Projeto Final

Disciplina Métodos Estatísticos de Previsão

Isabela Ferreira Ventura Cruz Nathalia Gabriella Ferreira dos Santos

29/11/2023

Índice

Capítulo 1

Descrição dos Dados

A série escolhida para realizar o trabalho foi obtida por meio da plataforma kaggle no seguinte link. O banco de dados é sobre o **número mensal de passageiros de voos aéreos do ano de 1949 a 1960**, ao todo contém 144 observações. A tabela Tabela 1.1 contém as 5 primeiras observações do conjunto de dados escolhido.

Tabela 1.1: Resumo das 5 primeiras observações do conjunto de dados

Mes	Passageiros
1949-01	112
1949-02	118
1949-03	132
1949-04	129
1949-05	121
1949-06	135

Como o foco do projeto é realizar previsões, optou-se por retirar as 12 observações finais do conjunto de dados. Ou seja, 132 observações serão usadas para *treino* dos modelos de previsão para séries temporais e depois será comparado quão bem cada modelo usado realizou as previsões.

A Figura 1.1 indica o comportamento dos dados ao longo do tempo. Nota-se *comportamento crescente* além de presença de grandes picos de *sazonalidade*, a qual parece se comportar como um modelo em que a sazonalidade seja aditiva e/ou multiplicativa, ou seja, a sazonalidade muda ao longo do tempo, esse caso será melhor avaliado na Seção 2.2.

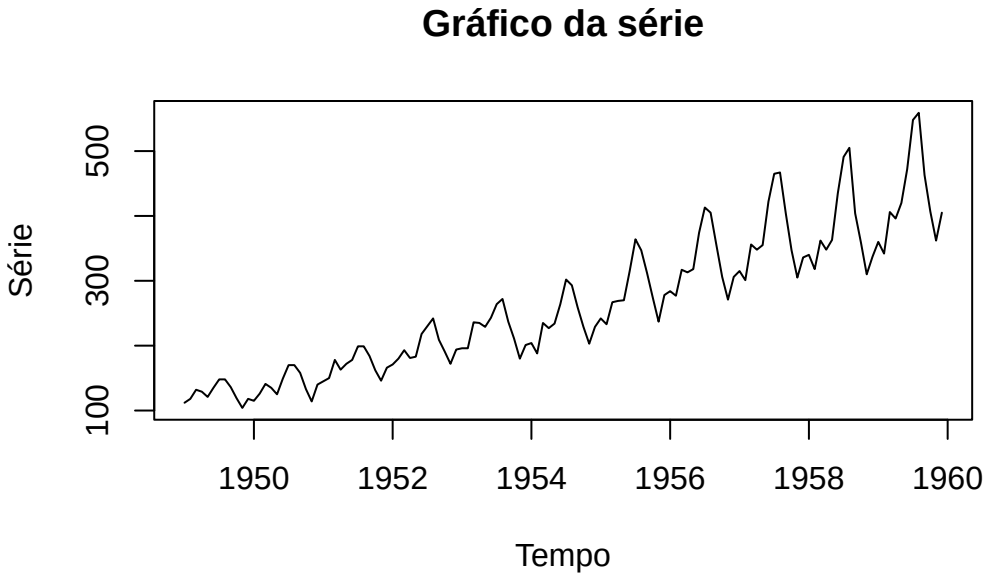


Figura 1.1: Gráfico da série

Capítulo 2

Ajustes de Modelos

2.1 Modelo ARIMA

2.1.1 Estacionariedade

Os modelos ARIMA de Box e Jenkins partem do pressuposto de que a série é estacionária, desse modo, o primeiro passo para uma modelagem acertiva é verificar a estacionariedade da série. Para isso foi utilizado o teste aumentado de Dickey-Fuller, em que a hipótese nula é que os dados não são estacionários, o valor-p do teste é 0,01, ou seja, rejeitamos essa hipótese e não precisaremos tomar diferenças na série (d=0).

```
    Augmented Dickey-Fuller Test

data:  serie_train
Dickey-Fuller = -7.1692, Lag order = 5, p-value = 0.01
alternative hypothesis: stationary
```

2.1.2 Identificação do modelo

Para identificar o modelo foram analisados os gráficos de autocorrelação (ACF) e autocorrelação parcial (PACF). Na Figura 2.1 podemos notar que há um decaimento rápido e picos significativos nos lags multiplos de 12, esse é um forte indicativo de sazonalidade (S =12). Já em Figura 2.2 podemos notar um pico significativo no lag 1, e picos significativos nos lags 8 e 10 e 12. Como a significância adotada neste trabalho é de 5%, espera-se que 5% dos valores de Φ_{kk} estejam fora dos limites esperados, desse modo, consideraremos apenas os lags 1, e 12 significativos. Desse modo, o pico no lag 12 é mais um indício de sazonalidade de ordem 12. Juntando as informações trazidas pelos gráficos da ACF e da PACF, iniciamos a análise partindo de um modelo $SARIMA(1,0,0)(0,0,1)_{12}$ e o sobrefixamos, aumentando a ordem dos componentes um por um.

2.1.3 Modelo inicial:

O modelo inicial é um $SARIMA(1,0,0)(0,0,1)_{12}$. Todos os coeficientes são significativos e o AIC é 1187.266

```
z test of coefficients:

            Estimate Std. Error z value  Pr(>|z|)
ar1         0.959137   0.023787 40.3224 < 2.2e-16 ***
sma1        0.855174   0.086748  9.8581 < 2.2e-16 ***
intercept 274.769648  64.800793  4.2402 2.233e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

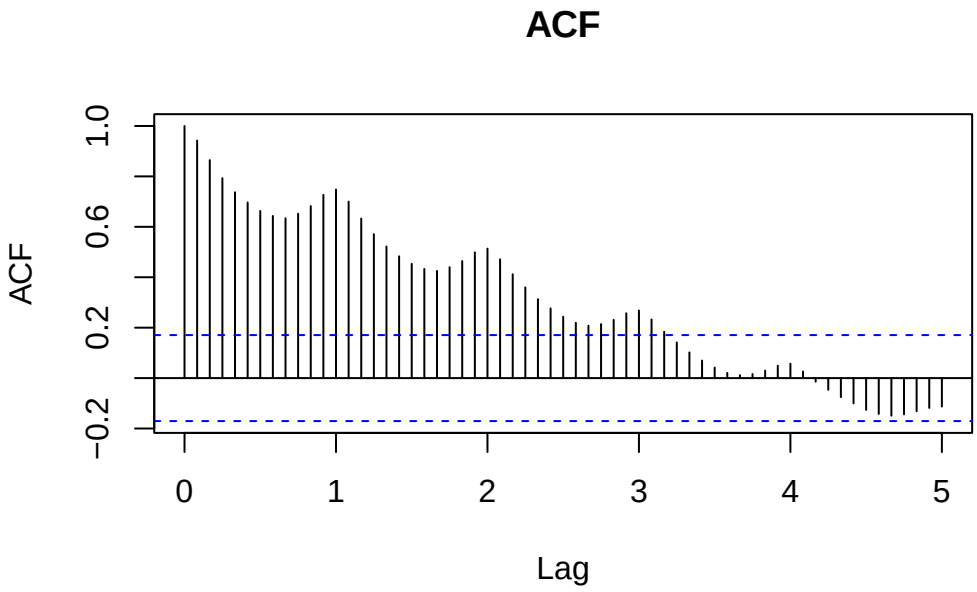


Figura 2.1: ACF

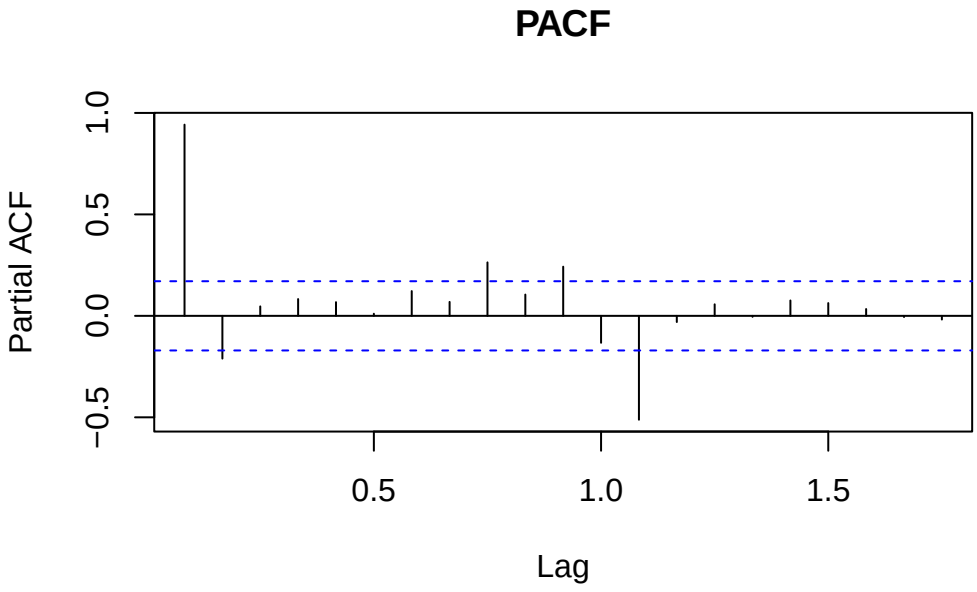


Figura 2.2: PACF

```
[1] 1187.266
```

2.1.4 Modelo 02

Aumentando a ordem p chegamos a um modelo $SARIMA(2,0,0)(0,0,1)_{12}$. Todos os coeficientes são significativos e o AIC (1181.575) caiu.

```
z test of coefficients:

            Estimate Std. Error z value  Pr(>|z|)
ar1         1.192885   0.086110 13.8531 < 2.2e-16 ***
ar2        -0.244836   0.086967 -2.8153  0.004873 **
sma1        0.822361   0.076947 10.6873 < 2.2e-16 ***
intercept 273.042171  51.843982  5.2666  1.39e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2.1.5 Modelo 03

Aumentando a ordem p novamente chegamos a um $SARIMA(3,0,0)(0,0,1)_{12}$. Nem todos os coeficientes são significativos e o AIC (1182.402) aumentou em relação ao modelo 02. Desse modo, ficamos com $p = 2$ e o melhor modelo até então é o modelo 02.

```
z test of coefficients:

            Estimate Std. Error z value  Pr(>|z|)
ar1         1.219954   0.089341 13.6550 < 2.2e-16 ***
ar2        -0.364022   0.139832 -2.6033  0.009234 **
ar3         0.098112   0.090216  1.0875  0.276803
sma1        0.812231   0.075660 10.7353 < 2.2e-16 ***
intercept 274.630258  56.997937  4.8182 1.448e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2.1.6 Modelo 04

Aumentando a ordem q temos um modelo $SARIMA(2,0,1)(0,0,1)_{12}$. Todos os coeficientes são significativos e o AIC (1177.668) caiu

```
z test of coefficients:

            Estimate Std. Error z value  Pr(>|z|)
ar1         0.187247   0.068332  2.7403  0.006139 **
ar2         0.730845   0.068661 10.6442 < 2.2e-16 ***
ma1         0.981127   0.025184 38.9581 < 2.2e-16 ***
sma1        0.830113   0.079763 10.4073 < 2.2e-16 ***
intercept 273.526701  61.182435  4.4707 7.797e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2.1.7 Modelo 05

Aumentando a ordem q chegamos a um $SARIMA(2,0,2)(0,0,1)_{12}$. Nem todos os coeficientes são significativos e o AIC (1184.598) aumentou. Desse modo, ficamos com *q=1* e o melhor modelo até então é o modelo 04.

```
z test of coefficients:

            Estimate Std. Error z value  Pr(>|z|)
ar1         0.769208   0.373981  2.0568    0.0397 *
ar2         0.147288   0.354473  0.4155    0.6778
ma1         0.448210   0.362247  1.2373    0.2160
ma2         0.147031   0.148296  0.9915    0.3215
sma1        0.822551   0.077725 10.5828 < 2.2e-16 ***
intercept 273.277201  51.365597  5.3202 1.036e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2.1.8 Modelo 06

Aumentando a ordem P: $SARIMA(2,0,1)(1,0,1)_{12}$. Nota-se que o ajuste não convergiu. Desse modo, ficamos com *P=0*.

2.1.9 Modelo 07

Aumentando a ordem Q chegamos a um $SARIMA(2,0,1)(0,0,2)_{12}$. Nem todos os coeficientes são significativos mas o AIC (1137.312) caiu.

```
z test of coefficients:

           Estimate Std. Error z value  Pr(>|z|)
ar1         0.29377    0.72305  0.4063    0.6845
ar2         0.62763    0.68866  0.9114    0.3621
ma1         0.65155    0.72157  0.9030    0.3666
sma1        1.09624    0.12017  9.1221 < 2.2e-16 ***
sma2        0.66639    0.11167  5.9675 2.410e-09 ***
intercept 277.02688   65.06150  4.2579 2.063e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2.1.10 Modelo 08

Aumentando a ordem Q temos um SARIMA(2,0,1)(0,0,3)12. Nem todos os coeficientes são significativos mas o AIC caiu. A partir daqui os modelos começam a ficar cada vez mais complicados e de convergência mais lenta, desse modo, paramos com Q=3

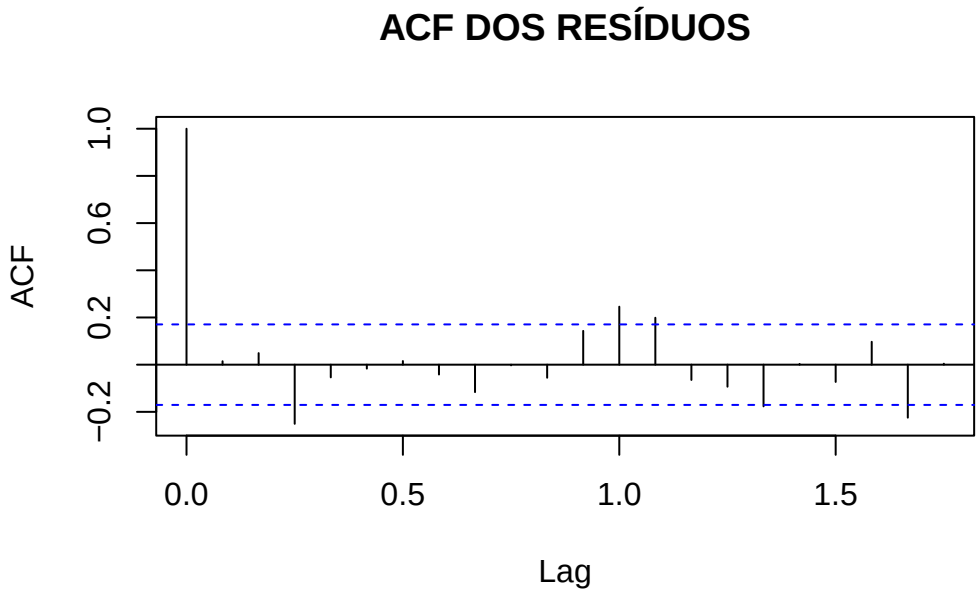
Tabela 2.1: Modelos SARIMA ajustados

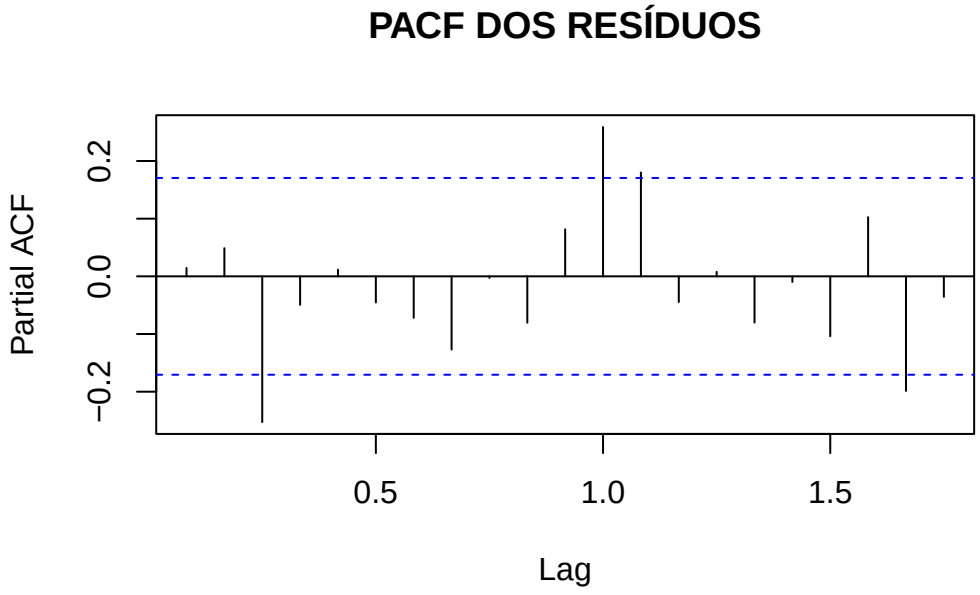
	Modelo	Número de coeficientes	Porcentagem de coeficientes significativos	AIC
01	SARIMA(1,0,0)(0,0,1)12	3	100 %	1187.266
02	SARIMA(2,0,0)(0,0,1)12	4	100 %	1181.575
03	SARIMA(3,0,0)(0,0,1)12	5	80 %	1182.402
04	SARIMA(2,0,1)(0,0,1)12	5	100 %	1177.668
05	SARIMA(2,0,2)(0,0,1)12	6	50 %	1184.598
06	SARIMA(2,0,1)(1,0,1)12	-	-	-
07	SARIMA(2,0,1)(0,0,2)12	6	50 %	1137.312
08	SARIMA(2,0,1)(0,0,3)12	7	86 %	1096.508

Comparando todos os modelos acima, escolhemos o melhor considerando o menor AIC e maior número de coeficientes significativos. Os modelo com menor AIC é o 08 com 6 coeficientes significativos em 7, o modelo com segundo menor AIC é o modelo 07 e possui apenas metade dos coeficientes significativos, já o modelo 04 possui todos os coeficientes significativos mas AIC elevado se comparado com os modelos 07 e 08. Diante disso, optamos por escolher o modelo 08, pois apenas um coeficiente não é significativo e o AIC é menor dentre todas as opções abordadas.

2.1.11 Análise de resíduos

Uma das suposições dos modelos de Box e Jenkins é a que os resíduos são um ruído branco.A suposição de independência dos resíduos pode ser verficada por meio dos gráficos de ACF e PACF dos resíduos, bem como o teste formal de Ljung-Box cuja hipótese nula é H0: Não há autocorrelação nos resíduos. Apesar dos gráficos de ACF possuírem alguns picos significativos, o teste formal de Ljung-Box aponta fortemente que os resíduos são independentes.






```
    Box-Pierce test

data:  mod8$residuals
X-squared = 0.027968, df = 1, p-value = 0.8672
```

A partir do gráfico dos Figura 2.3 que observa *Resíduos x Tempo* para avaliar homocedasticidade. É possível notar um comportamento de maior amplitude no final da série, o que poderia ser indício de heterocedasticidade. Entretanto, como os valores estão oscilando em torno de 0 e a fim de progredir no estudo, sem perda de generalidade será tolerado esse aumento.

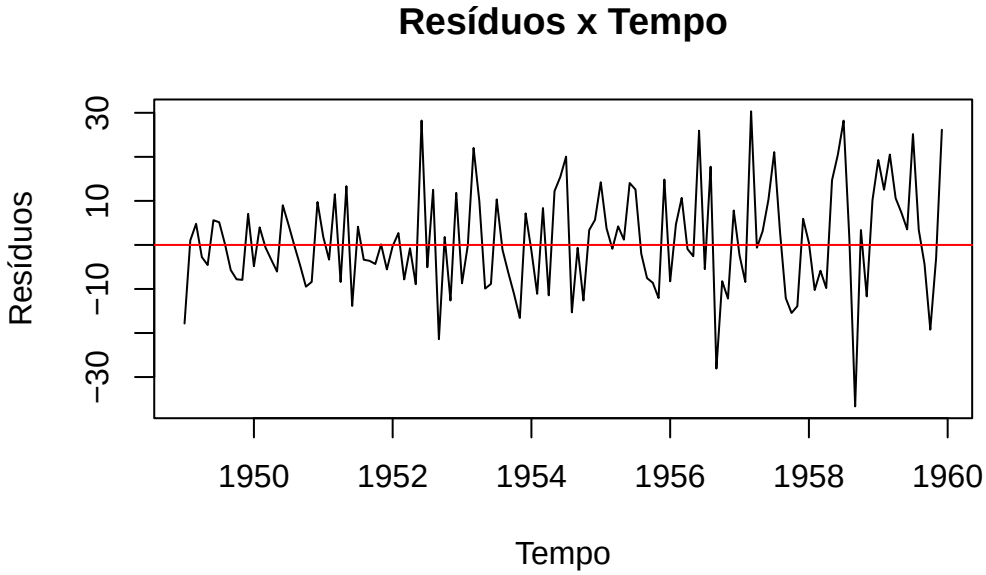


Figura 2.3: Gráfico para avaliar homocedastididade

Além disso pode ser verificado por meio do histograma em Figura 2.4 e do teste formal de shapiro-wilk de que há evidência de que os resíduos vieram de uma distribuição Normal. Logo, tem-se um Ruído Branco Gaussiano nesse caso.

```
    Shapiro-Wilk normality test

data:  mod8$residuals
W = 0.9854, p-value = 0.1716
```

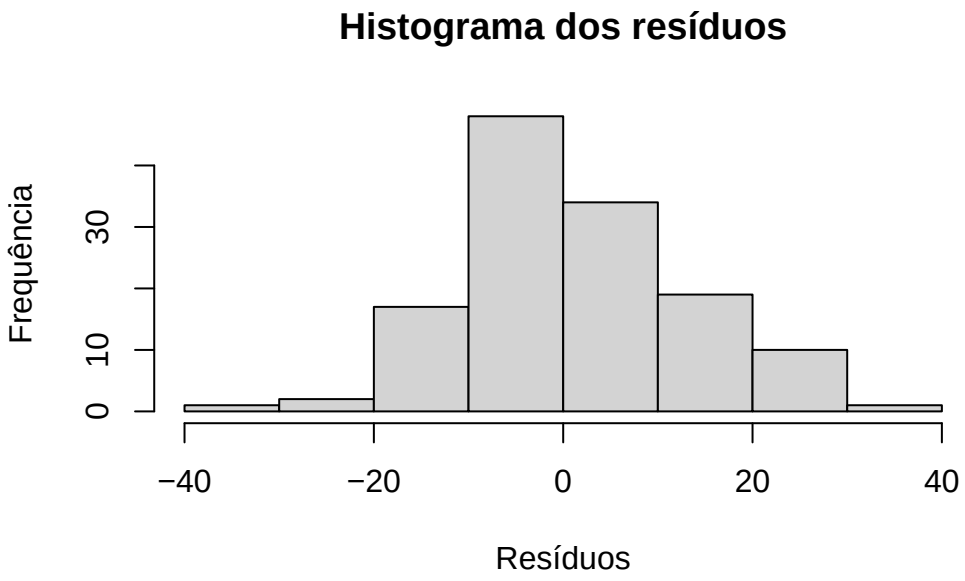


Figura 2.4: Normalidade dos resíduos

2.2 Modelo de Alisamento Exponencial

Como observado na Figura 1.1, devido a presença de sazonalidade aaditiva e/ou multiplicativa, os modelo de **Alisamento Exponencial de HOLT-WINTERS** aditivo e com fator multiplicativo foram considerados a fim de avaliar o Erro Quadrático Médio de Previsão para ambos ajustes e escolher o melhor.

As constantes de suavização encontradadas são:

Tabela 2.2: Constantes para ambos modelos

Constante	Modelo Aditivo	Modelo Multiplicativo
α	0.2465	0.8027
β	0.036	0.0085
γ	1	1

A partir das constantes encontradas, sobre os modelos, nota-se que:

- o α associado ao nível é maior no modelo multiplicativo. Ou seja, o modelo multiplicativo coloca cerca de 3 mais peso na influência das informações mais recentes sobre o nível do que o modelo aditivo.
- Em contrapartida, o β associado a tendência é 4 vezes maior no modelo aditivo, o que implica que o modelo aditivo coloca mais peso na influência das informações mais recentes sobre a tendência.

- Por fim, ambos colocam pesos iguais ao γ associado a sazonalidade. Como o peso é 1, indica que toda sazonalida pode ser explicada pelas informações recentes.

Além disso, foi calculado o Erro Quadrático Médio de Previsão para ambos ajuste, como:

$EQMP_{Aditivo} = 1.8127548 \times 10^4 < 2.0768381 \times 10^4 = EQMP_{Multiplicativo}$,

o modelo aditivo deve ser preferível para modelar a série, logo, ele será usado para previsão a ser feita na Seção 3.

Capítulo 3

Comparação de modelos

3.1 Previsão usando ARIMA

Os valores preditos usando o modelo final obtido na Seção 2.1 estão contidos na Tabela 3.1 , uma coluna foi adicionada indicando se o Intervalo de Confiança contém ou não o valor real da série. Nesse caso, apenas para o 7º mês foi observado que o valor real não estava contido no intervalo.

Tabela 3.1: Previsão modelo SARIMA(2,0,1)(0,0,3)12

real	fit	lwr	upr	ConclusaoIC
417	431	406	456	Contém ajuste
391	413	377	449	Contém ajuste
419	452	410	495	Contém ajuste
461	442	394	491	Contém ajuste
472	465	413	517	Contém ajuste
535	492	436	548	Contém ajuste
622	561	502	620	Não contém ajuste
606	560	498	621	Contém ajuste
508	486	422	550	Contém ajuste
461	445	379	511	Contém ajuste
390	409	341	477	Contém ajuste
432	448	378	517	Contém ajuste

Além disso, a Figura 3.1 indica o comportamento das previsões feitas junto a série usada para treino. Nota-se que os intervalos estão mais próximos do valor ajustado, o que indica que com mais certeza (menor variabilidade) o modelo consegue predizer o valor real.

3.2 Previsão usando Alisamento Exponencial

Os valores preditos usando o modelo final obtido na Seção 2.2 estão contidos na Tabela 3.2, uma coluna foi adicionada indicando se o Intervalo de Confiança contém ou não o valor real da série. Nesse caso, apenas para o 3º mês foi observado que o valor real não estava contido no intervalo.

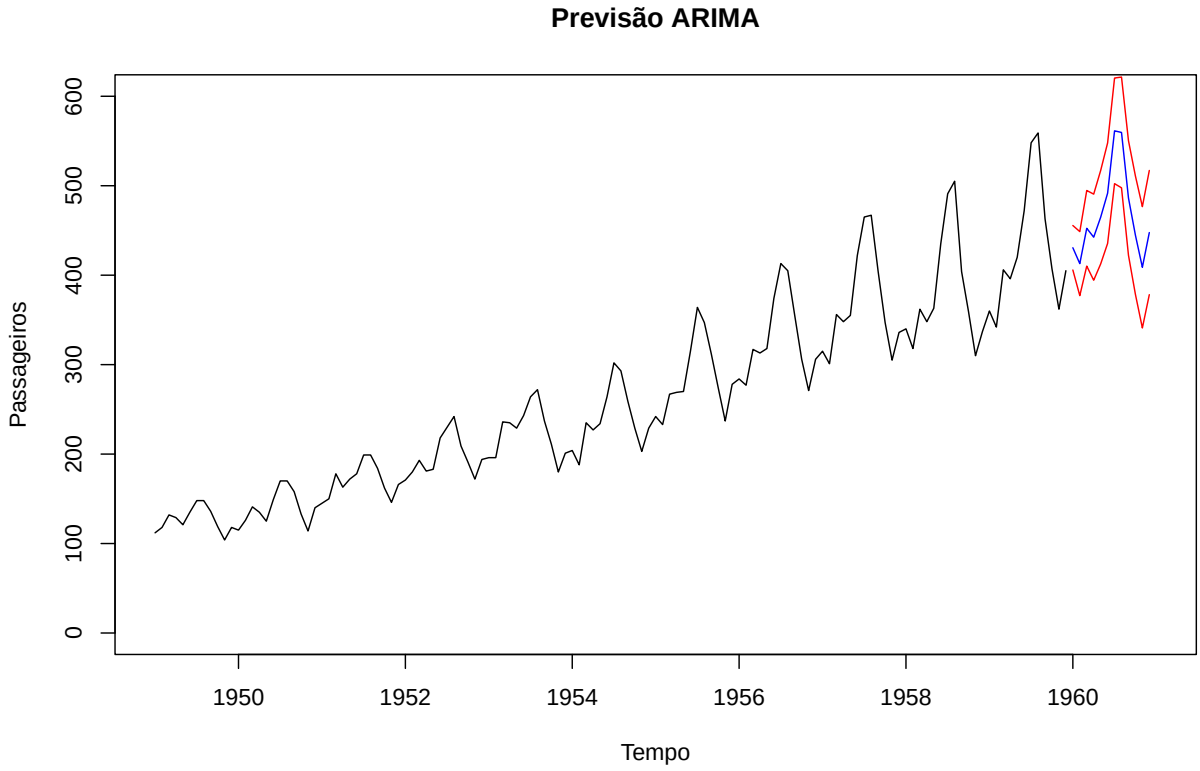


Figura 3.1: Valores preditos para o modelo SARIMA(2,0,1)(0,0,3)12

Tabela 3.2: Previsão AEWH-Aditivo

real	fit	lwr	upr	ConclusaoIC
417	417	393	441	Contém ajuste
391	399	374	423	Contém ajuste
419	460	434	485	Não contém ajuste
461	448	422	475	Contém ajuste
472	470	443	497	Contém ajuste
535	525	497	553	Contém ajuste
622	598	569	627	Contém ajuste
606	605	575	635	Contém ajuste
508	506	475	536	Contém ajuste
461	449	418	481	Contém ajuste
390	404	371	437	Contém ajuste
432	444	410	477	Contém ajuste

Além disso, a Figura 3.2 indica o comportamento das previsões feitas junto a série usada para treino. Nota-se que os intervalos estão bem próximos do valor ajustado, o que indica que com mais certeza (menor variabilidade) o modelo consegue predizer o valor real.

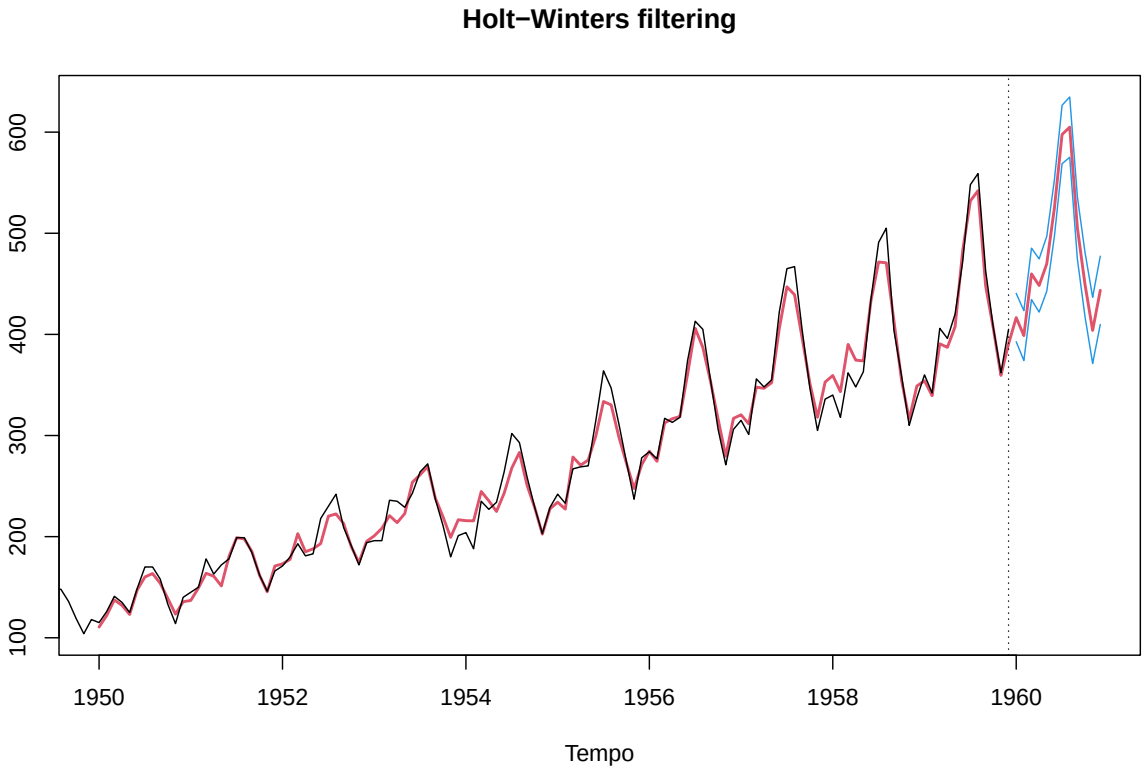


Figura 3.2: Valores preditos para modelo AEWH-Aditivo

3.3 Erro quadrático médio de previsão

Para definir qual o melhor modelo comparamos o erro quadrático médio de previsão. O EQMP do modelo aditivo é 18127.55 e o do modelo SARIMA é 935.43, desse modo, o modelo que apresenta previsões mais acertivas e portanto o melhor modelo é o SARIMA, pois possui menor EQMP.

$EQMP_{Sarima} = 935.43 < EQMP_{Aditivo} = 1.8127548 \times 10^4 < 935.4295195 = EQMP_{Multiplicativo}$,

Capítulo 4

Conclusão

Dado o discutido nesse trabalho, pelo observado das previsões obtidas na Seção 3, conclui-se que o modelo que resultou nas melhores previsões e portanto entende-se ser o melhor, foi o modelo $SARIMA(2,0,1)(0,0,3)_{12}$.