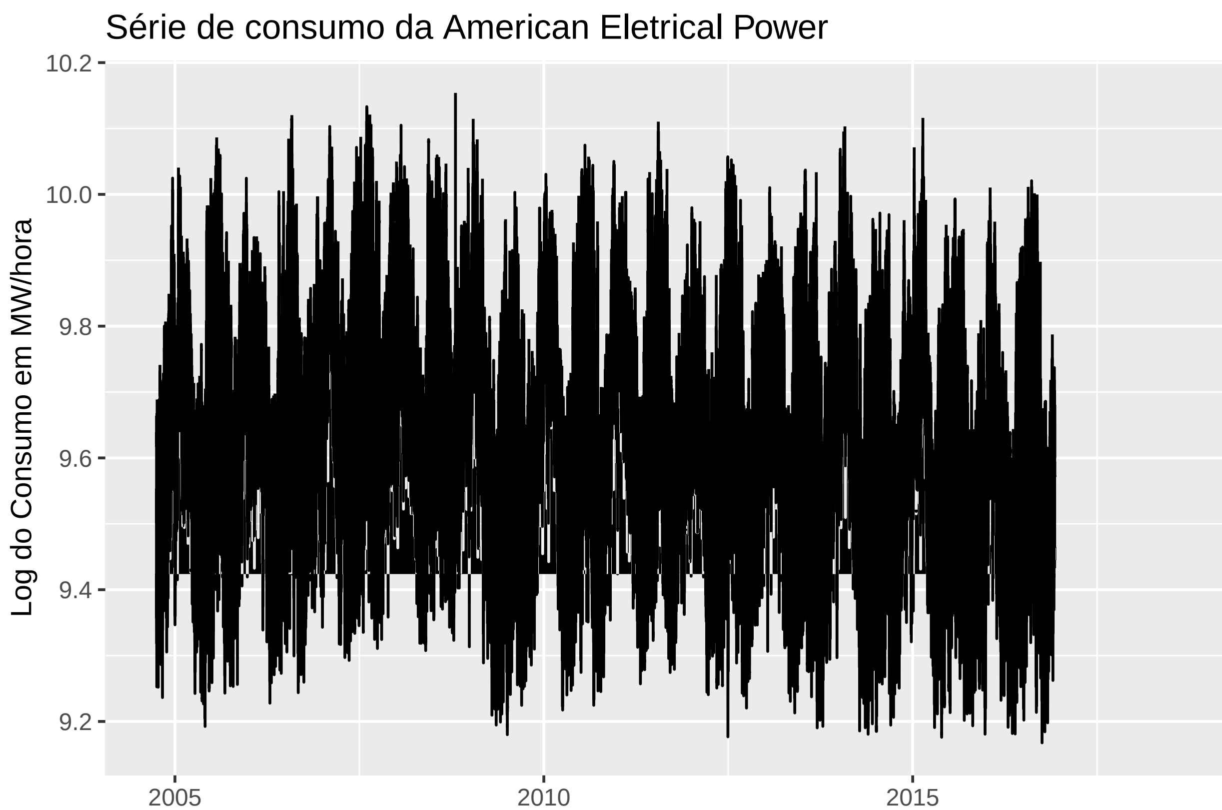
Trabalho de Análise de Série Temporais I

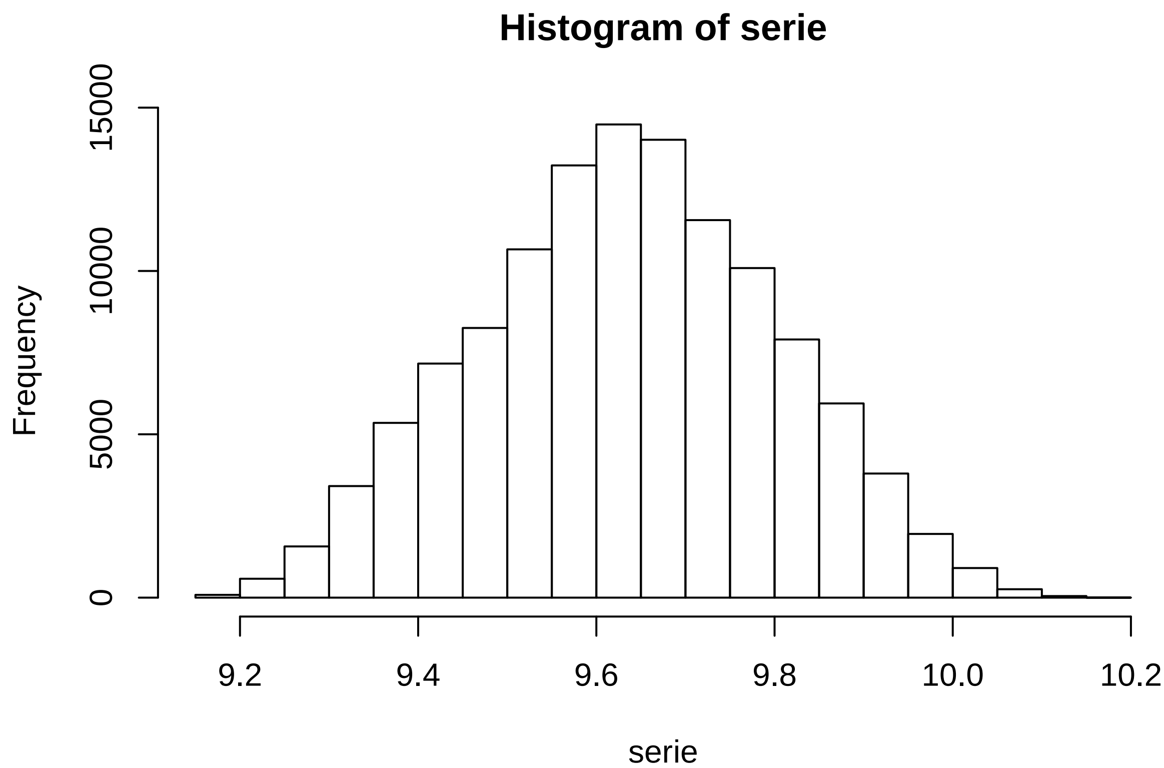
Pedro Cavalcante

Amostra

A base de dados usada contém séries históricas de consumo de energia elétrica por hora de subsidiárias regionais dos EUA. Como as grandezas envolvidas são grandes porém não particularmente informativas usarei logarítimo natural da série.



A série é de altíssima frequência e cobre um horizonte razoavelmente longo. Ao todo temos `121269` observações. Abaixo reproduzo deixo algumas estatísticas e visualizações descritivas simples. Elas são, em ordem, histograma, sumário da distribuição, quatro primeiros momentos amostrais centrais e gráfico de Cullen-Frey.



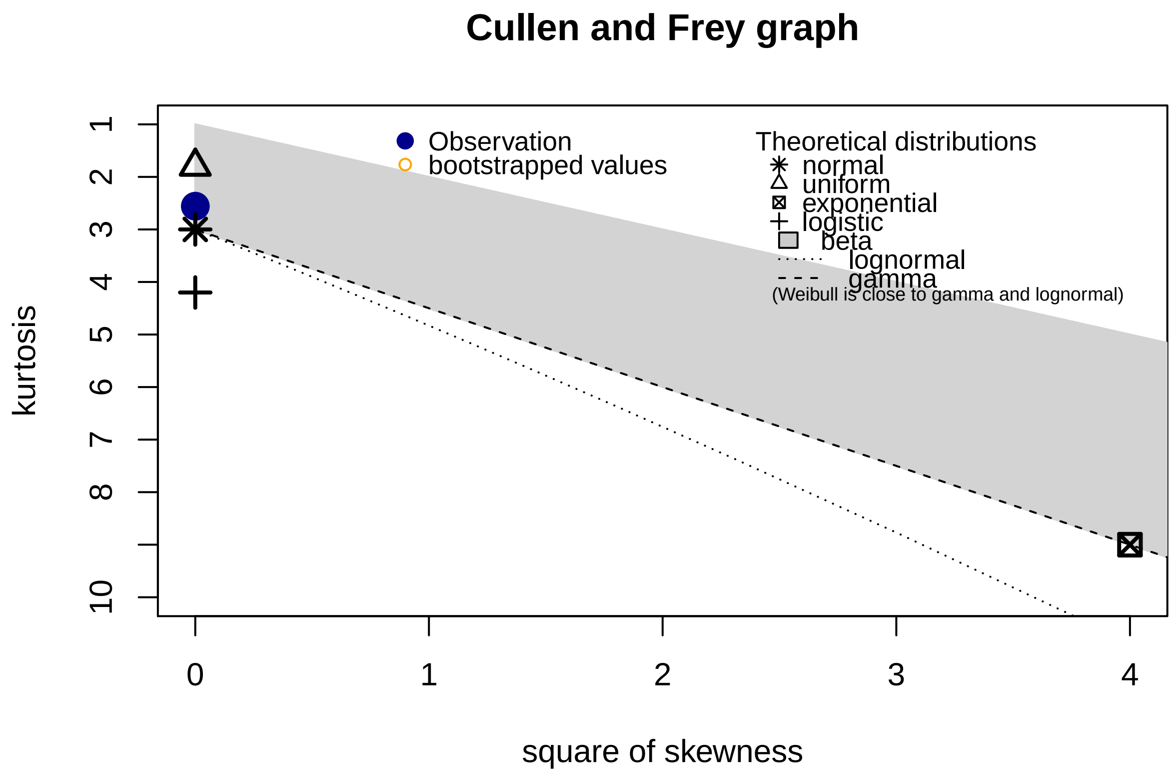

```
##       Index                            serie
##  Min.   :2004-10-01 01:00:00   Min.   : 9.168
##  1st Qu.:2008-03-17 14:00:00   1st Qu.: 9.520
##  Median :2011-09-02 02:00:00   Median : 9.636
##  Mean   :2011-09-02 01:55:58   Mean   : 9.635
##  3rd Qu.:2015-02-16 15:00:00   3rd Qu.: 9.753
##  Max.   :2018-08-03 00:00:00   Max.   :10.154
```

```
## [1] -4.184954e-16
```

```
## [1] 0.0278302
```

```
## [1] -6.921959e-05
```

```
## [1] 0.001981162
```




```
## summary statistics
## ------
## min:  9.167537   max:  10.15405
## median:  9.636261
## mean:  9.634683
## estimated sd:  0.1668245
## estimated skewness:  -0.01490939
## estimated kurtosis:  2.557954

## Warning in ks.test(x = as.numeric(serie), normal_simulada): p-value will be
## approximate in the presence of ties

##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  as.numeric(serie) and normal_simulada
## D = 0.015709, p-value = 2.016e-13
## alternative hypothesis: two-sided
```

A assimetria com viés para a direita da média e a *leve* cauda à direita, bem como o gráfico de Cullen-Frey sugerem que os dados seguem um processo gerador próximo de uma lognormal. Os resultados do teste Kolmogorov-Smirnof corroboram a não-normalidade dos dados.

Estacionariedade

Os resultados do teste ADF indicam ausência de raiz unitária.

```r
adf.test(serie,
         alternative = "explosive")
```

```
## Warning in adf.test(serie, alternative = "explosive"): p-value smaller than
## printed p-value
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  serie
## Dickey-Fuller = -22.736, Lag order = 49, p-value = 0.99
## alternative hypothesis: explosive
```

Como evidência extra a favor, vamos estimar uma random walk com os dados na forma:

$$y_t = \beta y_{t-1} + \epsilon_t$$

E depois com um Teste de Wald testar a hipótese nula $H0 : \beta = 1$.

```r
randomwalk = lm(serie ~ lag(serie))
randomwalk %>% summary()
```

```
##
## Call:
## lm(formula = serie ~ lag(serie))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.46137 -0.02090 -0.00216  0.02233  0.45356
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.239339   0.006129   39.05   <2e-16 ***
## lag(serie)  0.975159   0.000636 1533.25   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03695 on 121266 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.9509, Adjusted R-squared:  0.9509
## F-statistic: 2.351e+06 on 1 and 121266 DF,  p-value: < 2.2e-16
```

```r
wald.test(b = coef(randomwalk),
          Sigma = vcov(randomwalk),
          H0 = c(1),
          Terms = 2)
```

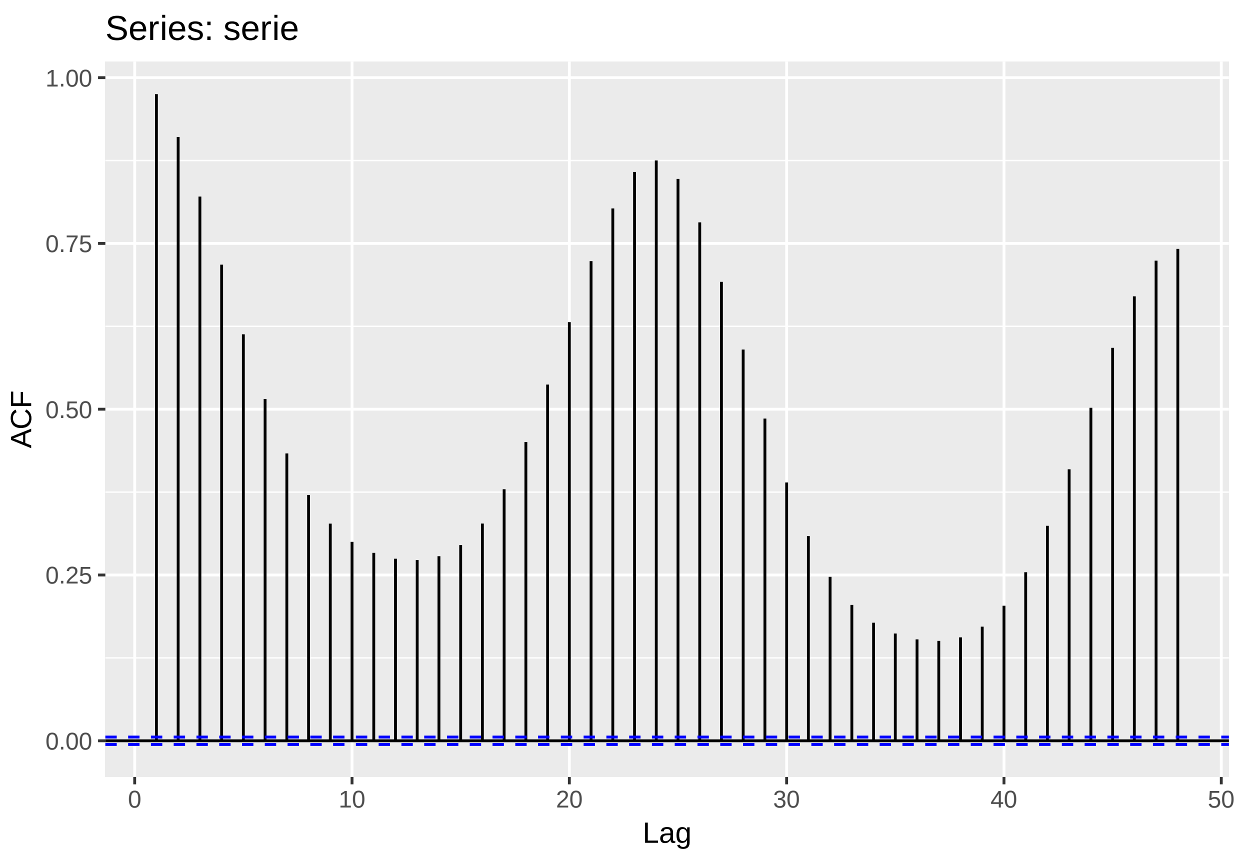
```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 1525.5, df = 1, P(> X2) = 0.0
```

Rejeitamos a hipótese com um baixíssimo p-valor. As evidências sugerem que de fato não temos raiz unitária.

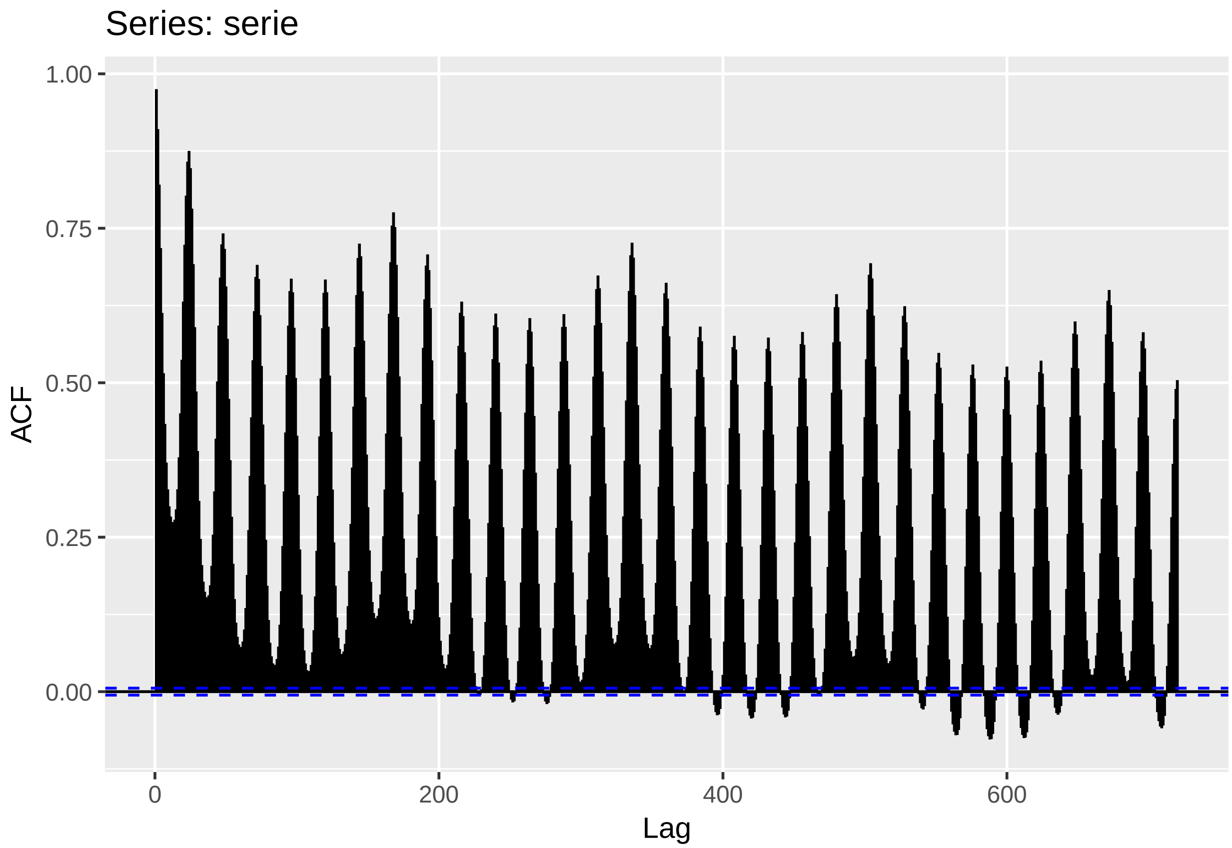
Autocorrelação

Graças à altíssima frequência da série podemos observar múltiplos padrões nos correlogramas. Existem padrões repetidos de consumo entre dias, semanas, meses e estações do ano.

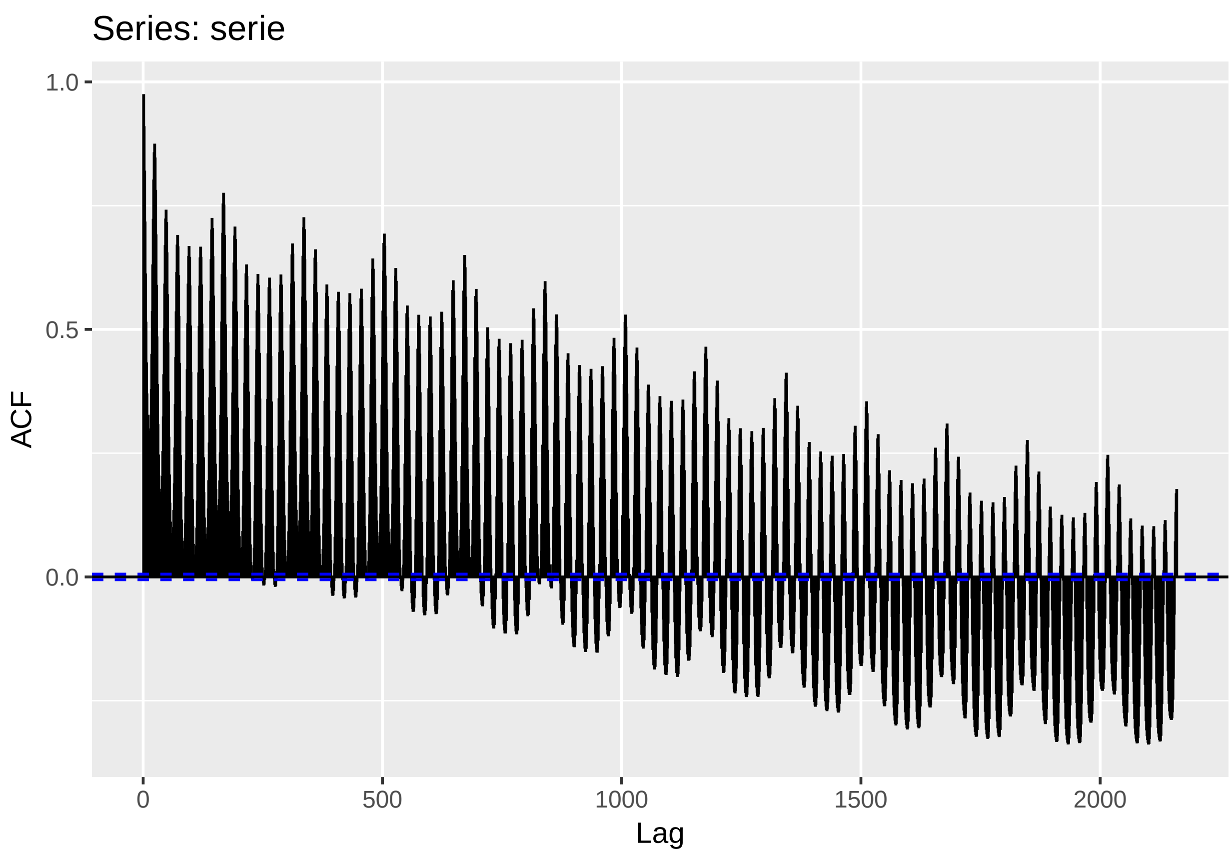
```r
ggAcf(serie, lag.max = 48) # dois dias
```



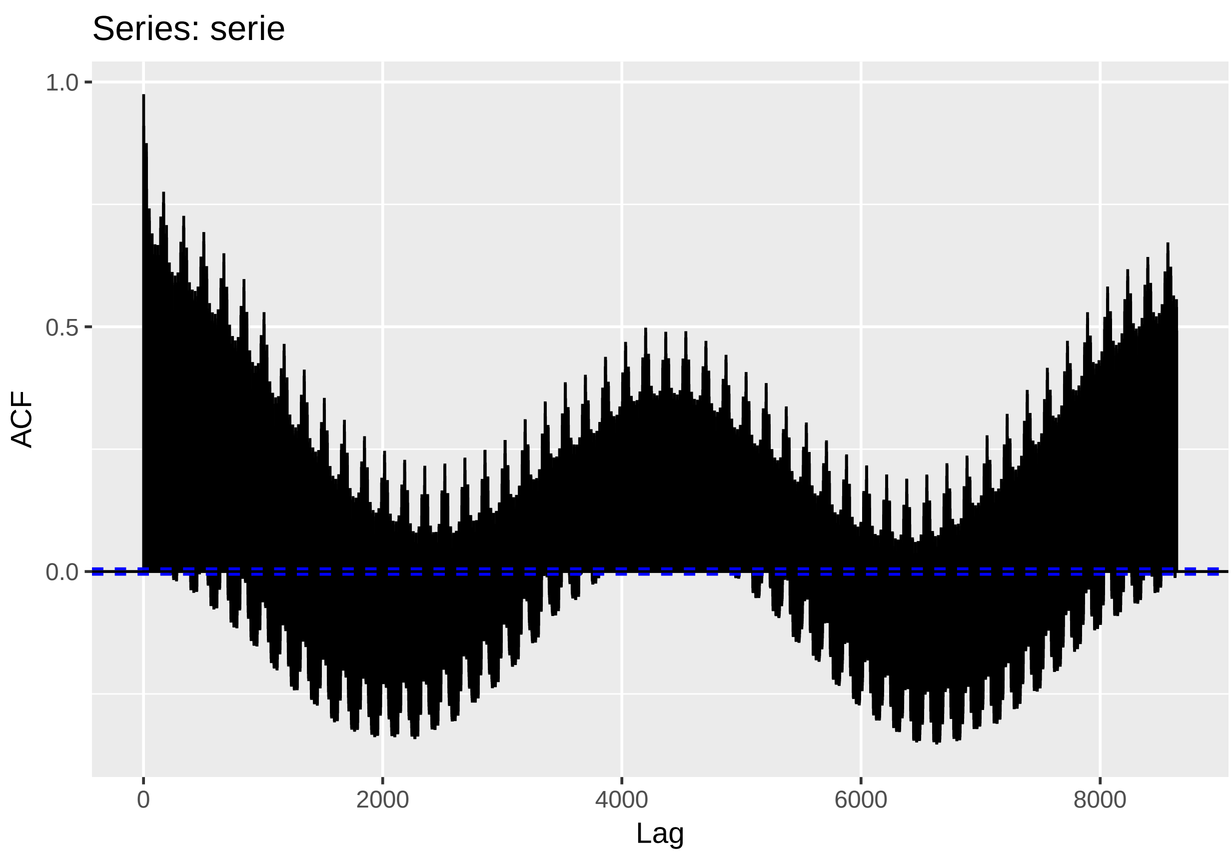

```r
ggAcf(serie, lag.max = (24*30)) # um mês
```



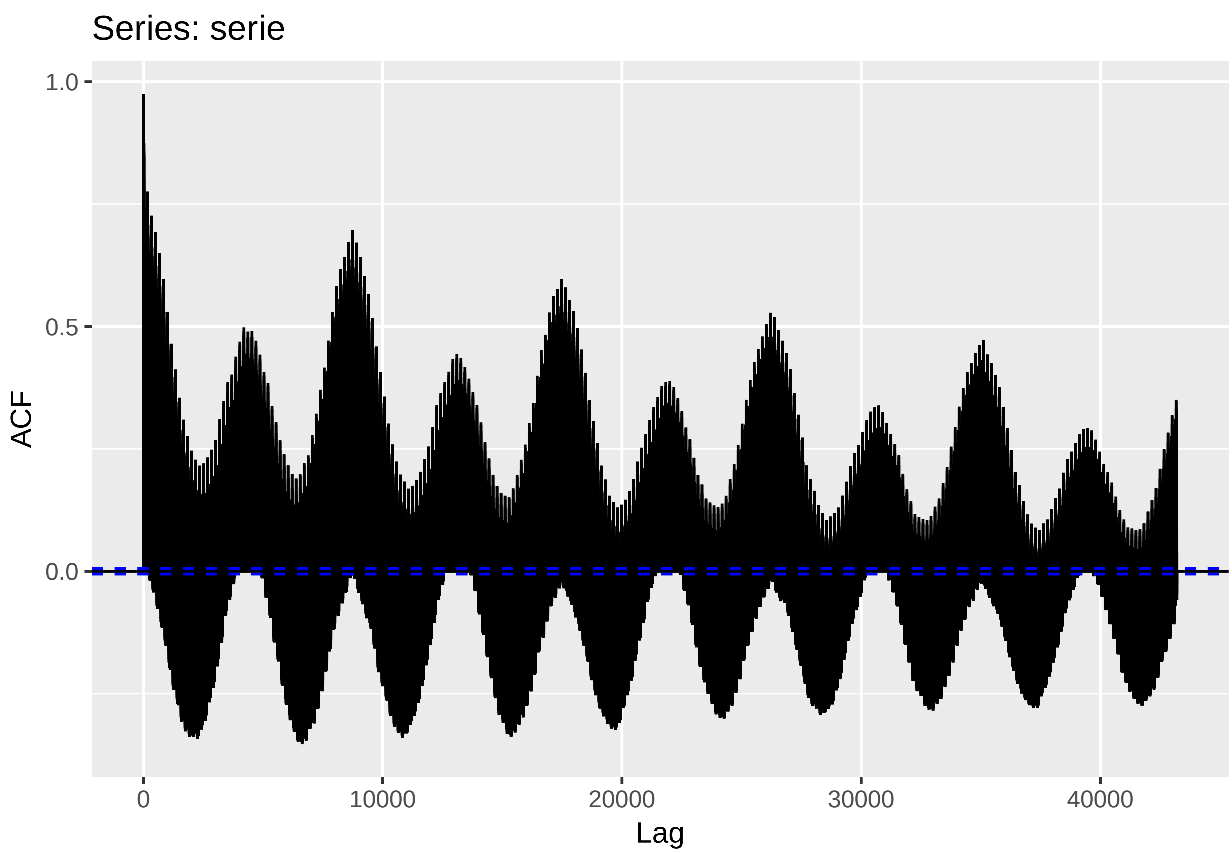

```
ggAcf(serie, lag.max = (24*30*3)) # um trimestre
```




```
ggAcf(serie, lag.max = (24*30*12)) # um ano
```

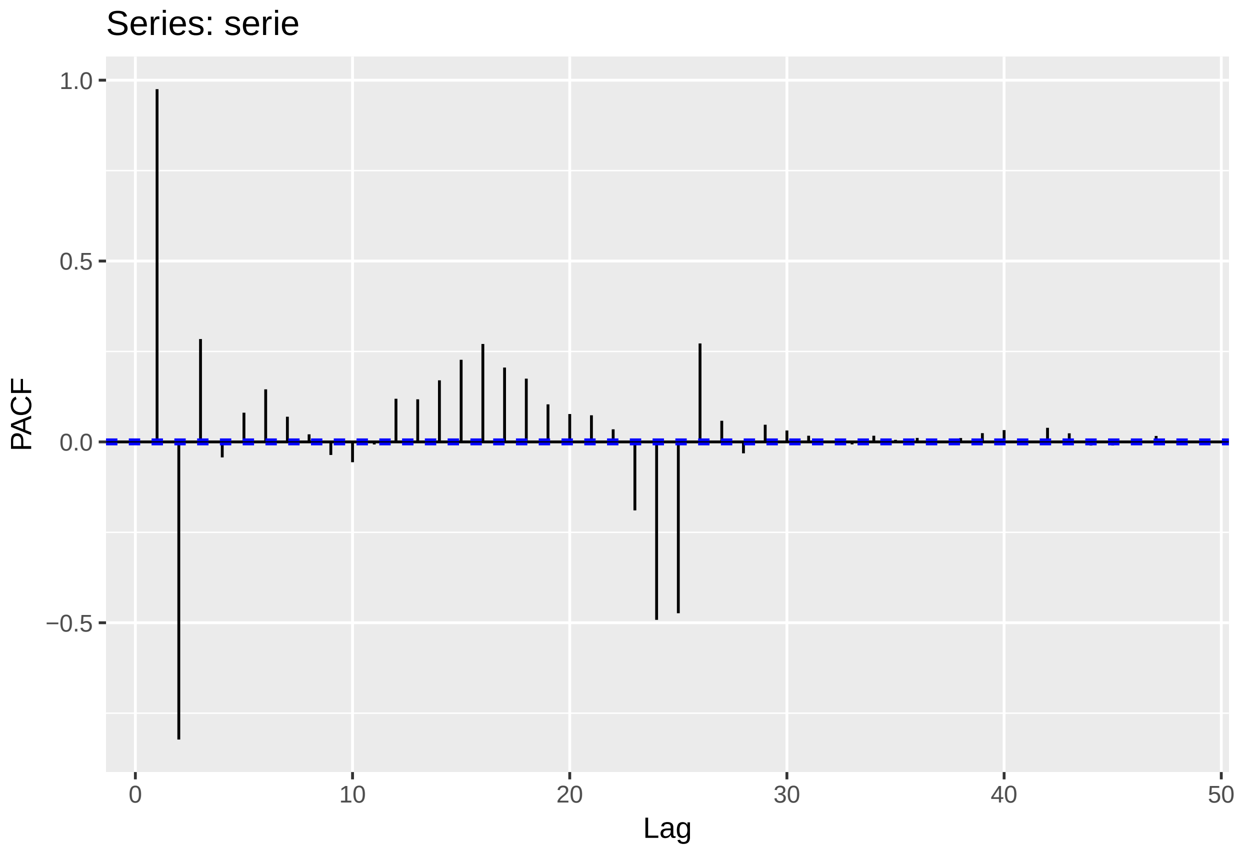



```
ggAcf(serie, lag.max = (24*30*12*5)) # cinco anos
```

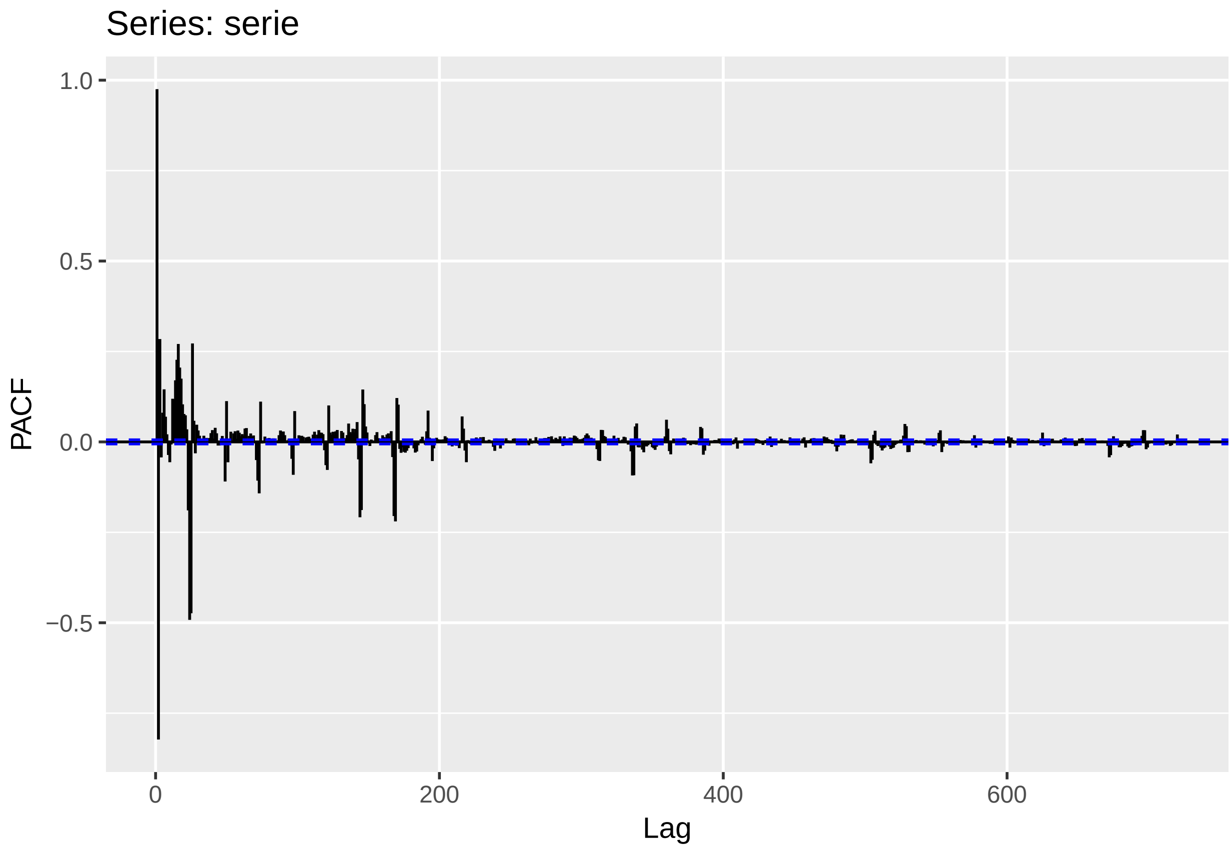


Como esperado de um processo AR de alta ordem, observamos um misto de decrescimento exponencial e comportamento sinoidal na função de autocorrelação. Padrões muito menos ricos aparecem nas autocorrelações parciais.

```
ggPacf(serie, lag.max = 48) # dois dias
```



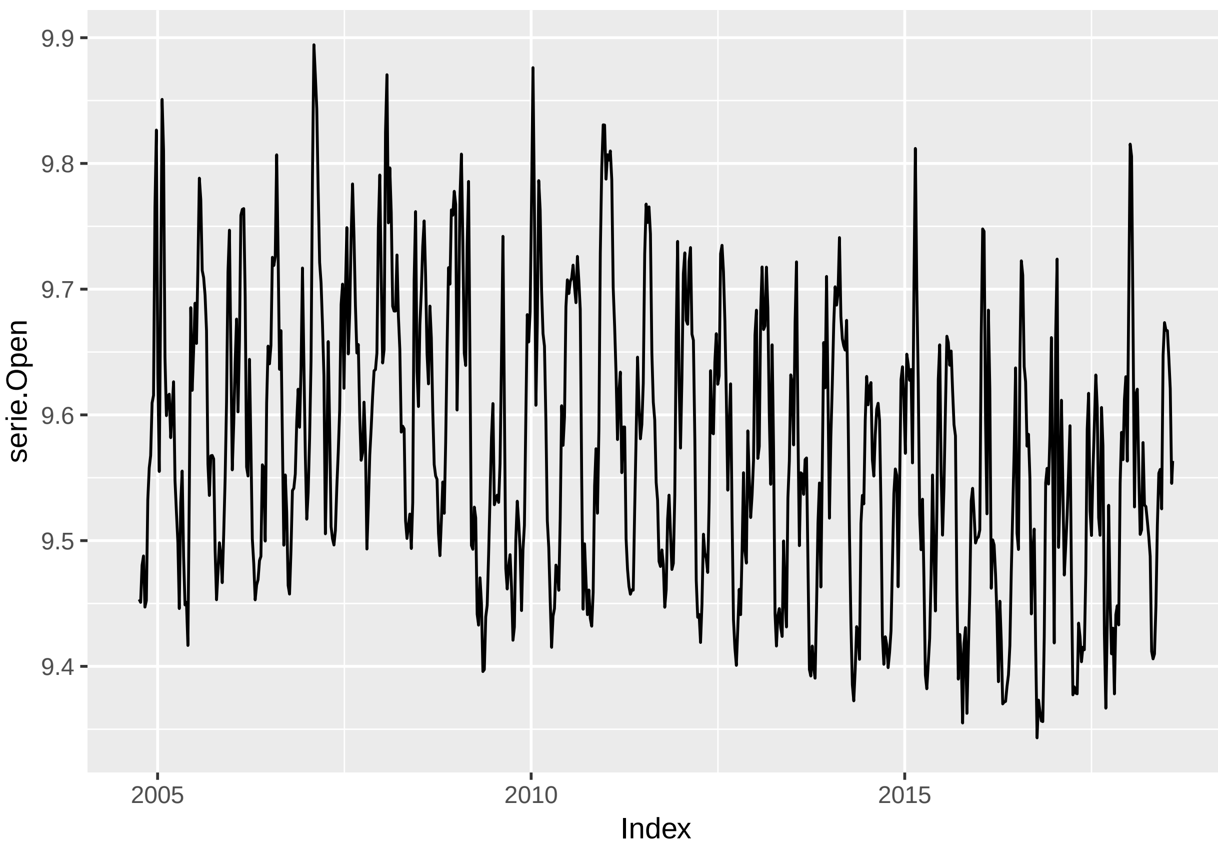

```
ggPacf(serie, lag.max = (24*30)) # um mês
```



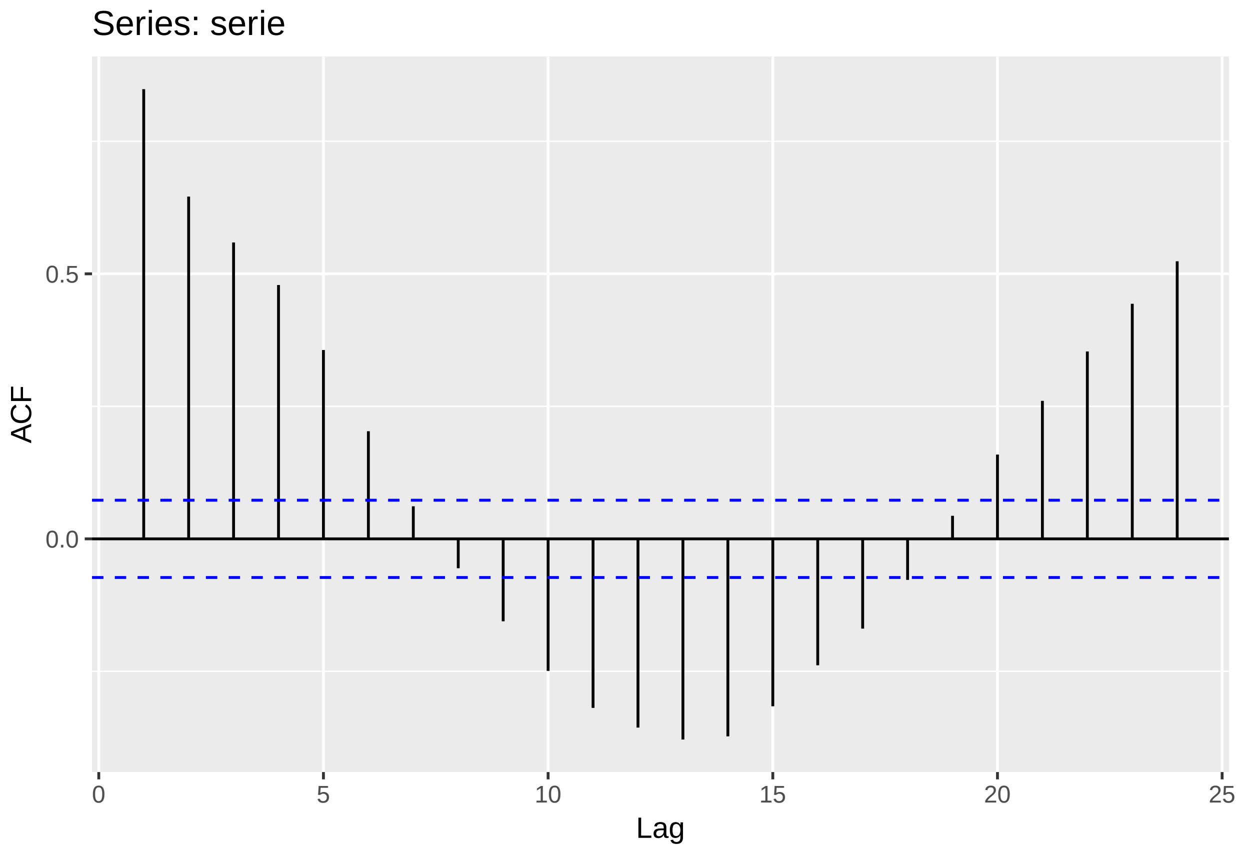
Uma pequena nota

A frequência da série é altíssima. Isso tende a piorar o *fitting* de modelos usualmente aplicados em dados de frequências mais baixas, por isso transformarei em uma série semanal.

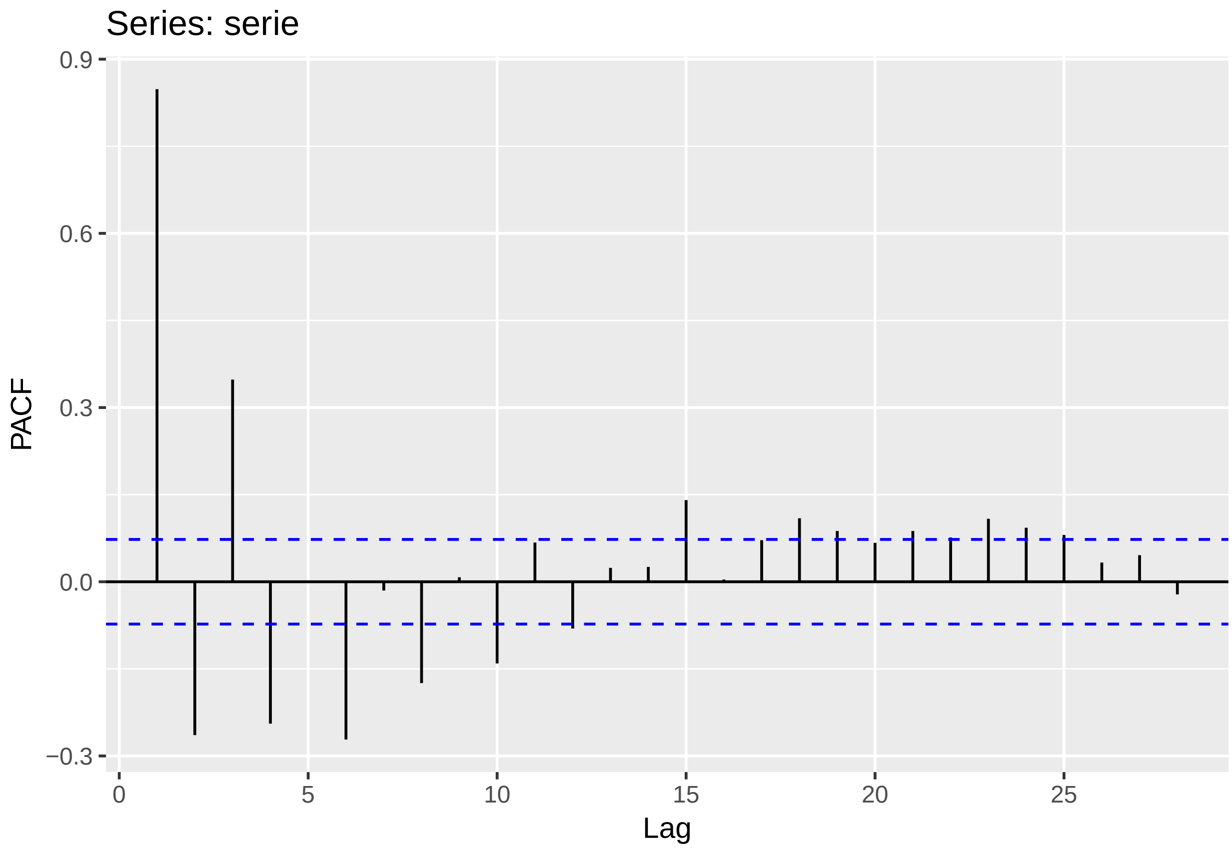
```
serie = (to.weekly(serie)$serie.Open + to.weekly(serie)$serie.Close)/2
autoplot(serie)
```




```r
ggAcf(serie, lag.max = 4*6) # um semestre
```




```r
ggPacf(serie)
```



Estimação e diagnósticos

A análise dos correlogramas e a série ser integrada de ordem 0 sugere que o melhor modelo é um ARIMA(2, 0, 3). Irei usar o algorítimo de seleção HK (Hyndman and Khandakar 2007) usando o Critério Bayesiano de Informação para colher evidências extras de que é um modelo apropriado. Como benchmark usarei um AR(1).

```
modelo = auto.arima(serie,
                    stationary = TRUE,
                    seasonal = TRUE,
                    allowmean = TRUE,
                    max.order = 2000,
                    ic = c("bic"))

summary(modelo)
```

```
## Series: serie
## ARIMA(2,0,3) with non-zero mean
##
## Coefficients:
##          ar1      ar2      ma1      ma2     ma3    mean
##       1.8879  -0.9435  -0.7209  -0.7602  0.6782  9.5743
## s.e.  0.0202   0.0198   0.0463   0.0269  0.0408  0.0063
##
## sigma^2 estimated as 0.002324:  log likelihood=1167.61
## AIC=-2321.21   AICc=-2321.06   BIC=-2289.13
```

```
##
## Training set error measures:
##                        ME       RMSE        MAE          MPE      MAPE
## Training set 2.277081e-05 0.04800789 0.03762043 -0.002432759 0.3924046
##                   MASE       ACF1
## Training set 0.7916707 0.08824254
```

```
benchmark = lm(serie ~ lag(serie))
benchmark %>% summary()
```

```
##
## Call:
## lm(formula = serie ~ lag(serie))
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.206862 -0.037475 -0.001666  0.035114  0.221597
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.45202    0.18838   7.708 4.25e-14 ***
## lag(serie)   0.84836    0.01967  43.123  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0592 on 720 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.7209, Adjusted R-squared:  0.7205
## F-statistic:  1860 on 1 and 720 DF,  p-value: < 2.2e-16
```

```
RMSE_benchmark = sqrt(sum(benchmark$residuals^2) / benchmark$df)
RMSE_arima = sqrt(sum(modelo$residuals^2) / (modelo$nobs - length(modelo$coef)))

melhora = (round(RMSE_benchmark/RMSE_arima, digits = 3) - 1)*100
```

O modelo ARIMA apresenta um RMSE `22.8` por cento abaixo do benchmark. Procuramos agora por autocorrelaçãos serial e normalidade dos resíduos.

Para autocorrelação, uma escolha natural seria usar o teste Ljung-Box, mas como mostra Hayashi (2000), a estatística Q do teste converge assintoticamente em distribuição para uma chi-quadrado se e somente se $\mathbb{E}[y_t|y_{t-1}, y_{t-2}, ...] = 0$, o que claramente não procede para modelos ARMA(P, Q) se $P > 0$. Maddala and Lahiri (1992) sugere o uso do teste de Breusch and Pagan (1979).

```
primeirolag = stats::lag(serie)
segundo = stats::lag(serie, k = 2)
terceiro = stats::lag(serie, k = 3)

bptest(serie ~ primeirolag + segundo + terceiro)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  serie ~ primeirolag + segundo + terceiro
## BP = 7.2179, df = 3, p-value = 0.06527
```

```
jarque.bera.test(modelo$residuals)
```

```
##
##  Jarque Bera Test
##
## data:  modelo$residuals
## X-squared = 6.1273, df = 2, p-value = 0.04672
```

Não rejeitamos a hipótese nula de homocedasticidade do Teste de Breusch-Pagan a 5% de significânia e marginalmente rejeitamos normalidade nos resíduos no mesmo nível de significância. A 1% de significância, temos homocedasticidade e normalidade dos resíduos. Dado o tamanho atípico da amostra, parece ser razoável requerer níveis menores que o usual de significância.

Referências

Breusch, Trevor S, and Adrian R Pagan. 1979. "A Simple Test for Heteroscedasticity and Random Coefficient Variation." *Econometrica: Journal of the Econometric Society.* JSTOR, 1287–94.

Hayashi, Fumio. 2000. "Econometrics." *Princeton University Press* 1: 60–69.

Hyndman, Rob J, and Yeasmin Khandakar. 2007. *Automatic Time Series for Forecasting: The Forecast Package for R.* 6/07. Monash University, Department of Econometrics; Business Statistics.

Maddala, Gangadharrao S, and Kajal Lahiri. 1992. *Introduction to Econometrics.* Vol. 2. Macmillan New York.