fev 1979** | 7659.204 | 360.5742 |
| **Mar 1979** | 8486.608 | 403.5404 |
| **Abril 1979** | 8828.264 | 444.3038 |
| **Maio 1979** | 9746.304 | 483.4221 |
| **Jun 1979** | 10170.902 | 521.2662 |
| **Jul 1979** | 11253.943 | 558.0954 |
| **Agos 1979** | 10486.042 | 594.0988 |
| **Set 1979** | 9600.179 | 629.4183 |

| | | |
|---|---|---|
| **Out 1979** | 9875.404 | 664.1633 |
| **Nov 1979** | 9421.648 | 698.4195 |
| **Dez 1979** | 9944.561 | 732.2559 |
| **Jan 1980** | 9174.757 | 815.0840 |
| **Fev 1980** | 8445.318 | 870.2077 |
| **Mar 1980** | 9293.929 | 924.0025 |
| **Abril 1980** | 9656.791 | 976.6884 |
| **Maio 1980** | 10596.037 | 1028.4361 |
| **Jun 1980** | 11041.841 | 1079.3807 |
| **Jul 1980** | 12146.008 | 1129.6311 |
| **Agos 1980** | 11399.394 | 1179.2763 |
| **Set 1980** | 10534.736 | 1228.3898 |
| **Out 1980** | 10831.167 | 1277.0332 |
| **Nov 1980** | 10398.618 | 1325.2585 |
| **Dez 1980** | 10942.737 | 1373.1099 |

# Comparação dos dois modelos




Comparando os dois gráficos das previsões já referidos acima, observamos que os valores previstos se aproximam mais dos valores observados para a diferenciação de ordem 1, pelo que este será o melhor modelo. No entanto, este modelo falha um dos pressupostos do ruído branco como visto anteriormente. Concluímos assim que no limite escolheríamos o modelo $\boldsymbol{ARIMA(0,1,1)\times(0,1,1)_{s=12}}$.

# Conclusão

Este projeto foi realizado em conjunto pelo grupo, desde a pesquisa e recolha de informação sobre a série temporal accdeaths, até à realização deste documento, passando pela realização em si do script do R.

Também pensamos em conjunto qual seria a estrutura do trabalho, bem como de que forma poderíamos tornar este mais apelativo e sólido. Em adição achamos que este trabalho de projeto contribuiu para o desenvolvimento das nossas capacidades de pesquisa, de utilização do R e a respetiva realização de gráficos, criatividade e sobretudo no desenvolvimento da aplicação da matéria lecionada na UC de Séries Temporais. Com este trabalho foi possível estudar de que forma evolui o número de mortes acidentais nos Estados Unidos da América, ao longo dos anos, no nosso caso desde 1973 a 1978. Também fomos capazes de realizar uma modelação e previsão da nossa série, no nosso caso com diferentes ordens de diferenciação (d=1 e d=2). Percebemos que este trabalho reforçou bastante a nossa aprendizagem e deu-nos a oportunidade de explorar a matéria de uma forma mais interessante.

Após a realização deste, podemos afirmar que enriquecemos o nosso conhecimento. Além disso, o trabalho dinamizado permitiu também o desenvolvimento do estímulo do raciocínio, a capacidade de trabalho em grupo, a responsabilidade, criatividade e autonomia.

Em suma, consideramos que os objetivos do mesmo foram alcançados e corresponderam às expectativas da professora responsável pela unidade curricular e do grupo nomeadamente por ter potenciado espaços/ momentos de partilha, aprendizagem e reflexão conjunta no contexto académico, no desenvolvimento de competências e atitudes em termos de partilha e colaboração. A partir da voz da professora foi também possível recolher informações relevantes que permitissem uma melhor compreensão das oportunidades de aprendizagem no contexto de trabalho.

Tendo em vista o exposto, considera-se possível formular um modelo ARIMA adequado aos dados, de forma a obter boas previsões. Para isto, foi fundamental aplicar o conhecimento adquirido ao longo desta UC.

# Webgrafia

https://facilitaseguros.com.br/blog/seguro-morte-acidental/

https://pt.wikipedia.org/wiki/Morte_acidental