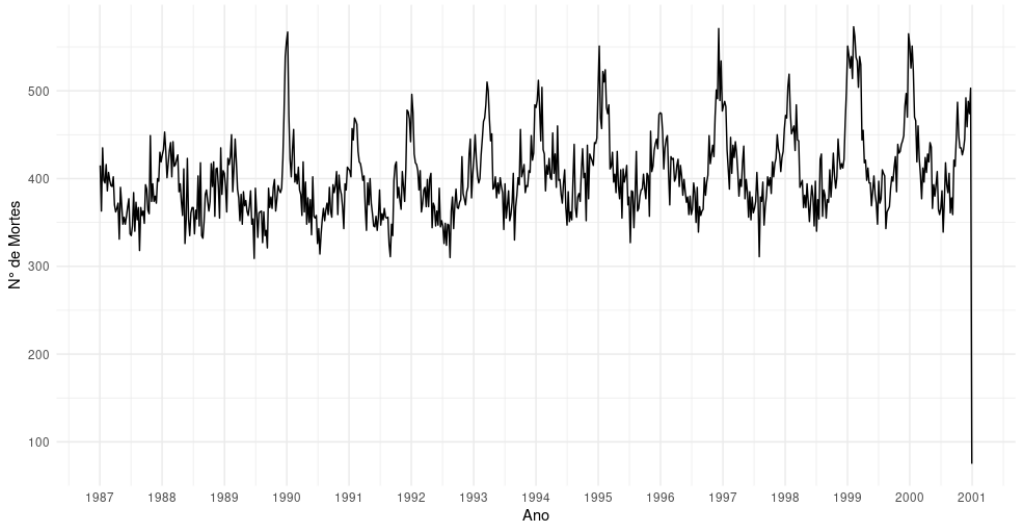
MI411 SÉRIES TEMPORAIS
Prova 2 / Questao 5

Adriel Wesley Nascimento Melo
RA: 258000

Campinas, 2024

O objetivo deste estudo é modelar uma série temporal de mortalidade de pessoas com mais de 75 anos na cidade de Chicago, EUA, utilizando dados semanais. A série abrange 731 observações, que representam o total de mortes em cada semana, desde 1987 até 2000. A visualização dessa série pode ser observada na Figura 1:

Figura 1: Série temporal de mortalidade semanal de pessoas com mais de 75 anos em Chicago (1987-2000)



Fonte: Autor

A Figura 1 revela uma leve tendência ao longo dos anos e uma notável sazonalidade anual. Observa-se um padrão de diminuição das mortes na metade do ano e um aumento no final e início do ano. Para identificar outliers, foram calculados a média e o desvio padrão da série, considerando como outliers aqueles que estão a mais de 2 desvios padrão da média.

Na análise da série temporal de mortalidade, foram identificados outliers utilizando o método do desvio padrão. Os valores considerados outliers são aqueles que se encontram a mais de três desvios padrão da média da série. A tabela abaixo apresenta esses valores. A média obtida foi de 402,651 e o desvio padrão foi de 48,575. Assim, os valores discrepantes podem ser observados na Tabela 1.

Tabela 1: Outliers identificados na série temporal de mortalidade

Nº da Observação	Semana	Nº de Mortes
1	157	556
2	158	567
3	419	551
4	519	571
5	627	551
6	632	573
7	633	562
8	678	565
9	679	552
10	681	551
11	731	75

Fonte: Autor

Esses valores extremos podem influenciar a análise e a modelagem da série temporal, levando a interpretações errôneas dos resultados. Para lidar com essa questão, será empregado o logaritmo natural (log) para estabilizar esses valores e reduzir a variabilidade. A aplicação do logaritmo pode ajudar a normalizar a distribuição dos dados e atenuar o efeito dos outliers, facilitando uma modelagem mais robusta e confiável da série temporal.

Para a modelagem da série, foram descartadas as 13 últimas observações, que serão utilizadas como dados de teste; assim, a série passa a ter um total de $n = 718$ observações.

Dado que a série temporal apresenta um ciclo anual, o modelo de regressão harmônica foi escolhido. A frequência do ciclo é de 52 semanas, e a fórmula do modelo de regressão harmônica é:

$$y_t = T_t + \sum_{j=1}^{J} \left(a_j(2\pi\nu_j t) + \beta_j \cos(2\pi\nu_j t)\right) + \epsilon_t,$$

onde T_t é a semana, ou seja, o tempo, $P = 52$ semanas e $J = \frac{52}{2} = 26$. Assim, temos um total de 26 pares de $sen()$ e $cos()$. Foram realizadas três tentativas de modelagem, para $J = 26$, $J = 5$ e $J = 4$. Os respectivos R^2 foram de $0,608$, $0,600$ e $0,591$. Para o modelo com $J = 26$, nem todos os coeficientes foram significativos; a partir da ordem 6, os coeficientes deixaram de ser significativos. Já para os modelos com $J = 4$ e $J = 5$, os únicos coeficientes que não foram significativos foram os do $cos()$ de ordens 3 e 4. Como o modelo $J = 5$ foi o que obteve o maior R^2, foi o modelo selecionado.

Os coeficientes do modelo ajustado são apresentados na Tabela 2.

Tabela 2: Estimativas dos parâmetros do modelo selecionado.

Coeficientes	Estimativas	Estatística t	P-valor
Intercepto	375.2	2.237	< 2e-16 ***
T_t	0.0755	0.0054	< 2e-16 ***
$\sin(2\pi\frac{1}{52}t)$	38.89	1.584	< 2e-16 ***
$\cos(2\pi\frac{1}{52}t)$	22.88	1.576	< 2e-16 ***
$\sin(2\pi\frac{2}{52}t)$	4.265	1.577	0.007008 **
$\cos(2\pi\frac{2}{52}t)$	7.159	1.581	6.98e-06 ***
$\sin(2\pi\frac{3}{52}t)$	0.5318	1.580	0.336
$\cos(2\pi\frac{3}{52}t)$	5.491	1.578	0.000531 ***
$\sin(2\pi\frac{4}{52}t)$	1.143	1.580	0.724
$\cos(2\pi\frac{4}{52}t)$	4.686	1.578	0.003085 **
$\sin(2\pi\frac{5}{52}t)$	4.003	1.578	0.011424 *
$\cos(2\pi\frac{5}{52}t)$	3.203	1.579	0.042939 *

Fonte: Autor

Análise dos Coeficientes:

- **Intercepto**: O coeficiente é 375,3 com uma estatística t de 2,241 e um p-valor inferior a 2 × 1016, indicando ser estatisticamente significativo e sugere um valor base considerável para a variável dependente.

- T_t: O coeficiente é 0,075 com uma estatística t de 0,005 e um p-valor inferior a 2 × 1016. Isso indica uma tendência linear significativa ao longo do tempo. Dada a informação de que o coeficiente é positivo, deve-se discordar do pesquisador que afirma que a mortalidade está diminuindo. Em vez disso, a análise sugere que a mortalidade, de fato, está aumentando ao longo do tempo, com um coeficiente positivo que indica essa tendência.

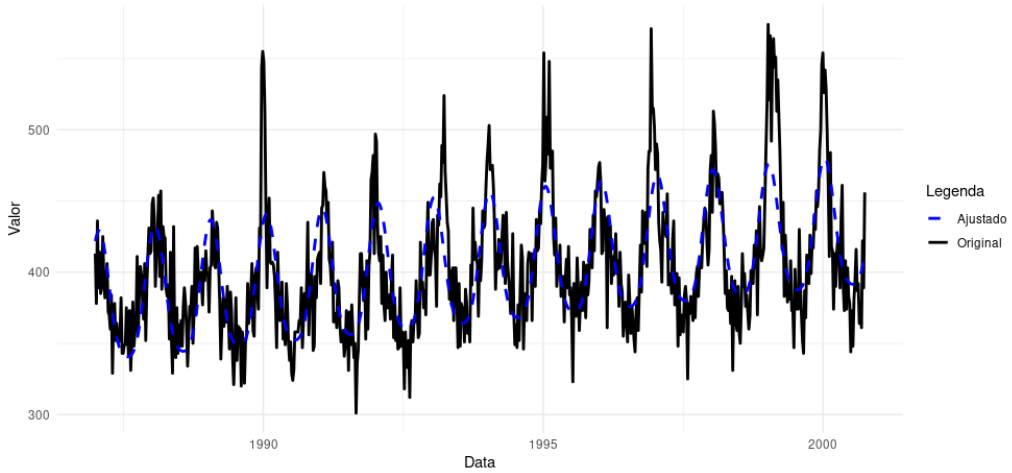
- **Componentes Sazonais**:

  - $\sin\left(2\pi\frac{1}{52}t\right)$: O coeficiente é 38,20 com uma estatística t de 1,586 e um p-valor muito baixo, confirmando a presença significativa desta componente sazonal.
  - $\cos\left(2\pi\frac{1}{52}t\right)$: O coeficiente é 24,20 com uma estatística t de 1,579 e um p-valor também muito baixo, indicando que esta componente cossenoidal semanal é significativa.
  - $\sin\left(2\pi\frac{2}{52}t\right)$: O coeficiente é 4,03 com uma estatística t de 1,580 e um p-valor de 0,0109, sugerindo significância, mas com menor intensidade comparada às primeiras ordens.

- $\cos\left(2\pi\frac{2}{52}t\right)$: O coeficiente é 7,14 com uma estatística t de 1,584 e um p-valor muito baixo, confirmando a importância desta componente cossenoidal de segunda ordem.
- $\sin\left(2\pi\frac{3}{52}t\right)$: O coeficiente é 0,5318 com uma estatística t de 0,336 e um p-valor de 0,7366, indicando que esta componente não é significativa.
- $\cos\left(2\pi\frac{3}{52}t\right)$: O coeficiente é 5,491 com uma estatística t de 1,578 e um p-valor de 0,000531, sugerindo que esta componente é significativa.
- $\sin\left(2\pi\frac{4}{52}t\right)$: O coeficiente é 1,143 com uma estatística t de 0,724 e um p-valor de 0,4694, o que indica que esta componente não é significativa.
- $\cos\left(2\pi\frac{4}{52}t\right)$: O coeficiente é 4,686 com uma estatística t de 1,578 e um p-valor de 0,003085, confirmando a significância desta componente cossenoidal.
- $\sin\left(2\pi\frac{5}{52}t\right)$: O coeficiente é 4,003 com uma estatística t de 1,578 e um p-valor de 0,011424, sugerindo que esta componente é significativa.
- $\cos\left(2\pi\frac{5}{52}t\right)$: O coeficiente é 3,203 com uma estatística t de 1,579 e um p-valor de 0,042939, indicando a importância desta componente cossenoidal.

Na Figura 3 pode-se observar os valores reais e os ajustados para série de mortalidade. Observa-se que o modelo consegue acompanhar o movimento da tendência e sazonalidade, mas não consegue capturar bem os picos.

Figura 2: Figura 3: Valores ajustados



Fonte: Autor

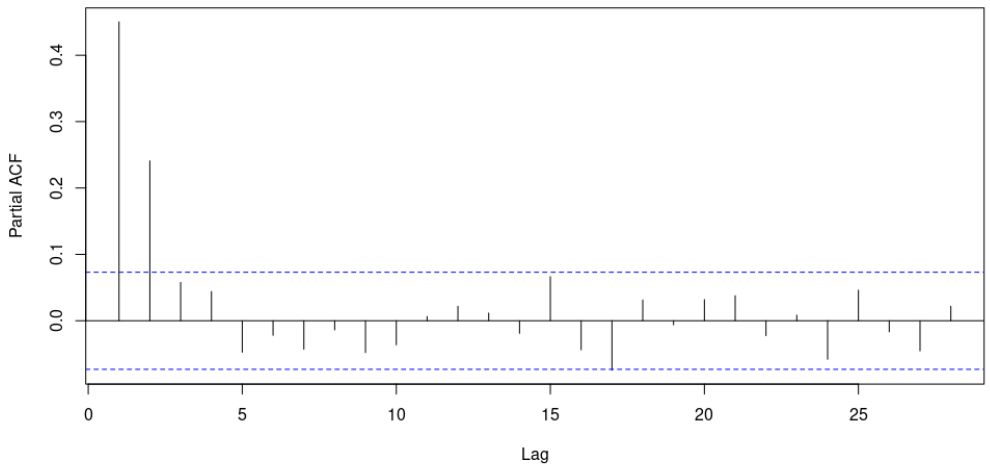
Na Figura 3, pode-se observar o qq-plot para os resíduos do modelo, pode-se observar que os pontos nas extremidades não estão sob a reta, isso pode indicar que

os resíduos não seguem uma distribuição normal. Verificando a Figura 4, parece haver ainda estrutura de dependência.

O teste de Ljung-Box para a independência dos resíduos com defasagem de ordem 20 obteve um p-valor de 2,2e-16, que indica haver uma forte evidência de autocorrelação nos resíduos do modelo. Nesse caso, temos um modelo de regressão com erros de séries temporais, e as estimativas MQ podem não ser consistentes.

O modelo é inadequado, como mostrado na Figura 3, que apresenta o gráfico da PACF (Função de Autocorrelação Parcial) dos resíduos. Em particular, a PACF amostral dos resíduos é altamente significativa até a defasagem de ordem 2, mostrando que os resíduos são correlacionados e se faz necessário tratar essa dependência.

Figura 3: Figura 3: Série residual de regressão linear



Fonte: Autor

Pela Figura 3, essa fraca dependência serial nos resíduos pode ser modelada usando um AR(p). A partir da PACF amostral dos resíduos mostrada na Figura 3, especifica-se um modelo AR(2) para os resíduos e modifica-se o modelo de regressão linear para:

$$y_t = T_t + \sum_{j=1}^{J} \left(a_j(2\pi\nu_j t) + \beta_j \cos(2\pi\nu_j t)\right) + \epsilon_t \ , \ \epsilon_t = \phi_1 a_{t-1} + \phi_2 a_{t-2} + a_t,$$

onde a_t é considerado uma série de ruído branco. Em outras palavras, simplesmente usamos um modelo AR(2), sem o termo constante, para capturar a dependência serial no termo de erro do modelo de regressão. O modelo resultante é um exemplo simples de regressão linear com erros de séries temporais. Para fim de exploração, foram considerados três modelos autorregressivos para os resíduos, AR(1), AR(2) e AR(3).

A seleção dos modelos foi baseada no Critério de Informação de Akaike (AIC), Schwarz, Hanna - Quinn e na análise de diagnóstico dos resíduos. Com base nesse critério, concluiu-se que o modelo AR(2) é o que mais se destaca na análise de diagnóstico.

Tabela 3: Comparação de Critérios de Informação para Modelos AR(1), AR(2) e AR(2)

Critério	AR(1)	AR(2)	AR(2)
AIC	6996,253	6862,042	**6852,00**
BIC	7009,982	6880,348	**6852,00**
Hanna-Quinn	6997,253	**6853,301**	6857,068

Fonte: Autor

Pela Figura 4 observam-se os modelos propostos não conseguiram capturar toda dependência, em que, várias autocorrelações estatísticas significativas na defasagem de ordem 4 e a partir da ordem 17. O modelo escolhido para capturar a dependência temporal nessa análise foi um AR(3), pois, para capturar toda essa correlação, seria necessário um modelo AR(17), porém, seria um modelo com mais coeficientes e com um processamento mais trabalhoso.

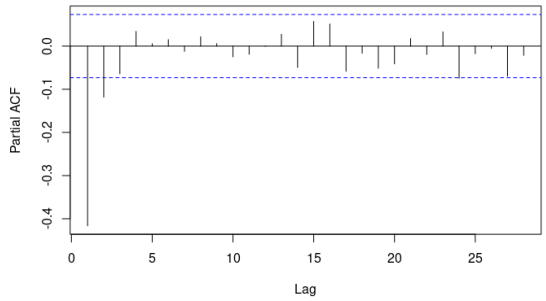
Logo o modelo escolhido para modelar os resíduos foi um AR(3), que possui os seguintes coeficientes:

$$\epsilon_t = 0.524_1 a_{t-1} + 0,343_2 a_{t-2} + 0.1295327 a_{t-3} + a_t,\ a_t \sim RB(0, 800, 01).$$

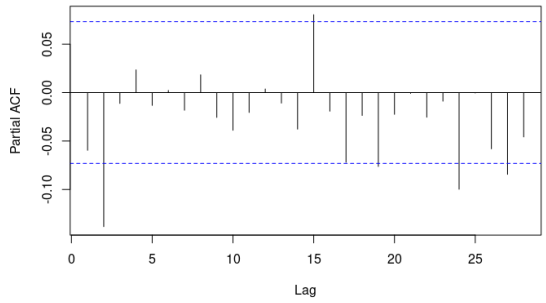
Mas como a série original foi aplicada o log, é necessário aplicar a exponencial para voltar a natureza dos dados, que é dada da seguinte forma

$$\exp y_t = \exp\{T_t + \textstyle\sum_{j=1}^{J} (a_j(2\pi\nu_j t) + \beta_j \cos(2\pi\nu_j t)) + \epsilon_t\}.$$

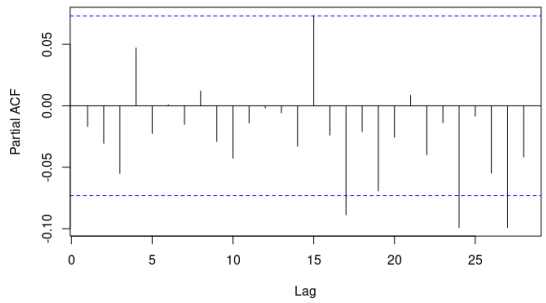
Figura 4: Figura 4: PACF dos resíduos dos modelos AR(1),AR(2) e AR(3)



((a)) PACF AR(1)



((b)) PACF AR(2))



((c)) PACF AR(3)

Figura 5: Fonte: Autor