Modelos ARIMA com metodologia Box-Jenkins

Universidade Federal do Ceará
Centro de Ciências
Departamento de Estatística e Matemática Aplicada
Bacharelado em Estatística
Análise de Séries Temporais

Antônio Arthur Silva de Lima

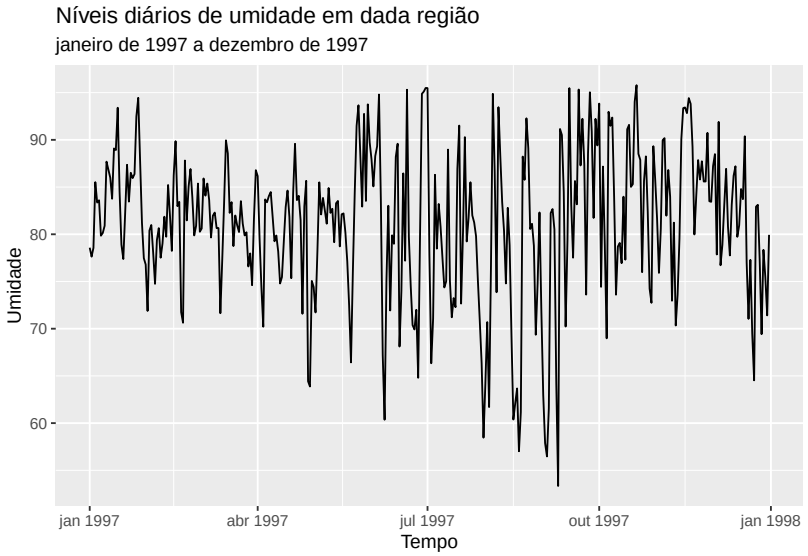
14 de julho de 2024

Sumário

Série

A série considerada para análise é a de **Umidade**, presente no conjunto *Atmosfera* e disponível neste link. A mesma constitui uma série diária, que vai de 01 de janeiro de 1997 a 31 de dezembro do mesmo ano, e trata dos níveis de umidade em dada região.

Primeiramente, podemos visualizar a série e fazer algumas pontuações acerca do seu comportamento, tendo em mente a medologia Box-Jenkins para o processo de modelagem.



Pelo gráfico, é fácil perceber que não temos a presença de tendência ou sazonalidade, sendo razoável supor então que a série já seja estacionária, com média constante próxima de 80. Também é razoável supor que a série apresente baixa variância ou desvio padrão, pois o gráfico não aparenta ter muitas variações bruscas de mais (como várias quedas ou picos muito fortes).

Vamos atestar essas inferências a partir dos testes estatísticos *KPSS* e o *teste aumentado de Dickey-Fuller*.

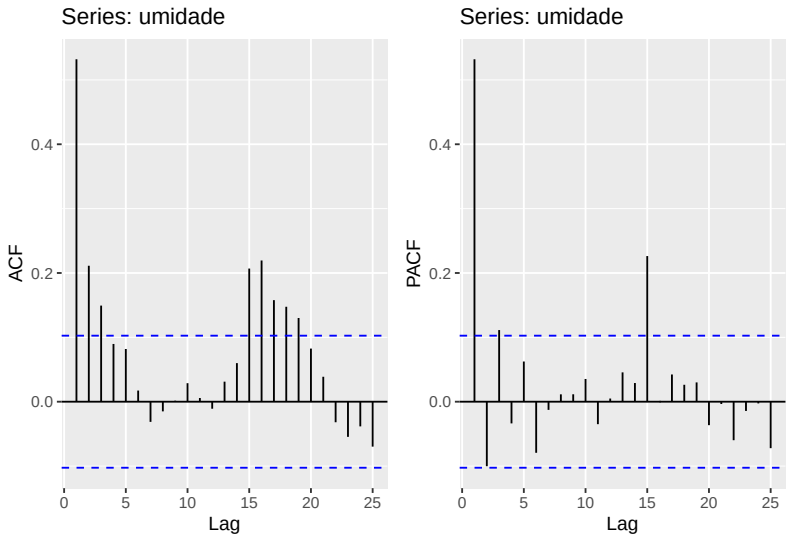
```
##
##  KPSS Test for Level Stationarity
##
## data:  umidade
## KPSS Level = 0.18604, Truncation lag parameter = 5, p-value = 0.1
```

Pelo teste *KPSS*, não temos evidências suficientes para rejeitar a hipótese nula de estacionariedade da série.

```
##
##  Augmented Dickey-Fuller Test
##
## data:  umidade
## Dickey-Fuller = -5.9163, Lag order = 7, p-value = 0.01
## alternative hypothesis: stationary
```

Da mesma maneira, o *teste aumentado de Dickey-Fuller* nos aponta que os dados sejam de fato estacionários.

Para a sugestão de modelos, é necessário também que analisemos as funções de autocorrelação e autocorrelação parcial abaixo, pois estas nos auxiliam a escolher a ordem de tais modelos.



Também não devemos esquecer de dividir a nossa série em conjunto de treino e teste, a fim de avaliar a performance do ajuste e de previsões:

```
treino = umidade['1997-01-01/1997-12-24']
teste = umidade['1997-12-25/']
```

Com isso, estamos prontos para ajustar alguns modelos.

Primeiro modelo

A partir do gráfico das funções de autocorrelação e autocorrelação parcial, notamos que o primeiro tem um comportamento aproximadamente sinusoidal, enquanto o segundo decai rapidamente para 0 logo após o primeiro lag, mas possuindo ainda dois lags significantes, ainda que um deles (lag 3) esteja próximo do intervalo de confiança.

Tal comportamento pode ser indicativo de um modelo ARIMA(1,0,0), ou simplesmente um modelo AR(1). Desta forma, ajustamos esse modelo no R, encontrando as estimativas dos parâmetros. É importante perceber valores altos para os critérios da informação.

```
fit1 = treino |> Arima(order = c(1,0,0)); fit1
```

```
## Series: treino
## ARIMA(1,0,0) with non-zero mean
##
## Coefficients:
##          ar1     mean
##       0.5306  81.2416
## s.e.  0.0446   0.7577
##
## sigma^2 = 45.83:  log likelihood = -1191.79
## AIC=2389.58   AICc=2389.65   BIC=2401.22
```

Testando a significância dos parâmetros através do *teste t* bicaudal, e utilizando um nível descritível de 5%, chegamos à conclusão de que ambos os parâmetros estimados são significativos para o modelo. A tabela a seguir apresenta as estimativas, estatísticas e valor crítico dos testes.

Coef	Val	S.E.	|Tcalc|	Ttab
ar1	0.5306	0.0446	11.8902	1.9666
intercept	81.2416	0.7577	107.2238	1.9666

Iremos agora checar a estacionariedade dos resíduos do modelo, a fim de atestar se os mesmos comportam-se como ruídos brancos, ou seja, com média e variância constantes.

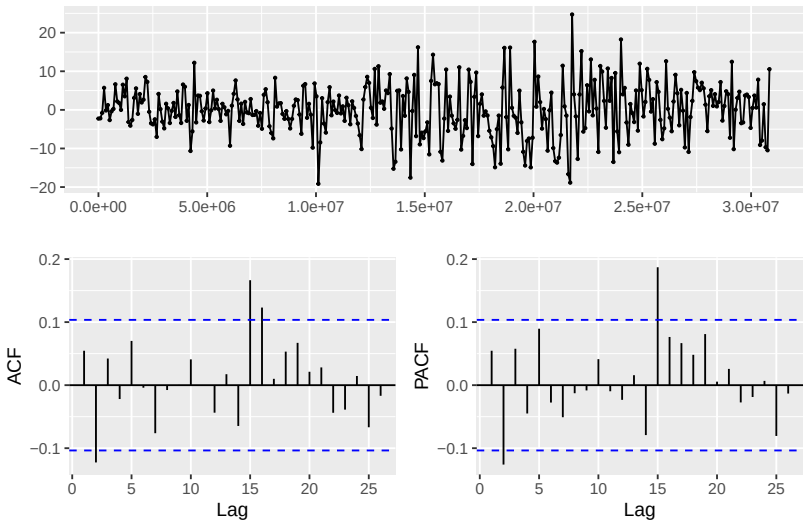
Realizando primeiro o *teste aumentado de Dickey-Fuller* e o teste *KPSS* abaixo, temos indícios de que os resíduos sejam estacionários. Também podemos aplicar o *teste Ljung-Box* e testar a hipótese de estacionariedade através dele.

```
##
##  KPSS Test for Level Stationarity
##
## data:  fit1$residuals
## KPSS Level = 0.16937, Truncation lag parameter = 5, p-value = 0.1
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  fit1$residuals
## Dickey-Fuller = -6.7864, Lag order = 7, p-value = 0.01
## alternative hypothesis: stationary
```

```
##
##  Box-Ljung test
##
## data:  fit1$residuals
## X-squared = 1.0759, df = 1, p-value = 0.2996
```

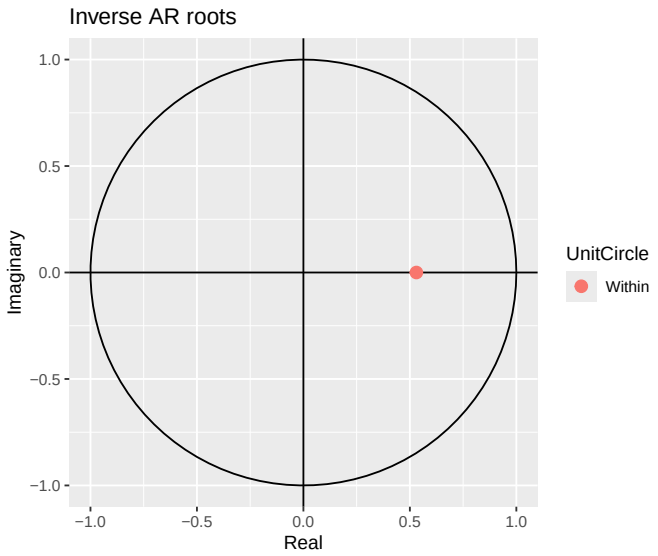
Devemos também checar os gráficos das funções de autocorrelação e autocorrelação parcial dos resíduos.



Vemos que apesar do comportamento da série realmente parecer a de um processo aleatório, não queremos que haja pontos significantes nas funções de autocorrelação, o que não é o caso para os resíduos do nosso modelo, que apresentam pelo menos dois desses pontos em cada gráfico.

Além disso, também devemos atestar a normalidade dos resíduos, e checar que a raíz inversa do polinômio esteja dentro do círculo unitário. Ambos os testes também atestam as suposições do nosso modelo.

```
##
##  Shapiro-Wilk normality test
##
## data:  fit1$residuals
## W = 0.99414, p-value = 0.1847
```

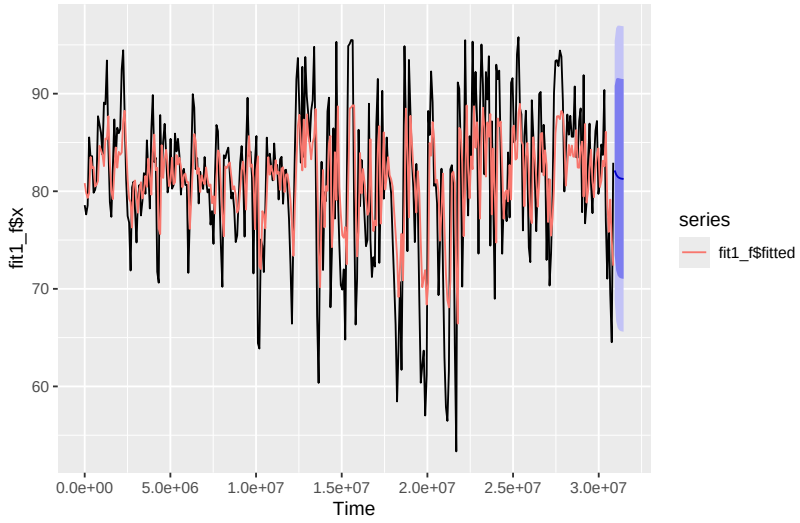


Por fim, com o modelo validado, podemos fazer uma previsão de horizonte igual a 7, compará-la ao conjunto de teste, extrair algumas medidas de erro e de qualidade de ajuste, para, posteriormente, fazer uma comparação com outras sugestões de modelos. Isso é realizado da seguinte maneira:

```
fit1_f = forecast(fit1, h = 7)
metricas_fit1 = accuracy(fit1_f, x = teste)
```

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Treino	0.0049	6.7505	5.2170	-0.7487	6.6453	0.0642	0.0546
Teste	-5.1009	6.6545	5.3831	-7.0275	7.3671	0.0663	—

Também podemos plotar um gráfico do ajuste e da previsão, junto com o intervalo de confiança para a previsão.



Assim, temos valores baixos para as medidas de erro no geral, tanto no conjunto treino quanto no de teste, porém, ainda veremos outros modelos possíveis capazes de modelar a série, e escolher aquele com melhor ajuste.

Segundo modelo

Como já visto anteriormente com os gráficos de autocorrelação da série original, temos no segundo gráfico (PACF) 3 lags significativos, sendo eles os lags 1, 3 e 15. Como já ajustamos um ARIMA(1,0,0), poderíamos desta vez ajustar um ARIMA(3,0,0), olhando agora para o terceiro lag. O 15º lag, por ser uma observação muito distante, possivelmente seja um outlier, e não indica necessariamente que um modelo com 15 parâmetros deva ser ajustado, afinal, devemos prezar pelo critério da parcimônia, e então, não o levamos em consideração para a análise neste momento.

```
fit2 = treino |> Arima(order = c(3,0,0)); fit2
```

```
## Series: treino
## ARIMA(3,0,0) with non-zero mean
##
## Coefficients:
##          ar1      ar2     ar3     mean
##       0.5954  -0.1628  0.1043  81.2391
## s.e.  0.0526   0.0611  0.0529   0.7597
##
## sigma^2 = 45.12:  log likelihood = -1188.02
## AIC=2386.04   AICc=2386.21   BIC=2405.45
```

Assim, temos uma leve redução na variância estimada, *AIC* e *AICc*. Testando a significância dos parâmetros, também obtemos a tabela a seguir, e vemos que todos eles são significativos para o modelo.

Coef	Val	S.E.	|Tcalc|	Ttab
ar1	0.5954	0.0526	11.3092	1.9667
ar2	-0.1628	0.0611	2.6635	1.9667
ar3	0.1043	0.0529	1.9727	1.9667
intercept	81.2391	0.7597	106.9365	1.9667

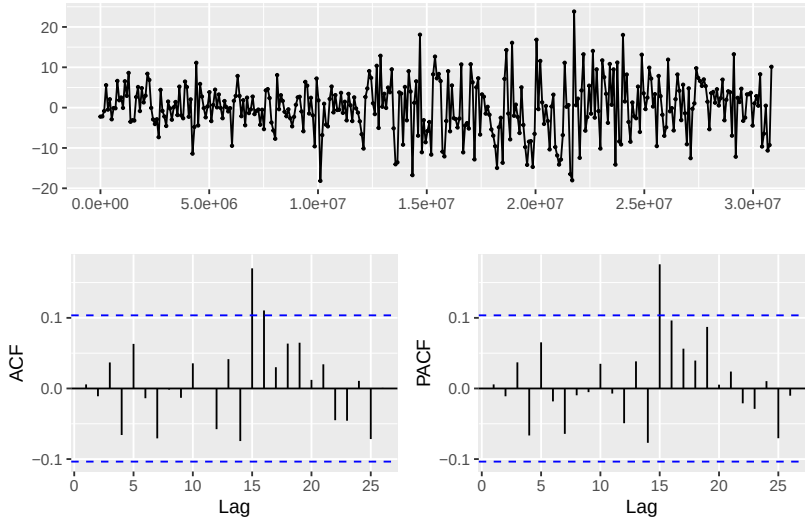
Realizamos os testes abaixo e confirmamos que os resíduos comportam-se como ruído branco.

```
##
##  KPSS Test for Level Stationarity
##
## data:  fit2$residuals
## KPSS Level = 0.16328, Truncation lag parameter = 5, p-value = 0.1
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  fit2$residuals
## Dickey-Fuller = -6.8857, Lag order = 7, p-value = 0.01
## alternative hypothesis: stationary
```

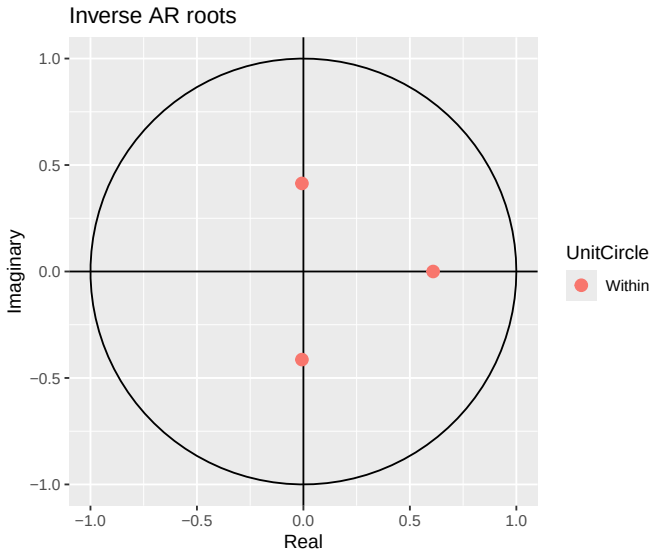
```
##
##  Box-Ljung test
##
## data:  fit2$residuals
## X-squared = 0.011985, df = 1, p-value = 0.9128
```

E plotando as funções de autocorrelação, junto da série residual, vemos que parece sim haver aleatoriedade, e também notamos menos lags significantes do que no modelo anterior.



Vemos então que o teste de *Shapiro-Wilk* aponta normalidade, e que as raízes inversas do polinômio estão dentro do círculo unitário.

```
##
##  Shapiro-Wilk normality test
##
## data:  fit2$residuals
## W = 0.99416, p-value = 0.1874
```

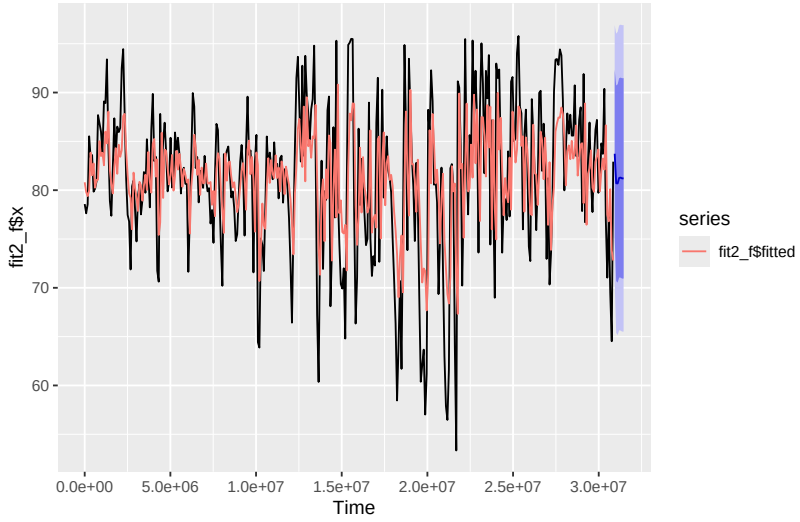


Agora, realizamos então uma previsão para o modelo de teste, e obtemos as medidas de erro a seguir.

```
fit2_f = fit2 |> forecast(h = 7)
metricas_fit2 = accuracy(fit2_f, x = teste)
```

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Treino	0.0054	6.6794	5.1538	-0.7322	6.5615	0.0634	0.0058
Teste	-5.0337	6.3240	5.0337	-6.9021	6.9021	0.0620	—

Vemos então uma leve redução na maioria dessas medidas para ambos os conjuntos. Plotando agora o gráfico do ajuste e da previsão, com o intervalo de confiança para a mesma, podemos ver que houve uma leve melhora.



Portanto, temos também um bom ajuste, levemente melhor comparado ao primeiro modelo, todavia, ainda temos lags significantes presentes nas funções de autocorrelação, o que será agora o objetivo principal no próximo modelo sugerido.

Terceiro modelo

Já sabendo que o 15° lag ainda possui influência mesmo em um ARIMA(3,0,0), podemos então levá-lo em consideração na modelagem, sem necessariamente adicionar os lags 4 ao 14, tendo em vista que desejamos um modelo o mais simples possível. Isso é feito no R com o código abaixo.

```
fit3 = treino |> Arima(order = c(15,0,0), fixed = c(NA, NA, NA, rep(0, 11), NA, NA)); fit3
```

```
## Series: treino
## ARIMA(15,0,0) with non-zero mean
##
## Coefficients:
##          ar1      ar2     ar3  ar4  ar5  ar6  ar7  ar8  ar9  ar10  ar11  ar12
##       0.5824  -0.1647  0.1088    0    0    0    0    0    0     0     0     0
## s.e.  0.0514   0.0596  0.0515    0    0    0    0    0    0     0     0     0
##       ar13  ar14    ar15     mean
##          0     0  0.1805  81.2506
## s.e.     0     0  0.0432   1.1401
##
## sigma^2 = 43.07:  log likelihood = -1179.56
## AIC=2371.12   AICc=2371.36   BIC=2394.41
```

Com isso, é possível perceber a grande diferença de redução nos critérios da informação e na variância, mesmo com um parâmetro a mais estimado. Também é importante mencionar que como não tomamos diferenças na série, há então uma média μ adicionada no modelo, que também é estimada.

Verificamos então a significância de todos os parâmetros estimados com a tabela abaixo, podendo ver que todos eles de fato são significativos para o modelo.

Coef	Val	S.E.	|Tcalc|	Ttab
ar1	0.5824	0.0514	11.3264	1.9667
ar2	-0.1647	0.0596	2.7644	1.9667
ar3	0.1088	0.0515	2.1110	1.9667
ar15	0.1805	0.0432	4.1767	1.9667
intercept	81.2506	1.1401	71.2692	1.9667

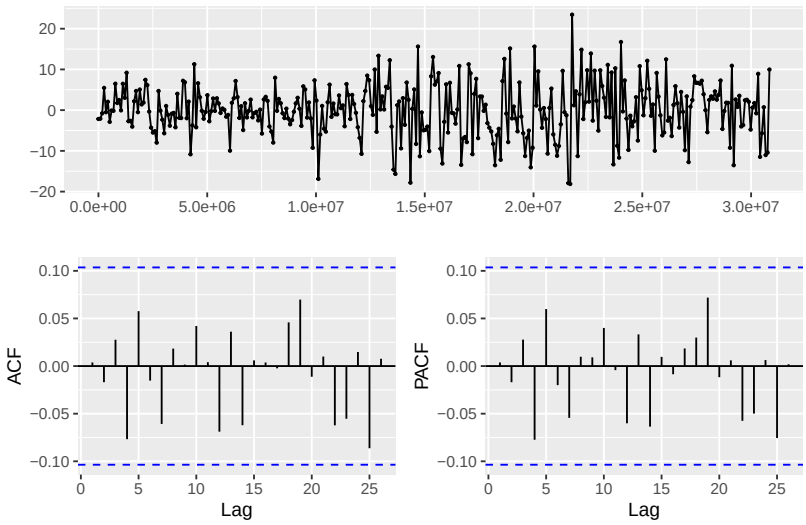
Em seguida, vamos usar os mesmos testes para verificar que há estacionariedade nos resíduos do nosso modelo.

```
##
##  KPSS Test for Level Stationarity
##
## data:  fit3$residuals
## KPSS Level = 0.084914, Truncation lag parameter = 5, p-value = 0.1
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  fit3$residuals
## Dickey-Fuller = -6.7754, Lag order = 7, p-value = 0.01
## alternative hypothesis: stationary
```

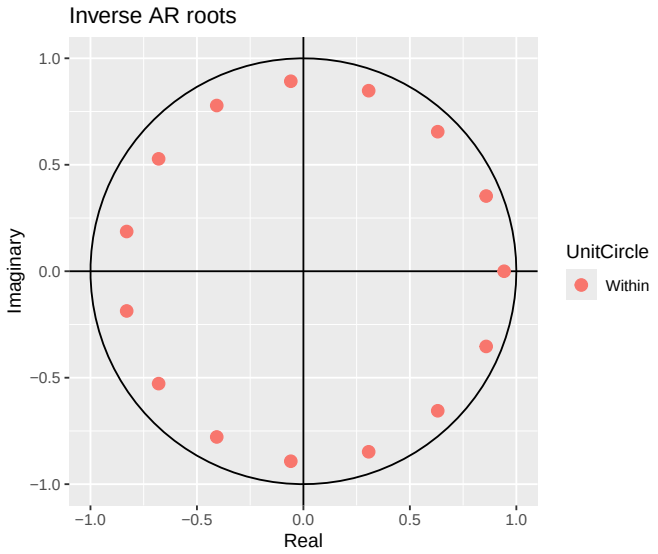
```
##
##  Box-Ljung test
##
## data:  fit3$residuals
## X-squared = 0.0051892, df = 1, p-value = 0.9426
```

Checamos os gráfico da série residual e autocorrelações, e vemos que finalmente não há nenhum lag significante em qualquer dos gráficos, além de que a série comporta-se como um ruído branco.



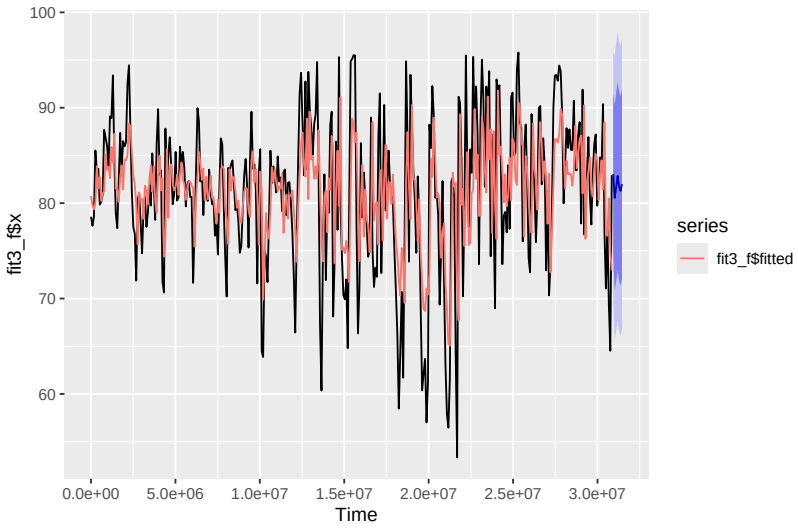
Utilizando o teste de normalidade e das inversas das raízes do polinômio, também temos esta parte da validação atendida.

```
##
##  Shapiro-Wilk normality test
##
## data:  fit3$residuals
## W = 0.99332, p-value = 0.1136
```



Logo, terminamos nossa validação extraindo as medidas de erro para o conjunto de teste, e plotamos em seguida o ajuste sobre a série, junto da previsão ao respectivo intervalo de confiança.

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Treino	-0.0114	6.5168	4.9945	-0.7167	6.3459	0.0615	0.0038
Teste	-5.4659	6.7805	5.4815	-7.4838	7.5025	0.0675	—



Modelo final

Ao encontrar um modelo que atende a todos os nossos pressupostos, encerramos o processo iterativo do método de Box-Jenkins, podendo então especificar o modelo, ajustá-lo a série completa, e realizar previsões. Logo, com o passo 3, encontramos um $ARIMA(15,0,0)$, com $\phi_4, \ldots, \phi_{14} = 0$, e média μ diferente de zero, de forma

$$(1 - \phi_1 B - \phi_2 B^2 - \phi_3 B^3 - \phi_{15} B^{15}) Z_t = \mu + a_t,$$

onde as estimativas para os coeficientes e a média são

- $\hat{\phi}_1 \approx 0,5824$
- $\hat{\phi}_2 \approx -0,1647$
- $\hat{\phi}_3 \approx 0,1088$
- $\hat{\phi}_{15} \approx 0,1805$
- $\hat{\mu} \approx 81,2506$

com $\hat{\sigma_a^2} \approx 43,07$.

Com a especificação do modelo, fazemos o ajuste considerando toda a série.

```
m = umidade |> Arima(order = c(15,0,0), fixed = c(rep(NA, 3), rep(0,11), rep(NA, 2))); m
```

```
## Series: umidade
## ARIMA(15,0,0) with non-zero mean
##
## Coefficients:
##          ar1      ar2     ar3  ar4  ar5  ar6  ar7  ar8  ar9  ar10  ar11  ar12
##       0.5839  -0.1679  0.1154    0    0    0    0    0    0     0     0     0
## s.e.  0.0507   0.0586  0.0506    0    0    0    0    0    0     0     0     0
##       ar13  ar14    ar15     mean
##          0     0  0.1795  81.0786
## s.e.     0     0  0.0430   1.1410
##
## sigma^2 = 42.76:  log likelihood = -1201.32
## AIC=2414.65   AICc=2414.88   BIC=2438.05
```

Realizamos então uma previsão, de uma semana, por exemplo.

```
m_prev = forecast(m, h = 7)
```

	prev	ic_l80	ic_u80	ic_l95	ic_u95
1998-01-01	81.8642	73.4844	90.2439	69.0484	94.6799
1998-01-02	82.2824	72.5788	91.9860	67.4420	97.1227
1998-01-03	80.7860	70.9748	90.5973	65.7810	95.7911
1998-01-04	78.9943	69.1330	88.8556	63.9128	94.0759
1998-01-05	79.3696	69.4673	89.2719	64.2253	94.5139
1998-01-06	78.3131	68.3968	88.2294	63.1475	93.4788
1998-01-07	76.5398	66.6198	86.4598	61.3685	91.7111

Temos então uma previsão que varia entre 81,8 e 76,5, indicando que, durante a primeira semana de 1998, a umidade relativa do ar foi decaindo, possivelmente indicando um tempo um pouco mais seco, porém ainda dentro do ideal (faixa entre 50% a 100%).