Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

Eduardo Moreira Araújo

Francisco Iago dos Reis Ferreira

Kassyano Kevyn Andrade de Souza

Análise de Séries Temporais

Decomposição MSTL e Modelo ARIMA

Brasília, DF

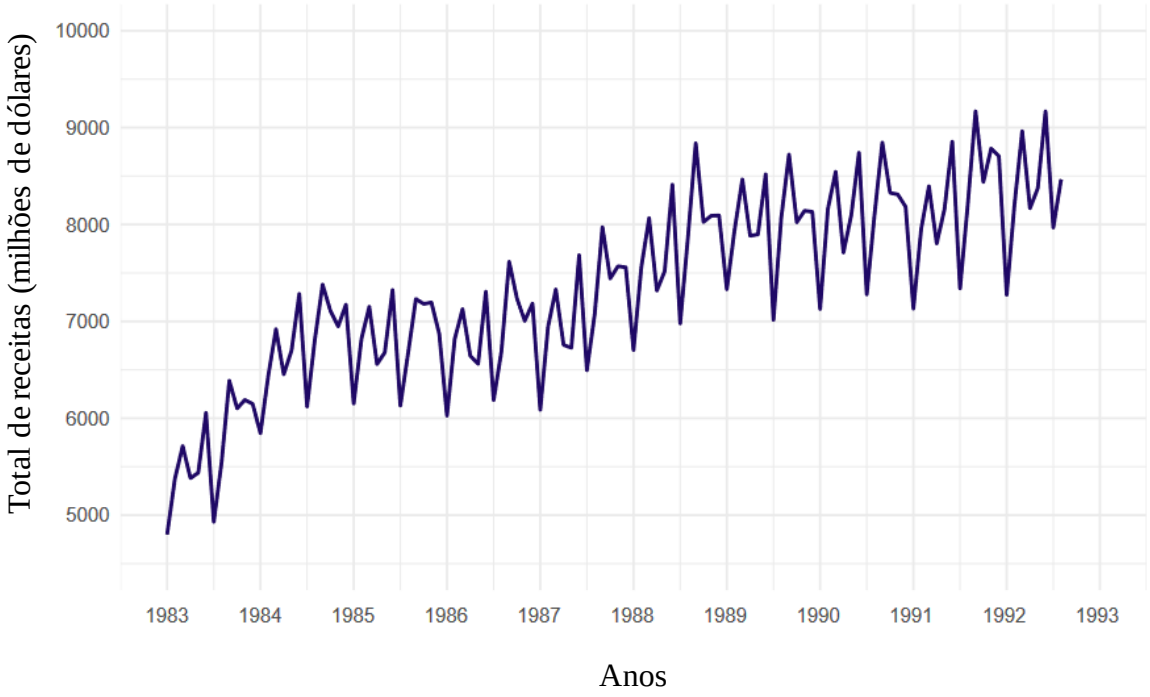
17 de maio de 2023

1. Considerações Iniciais

O presente trabalho tem como objetivo estudar uma série temporal selecionada do banco de dados da competição de previsão M3, disponibilizado no pacote *Mcomp* do R. Em um primeiro momento, será realizada uma decomposição da série temporal via MSTL e, posteriormente, um modelo ARIMA será selecionado para o estudo.

A série temporal escolhida foi a de número 2334. As 116 observações disponibilizadas se referem ao período entre janeiro de 1983 e agosto de 1992. Os valores da série indicam o total de receitas, em milhões de dólares, na venda de produtos desenvolvidos por empresas que fabricam e comercializam eletrônicos e equipamentos elétricos. Os produtos em questão são carregados e transportados em navios, além disso, são produzidos por meio da técnica de *manufacturing* (produção de bens por meio de maquinário, ferramentas, processos químicos e biológicos). A série temporal que será estudada pode ser visualizada a seguir:

Gráfico 1 – Série temporal do total de receitas, em milhões de dólares, na venda de produtos eletrônicos e equipamentos elétricos por empresas do setor - Estados Unidos, 1983-1992



Fonte: International Institute of Forecasters – M3 Competiton

2. Decomposição da série temporal via MSTL

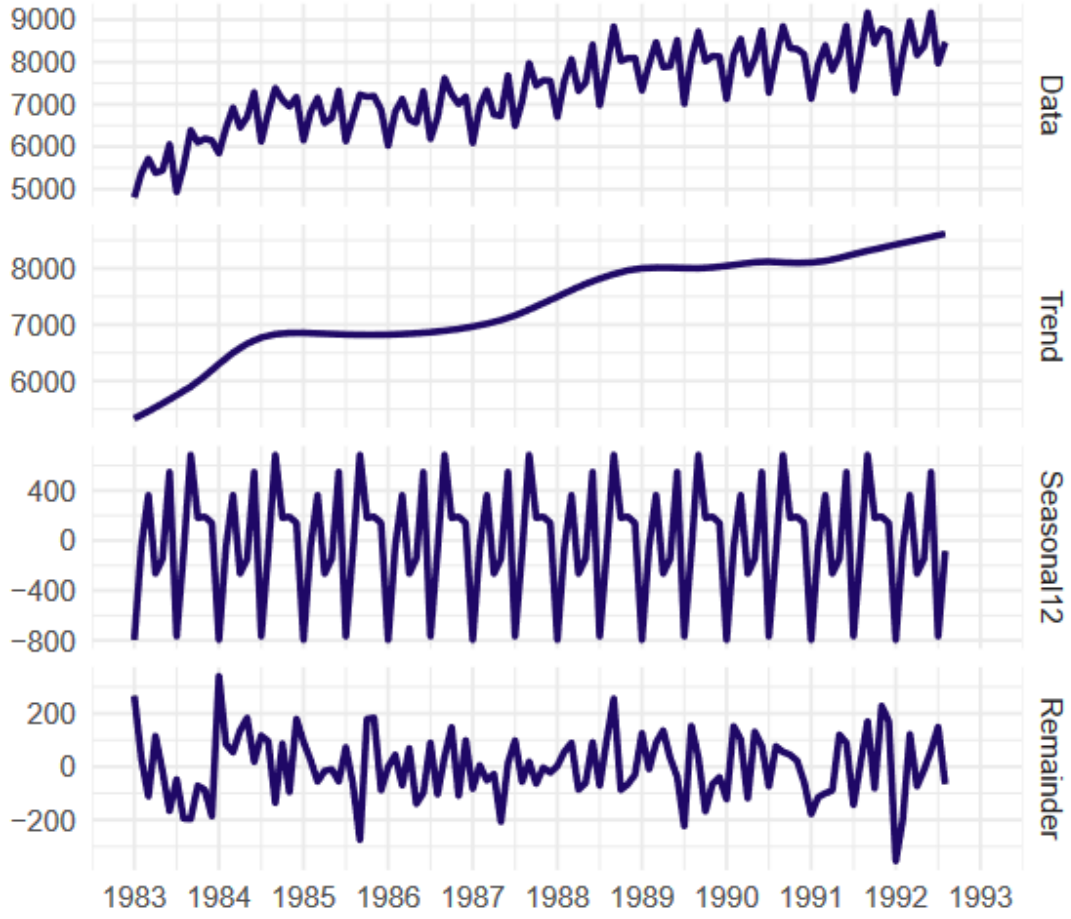
A decomposição MSTL nos permite dividir a estrutura da série temporal em três componentes: tendência, sazonalidade e resíduos. A tendência está atrelada a mudança no nível médio da série ao longo do tempo, ou seja, é possível observar comportamentos gerais como crescimento, constância ou decrescimento no desenvolvimento do fenômeno sob estudo. Por sua vez, a sazonalidade indica a repetição de padrões e ciclos que estão presentes em determinados recortes temporais da série como meses ou semanas, por exemplo. Por fim, a componente de resíduos incorpora a informação que resta na série após a extração das componentes de tendência e sazonalidade (Morettin e Toloi, 2018).

O termo MSTL se refere a Multiple Seasonal-Trend decomposition using LOESS. Nesse sentido, estamos trabalhando com uma técnica de decomposição que se baseia em um processo iterativo de regressão local. A decomposição MSTL é mais robusta do que outras técnicas, além disso, pode nos ajudar na identificação de múltipla sazonalidade. Portanto, apresentaremos a seguir a redução da série temporal selecionada em componentes mais simples de serem analisadas e que nos fornecerão informações para a modelagem dos dados.

Ao observarmos o gráfico 1 foi possível identificar que a série apresenta um desenho característico de séries com tendência de crescimento. Além disso, é perceptível a existência de um ciclo que se repete a cada 12 observações, ou seja, a série é mensal e possui um ciclo sazonal que volta a se repetir a cada ano. É interessante observar que essa sazonalidade parece ser mais ou menos constante, independente do nível global da série. Nessa perspectiva, os primeiros semestres de cada ciclo possuem um comportamento bem similar e o pouco de variação é mais nítido nos segundos semestres de cada ciclo.

Uma conclusão importante é que a série não é estacionária. Portanto, a série temporal em questão não se desenvolve ao redor de uma média constante e não reflete alguma forma de equilíbrio estável. Em suma, esse tipo de informação será essencial para a construção futura de um modelo como o ARIMA, capaz de descrever probabilisticamente uma série com tais características de tendência e sazonalidade. Vejamos abaixo o gráfico com o resultado da decomposição MSTL:

Gráfico 2 – Decomposição via MSTL da série original em componentes de tendência, sazonalidade e resíduos - Estados Unidos, 1983-1992



Fonte: International Institute of Forecasters – M3 Competiton

Conforme esperado, a tendência de crescimento foi confirmada. O total de vendas, em milhões de dólares, de produtos eletrônicos e equipamentos elétricos cresceu, de forma aproximadamente linear, ao longo dos quase 10 anos. De acordo com o Governo dos Estados Unidos (2023), esse setor foi aquecido nas últimas décadas a partir do desenvolvimento e melhoria da sua cadeia produtiva (capacidade de armazenamento e ritmo de produção dos produtos, desenvolvimento tecnológico, logística de transporte e entrega, e, mais recentemente, marketing digital e vendas na internet, por exemplo).

A componente sazonal também foi bem ajustada via MSTL. É perceptível que a cada ciclo existem três grandes picos mensais na geração de receitas, em milhões de dólares. O primeiro momento ocorre em março, o segundo em junho e o terceiro começa em setembro. Após o nono mês os valores ainda permanecem em um patamar elevado até caírem entre o fim de dezembro e o mês de janeiro, no qual a arrecadação do valor monetário apresenta o menor desempenho.

Por fim, obtemos a componente de resíduos. Podemos observar que os erros apresentam um padrão mais ou menos aleatório, logo, não aparentam estar correlacionados. Nesse sentido, os resíduos incorporaram muito pouca informação de tendência e não há muitos resquícios de sazonalidade, por exemplo. Além disso, os erros estão em torno da média zero, o que auxilia na construção de modelos e em previsões não viesadas.

No próximo tópico vamos encontrar um modelo para a série temporal que estamos estudando. A decomposição via MSTL foi essencial para realizarmos uma análise descritiva da série e, consequentemente, obtermos informações importantes para a modelagem a seguir. Em suma, temos as seguintes configurações:

- A série não é estacionária.
- A série possui tendência.
- Existe comportamento sazonal.

Observação: A função *stl* do R também foi utilizada. Foram testados alguns valores para as janelas de ajuste para sazonalidade e tendência, entretanto, o resultado mais satisfatório e adequado para a decomposição ocorre quando as funções *stl* e *mstl* convergem para os mesmos parâmetros. Portanto, a decomposição MSTL foi escolhida para discussão, embora tenhamos o mesmo resultado via STL.

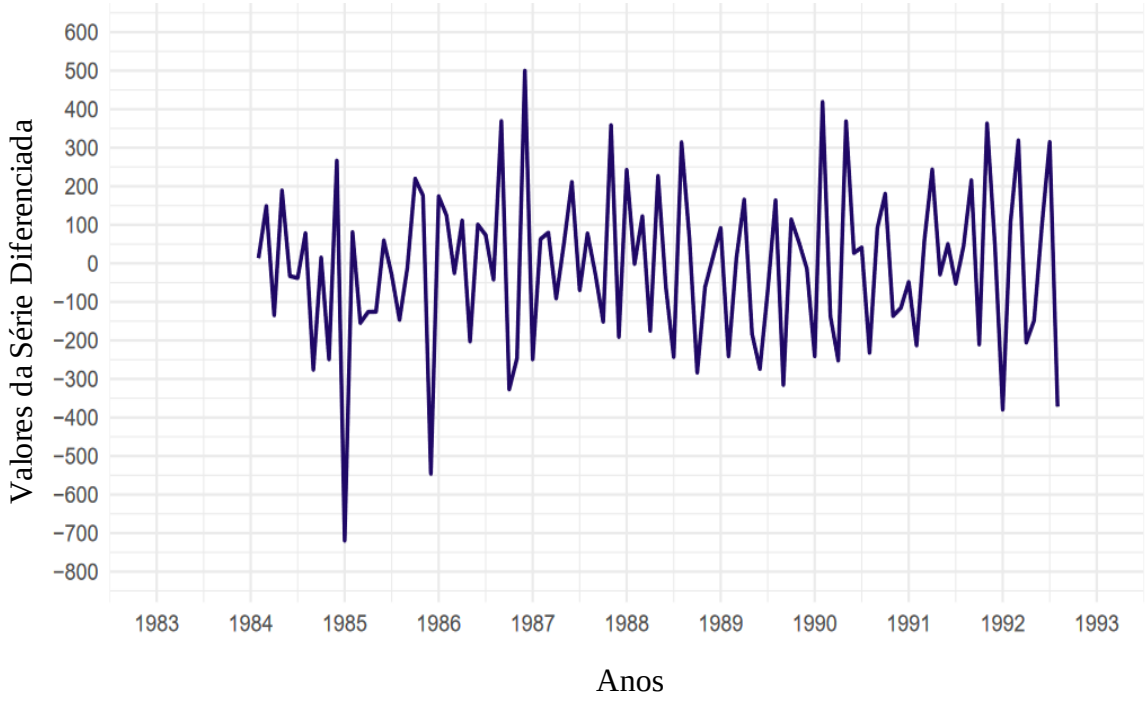
3. Escolha do modelo ARIMA

O primeiro passo para a escolha do modelo ARIMA consiste em identificar se a série é estacionária. Conforme análise acima, vimos que a série temporal sob estudo não apresenta tal característica, entretanto, precisamos de uma confirmação mais robusta. Fazendo o teste KPSS para concluir se a série é estacionária, obtem-se um p-valor = 0,01. De modo que, rejeita-se a hipótese nula de estacionaridade.

```
## Warning in kpss.test(serie_temp): p-value smaller than printed p-value
##
## KPSS Test for Level Stationarity
##
## data: serie_temp
## KPSS Level = 2.1897, Truncation lag parameter = 4, p-value = 0.01
```

De fato, podemos utilizar a função **ndiffs** e verificar que é necessário uma diferença para remover as raizes unitárias simples. E em seguida, com a utilização da função **nsdiffs**, também notamos que é suficiente uma diferença para remover as raízes sazonais (d=1, D=1). A série diferenciada pode ser observada abaixo:

Gráfico 3 – Diferenciação da série original após uma diferença simples e uma diferença sazonal - Estados Unidos, 1983-1992



Fonte: International Institute of Forecasters – M3 Competiton

Gráfico 4 – Gráfico ACF da série diferenciada - Estados Unidos, 1983-1992

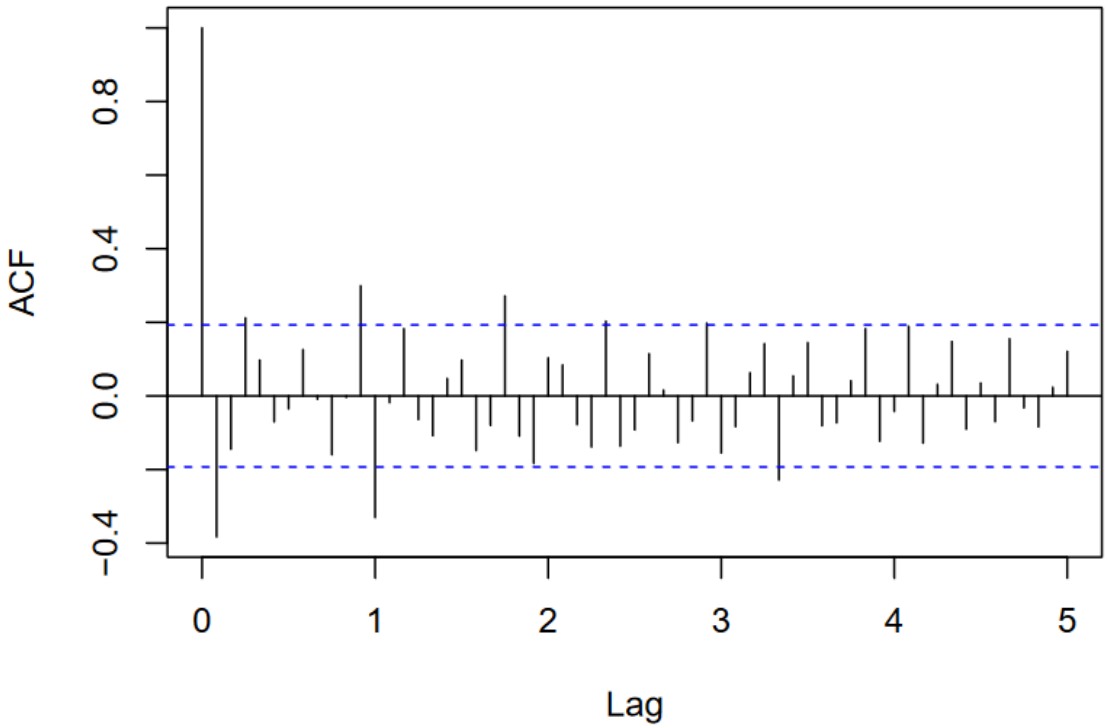
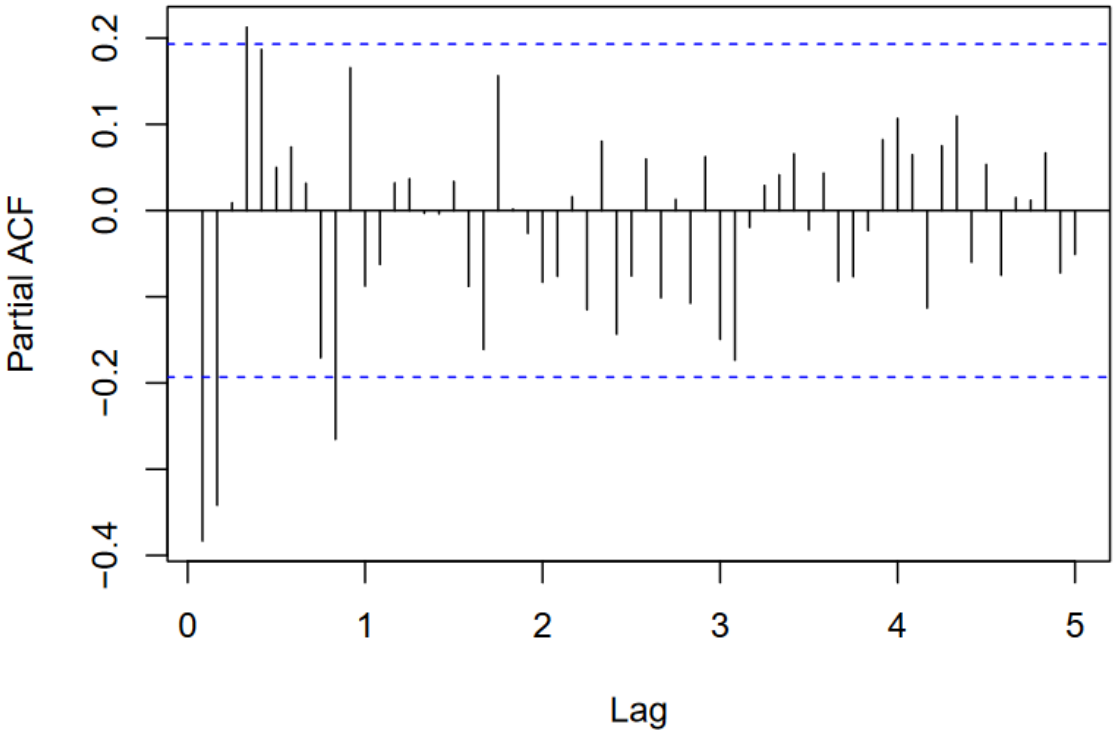


Gráfico 5 – Gráfico PACF da série diferenciada - Estados Unidos, 1983-1992



Percebe-se que, de fato, a série atingiu a estacionaridade após as diferenciações. Uma vez diferenciada, podemos tomar os gráficos ACF e PACF apresentados acima. A análise desses gráficos nos permitirá identificar possíveis candidatos a modelos para a série temporal, sobretudo, considerando que devemos verificar se precisaremos de parâmetros autoregressivos, de médias móveis, etc.

A partir do gráfico de autocorrelações ACF, visualizamos que a série não apresenta correlações que decaem para zero de forma amortizada, entretanto, também não é possível afirmar que existe uma quebra nos lags simples. No gráfico de autocorrelações parciais PACF também não há um padrão muito claro em relação aos lags simples, visto que o decaimento não é amortizado nem apresenta queda abrupta para o zero. Portanto, precisamos testar alguns valores para *p* e *q* do modelo ARIMA.

Considerando o gráfico de autocorrelações ACF, notamos que há uma correlação significativa no lag sazonal 1, seguido de uma quebra quando comparamos os demais níveis sazonais de 2 a 5. No gráfico PACF não é possível observar decaimento amortizado, o que caracterizaria diretamente um modelo com P = 0 e Q = 1, entretanto, também não existem quebras nos lags sazonais, visto que todos possuem correlações abaixo dos limites. Nessa perspectiva, o modelo com P = 0 e Q = 0 não faz muito sentido por causa da quebra no lag sazonal 1 no gráfico ACF. O modelo com P = 1 e Q = 1 é um candidato, entretanto, os decaimentos nos lags sazonais nos gráficos ACF e PACF não são amortizados, bem como poderíamos ter um modelo com mais parâmetros do que o necessário. Portanto, embora o decaimento no gráfico PACF não seja amortizado, o modelo com P = 0 e Q = 1 aparenta ser o candidato mais adequado para a série temporal.

Definido a forma do modelo, devemos encontrar agora as melhores combinações com os valores de p e q. Tomaremos os valores no intervalo {0,1,2,3}, e utilizaremos o critério de informação de Akaike para a seleção do melhor modelo.

```
##   p q Valor.de.AICc
## 1 0 0    1380.339
## 2 0 1    1362.526
## 3 0 3    1353.940
```

O melhor modelo, segundo o critério estabelecido, foi o modelo com os valores de $p = 0$ e $q = 3$. Dessa forma, o modelo escolhido será $\mathbf{SARIMA(0,1,3)_x(0,1,1)_{12}}$.

```
## Series: serie_temp
## ARIMA(0,1,3)(0,1,1)[12]
## Coefficients:
## ma1 ma2 ma3 sma1
## -0.5828 -0.0401 0.3869 -0.4537
## s.e. 0.1006 0.1035 0.1149 0.1524
##
## sigma^2 = 27149: log likelihood = -671.66
## AIC=1353.32 AICc=1353.94 BIC=1366.49
```

De modo geral, quando alguns modelos atendem todas as hipóteses é preferível escolher aquele com menos parâmetros. Além disso, deve-se também pensar em escolher o modelo com menor valor de critérios como o AIC, AICc e BIC. Esses critérios de parcimônia fazem um balanço entre a qualidade do ajuste e o grau de complexidade do modelo.

Portanto, a ideia geral do modelo selecionado é que não há dependência de valores atuais em relação aos anteriores, portanto, não se vê a necessidade de utilizar um termo autoregressivo. Além disso, se faz necessário o uso de três parâmetros de médias móveis e um parâmetro de médias móveis sazonal.

Observação: o modelo escolhido apresenta o melhor AIC corrigido e o menor valor para o BIC. Além disso, ele se destaca em relação a eventuais outros modelos por não apresentar mais parâmetros do que o necessário. Para ter certeza que a análise dos gráficos ACF e PACF estavam corretos, fizemos o teste para outros valores de P e Q entre 0 e 3, incluindo os modelos com P = 0 e Q = 0 e P = 1 e Q = 1. Considerando os critérios acima, o modelo $\mathbf{SARIMA(0,1,3)_x(0,1,1)_{12}}$ novamente foi considerado o mais adequado. Com a análise de resíduos será possível observar que o modelo ARIMA escolhido está bem ajustado. Embora não tenha sido requisitado, a previsão pela função *forecast* foi realizada e seu resultado nos indicou que as previsões, considerando tal modelo, seguem um padrão condizente com a série temporal.

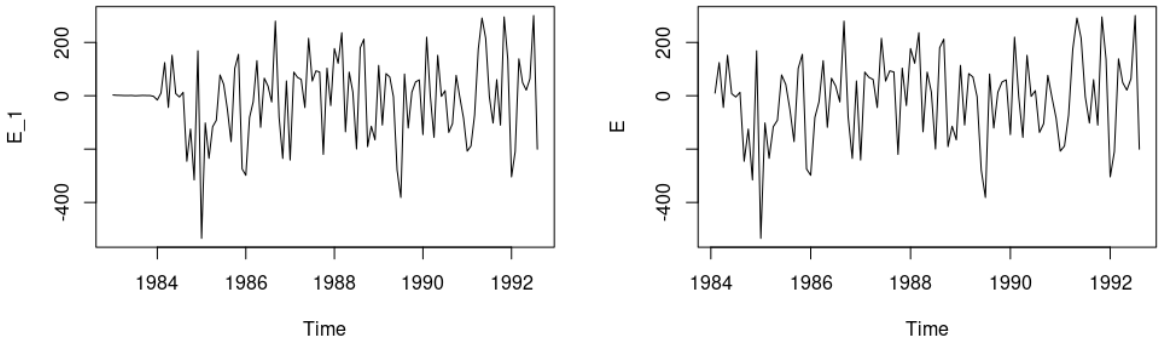
Em suma, podemos representar o algoritmo para a escolha do modelo da seguinte forma, conforme discutido em sala de aula (método de Box-Jenkins):

- Conferir a necessidade em remover raízes unitárias. Nesse sentido, devemos conferir quantas diferenças simples e sazonais são necessárias e aplicá-las.
- Identificar candidatos a modelos ARIMA, considerando a interpretação dos gráficos ACF e PACF.
- Seleção dos valores de p, q, P e Q, conforme os seguintes critérios: número de parâmetros, AICc e BIC.
- Análise de resíduos, assunto do próximo tópico.

4. Análise de Resíduos

Para finalizar o processo de escolha do modelo **SARIMA(0,1,3)$_x$(0,1,1)$_{12}$** realizamos uma análise de resíduos. O foco nessa etapa está em observar se o modelo está bem ajustado, considerando que as hipóteses levantadas sobre os erros $\{\varepsilon_t\}$ devem ser atendidas: média em torno do zero, variância constante, autocorrelação nula e normalidade. Além disso, também testamos a estacionariedade.

Gráficos 6 e 7 – Resíduos totais e ajustados do modelo **SARIMA(0,1,3)$_x$(0,1,1)$_{12}$** para a série 2334 - Estados Unidos, 1983-1992

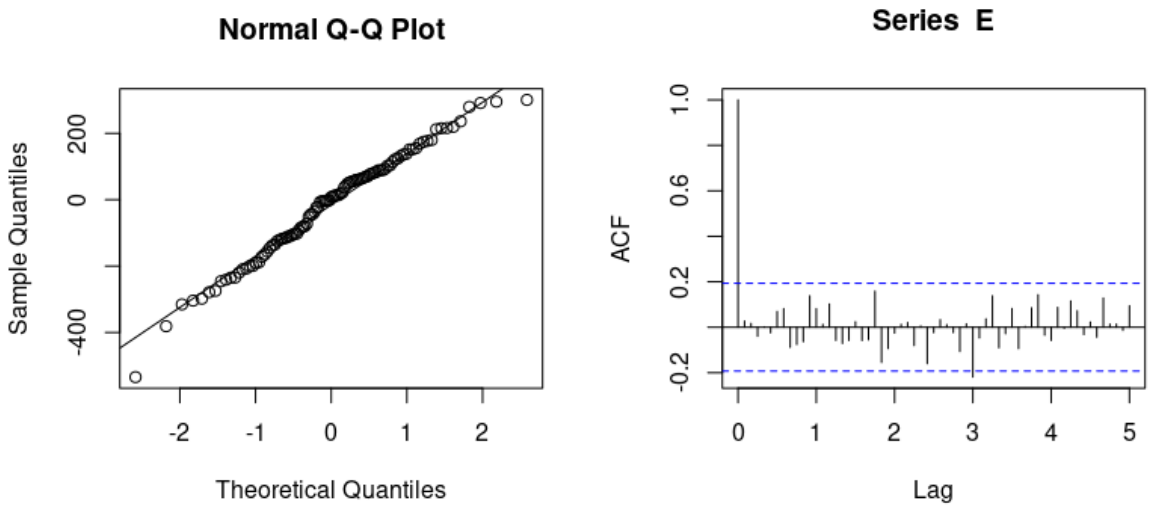


Fonte: International Institute of Forecasters – M3 Competiton

Os dados disponíveis não estão completos para anos anteriores a 1984, o que zerou os valores para todos eles durante o período.

Observando os valores onde os dados anuais estão completos, o comportamento explicitado pelo gráfico apresenta várias oscilações que, em média, são centradas em torno de zero. As extremidades, tanto superiores quanto inferiores, estão equidistantes em relação à média, e não se vê nenhum comportamento que aponte que a variância esteja correlacionada com o tempo. Ao invés disso, ela aparenta ser constante independente do período, comportamento semelhante ao observado em ruídos brancos.

Gráficos 8 e 9 – QQ Plot para teste de normalidade e ACF dos resíduos do modelo **SARIMA$(0,1,3)_x(0,1,1)_{12}$** - Estados Unidos, 1983-1992



Fonte: International Institute of Forecasters – M3 Competiton

A comparação dos quantis da distribuição normal com os pontos distribuídos dos erros mostra uma proximidade muito grande entre as duas distribuições. A reta coincide quase que perfeitamente com os resíduos aleatórios, com os pontos tendo maior concentração em torno de 0 e menor concentração nas extremidades.

No gráfico de autocorrelação, por sua vez, para as defasagens diferentes de zero (onde a autocorrelação é sempre 1), o comportamento geral é a autocorrelação contida dentro do intervalo. Não há indícios gráficos de autocorrelação entre os resíduos.

```
KPSS Test for Level Stationarity

data:  E KPSS Level = 0.2152, Truncation lag parameter = 4, p-value = 0.1
```

O p-valor obtido de 0,1 é maior do que o nível de significância de 0,05. Apesar de próximo, ainda não é suficiente para assumir diferenças significativas, de modo que não há evidências contra a hipótese de estacionariedade dos resíduos.

```
Shapiro-Wilk normality test

data:  E
W = 0.984, p-value = 0.2503
```

O p-valor de 0,2503 põe a estatística do teste fora dos intervalos de rejeição. Assim, não há evidências que atestem contra a normalidade dos resíduos, confirmando o que foi observado no gráfico de comparação dos resíduos com os quantis da distribuição normal.

```
Box-Ljung test

data:  E
X-squared = 8.4121, df = 15, p-value = 0.9062
```

```
Box-Ljung test

data:  E
X-squared = 10.378, df = 20, p-value = 0.9608
```

Os dois testes forneceram p-valores muito altos, de 0,9062 e 0,9608, para nível de defasagem de 15 e 20, respectivamente. Nos dois casos, a hipótese de independência permanece forte, de modo que não há motivos para assumir qualquer dependência entre os resíduos.

Em suma, a normalidade, a independência, a estacionariedade e a inexistência de autocorrelação estão todas satisfeitas na análise dos resíduos.

5. Equação do Modelo

Por fim, podemos escrever a equação do modelo **SARIMA$(0,1,3)_x(0,1,1)_{12}$**. Nesse sentido, vamos utilizar os parâmetros estimados e os operadores de retardo. Portanto, a equação encontrada possui a seguinte configuração:

$$(1 - B^{12})(1 - B)\, X_t = (1 - 0{,}454B^{12})(1 - 0{,}583B - 0{,}040B^2 + 0{,}387B^3)\, \varepsilon_t$$

em que $\{\varepsilon_t\}$ é um processo i.i.d. com distribuição Normal com média igual a 0 e desvio-padrão de 164,77. Para a inicialização do modelo precisamos considerar que o termo passado mais distante será o ε_{t-15}, logo, precisamos de $t > 15$. Dessa forma, temos que t = 16, 17, ...

6. Referências Bibliográficas

Estados Unidos da América, Census Bureau. Full Report on Manufacturers' Shipments, Inventories, and Orders. 2020.
Disponível em: https://www.census.gov/manufacturing/m3/data/index.html

Morettin, P. A e Toloi, C. M., Análise de Séries Temporais, 3ª edição, Projeto Fisher, 2018.