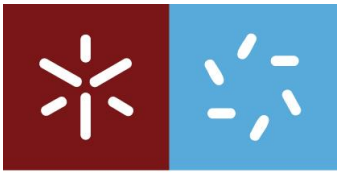

Universidade do Minho
Escola de Ciências

Modelação e Previsão da Série USAccDeaths: Mortes Acidentais Mensais nos EUA (1973–1978)

Mestrado em Estatística para a Ciência de Dados
Métodos de Previsão e Séries Temporais

Ano Letivo 2024/2025
Rui Miguel Pereira Alves PG55577

Junho de 2025

Conteúdo

1 Introdução

1.1 Problema e Objetivo do Estudo

O estudo de séries temporais é essencial para compreender fenómenos que evoluem ao longo do tempo e para realizar previsões com base em dados históricos. Um exemplo concreto e realista é o número mensal de mortes acidentais nos Estados Unidos da América, cuja evolução pode refletir padrões sazonais, tendências e variações aleatórias. A análise estatística destes dados permite não só descrever o comportamento passado, como também construir modelos que ajudem a antecipar a evolução futura.

O objetivo principal deste trabalho é aplicar a metodologia de Box-Jenkins à série temporal `USAccDeaths`, com vista à sua modelação e previsão. Serão consideradas transformações para estabilização da variância, avaliadas condições de estacionariedade, identificadas componentes sazonais, e ajustados modelos ARIMA sazonais. No final, pretende-se obter um modelo ajustado que permita realizar previsões para os 12 meses seguintes, avaliando a sua qualidade com base em métricas estatísticas apropriadas.

1.2 Descrição da Série Temporal

A série temporal `USAccDeaths` corresponde ao número de mortes acidentais registadas mensalmente nos Estados Unidos, no período compreendido entre janeiro de 1973 e dezembro de 1978, totalizando 72 observações. Trata-se de uma série com frequência mensal (`frequency = 12`), e que apresenta um padrão sazonal marcado, com valores mais elevados em determinados meses do ano. A série está disponível de forma nativa no software `R`, sendo frequentemente utilizada para fins pedagógicos em análise de séries temporais, tal como será utilizado no decorrer deste trabalho.

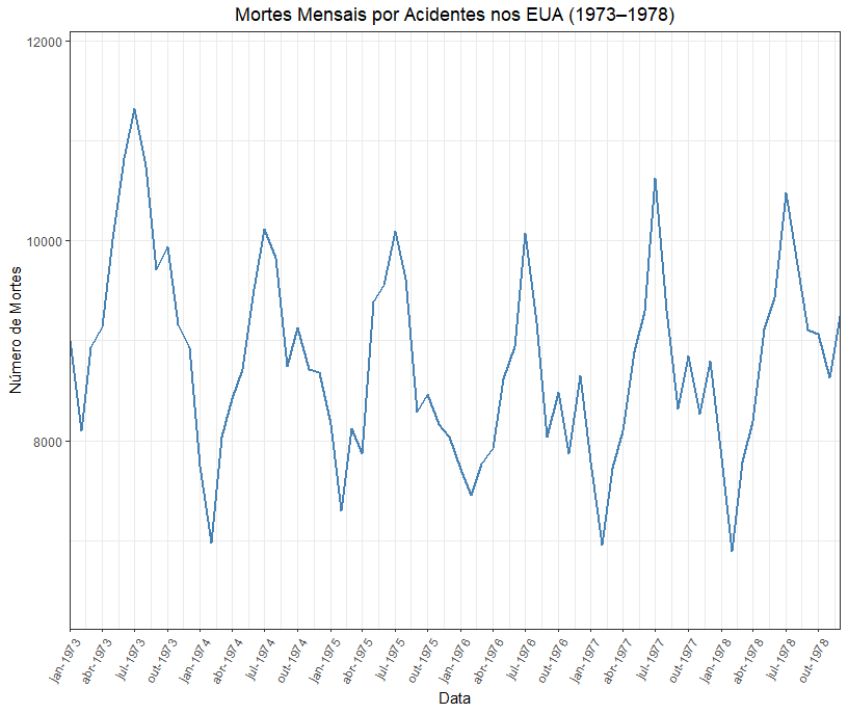

Figura 1.1: Evolução mensal do número de mortes acidentais nos EUA (1973–1978) — série `USAccDeaths`.

2 Análise Exploratória

2.1 Análise Descritiva e Gráfica

Antes da modelação, é importante compreender o comportamento geral da série. A Tabela 2.1 apresenta algumas medidas estatísticas descritivas da série `USAccDeaths`, incluindo o valor mínimo, máximo, média, mediana, desvio padrão e coeficiente de variação. Estes indicadores ajudam a caracterizar a variação global da série.

Tabela 2.1: Medidas estatísticas descritivas da série `USAccDeaths`.

Medida	Mínimo	Máximo	Média	Mediana	Desvio padrão	Coef. variação
Valor	6892	11317	8788.79	8728	957.75	0.11

Adicionalmente, foi construído um histograma (Figura 2.1) para visualizar a distribuição das observações. Observa-se uma assimetria ligeira à direita, com a maioria dos valores concentrados entre 8000 e 10000 mortes mensais. Este padrão sugere que os valores mais frequentes correspondem a meses de mortalidade moderada, enquanto os meses com valores muito elevados (acima de 10500) são menos comuns e podem estar associados a eventos sazonais ou pontuais. A distribuição não é perfeitamente simétrica, o que poderá ter implicações na análise dos resíduos e na escolha de modelos com suposições de normalidade.

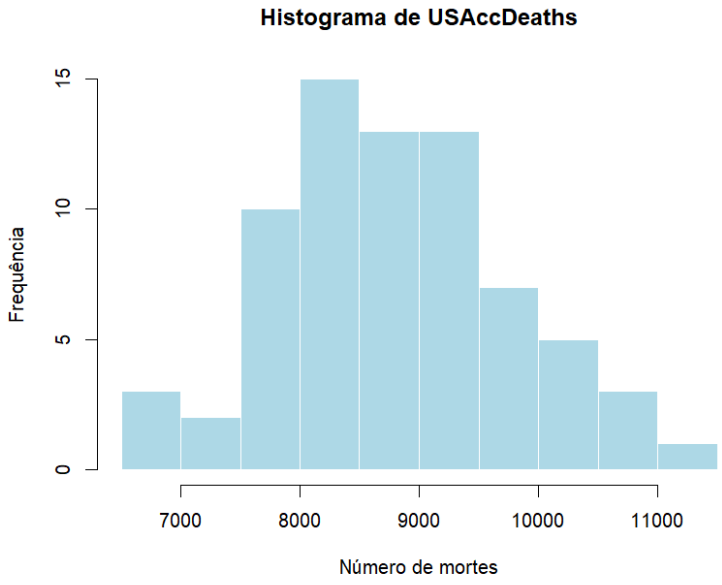

Figura 2.1: Histograma da série `USAccDeaths`.

Para complementar a análise descritiva da série, foram construídos três boxplots com diferentes objetivos.

O primeiro (Figura 2.2) apresenta a distribuição global da série `USAccDeaths`. Este gráfico evidencia a presença de um outlier com valor bastante elevado, identificado como correspondente ao mês de julho de 1973, que registou um número anormalmente alto de mortes acidentais. Este valor poderá influenciar a estimação de alguns parâmetros nos modelos ajustados e será considerado na análise residual.

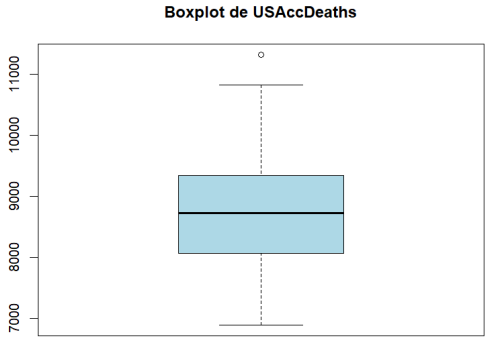

Figura 2.2: Boxplot global da série `USAccDeaths`. Observa-se um outlier em julho de 1973.

O segundo boxplot (Figura 2.3) agrupa os dados por mês do ano, permitindo observar a distribuição das mortes acidentais ao longo do ciclo anual. Verifica-se que os meses de verão, em particular julho (mês 7) e agosto (mês 8), apresentam medianas mais elevadas e maior dispersão, o que reforça a presença de uma componente sazonal. Já os meses de inverno (como fevereiro e março) concentram os valores mais baixos.

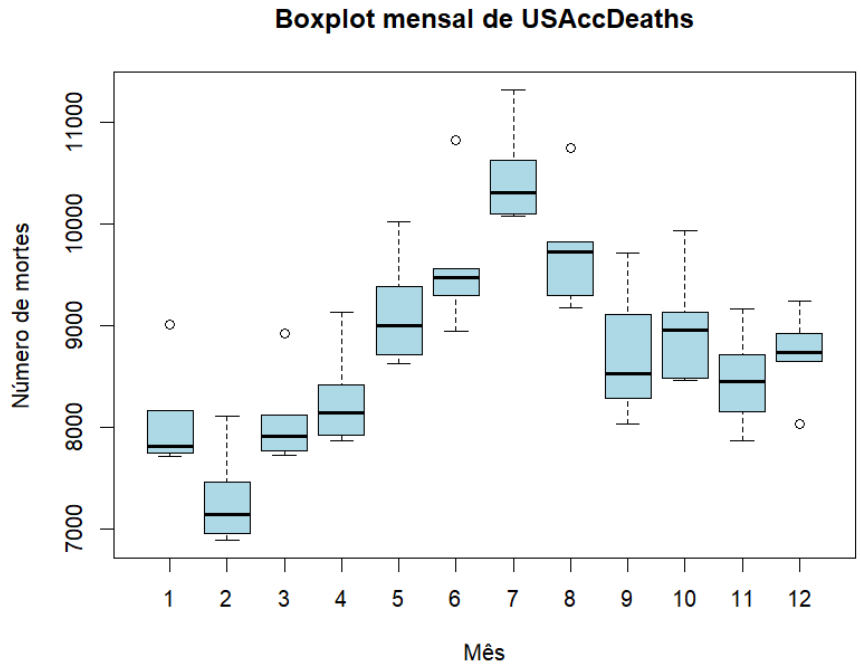

Figura 2.3: Boxplot mensal da série `USAccDeaths`. Evidencia-se um padrão sazonal claro, com picos nos meses de verão.

Por fim, a Figura 2.4 apresenta um boxplot agrupado por ano (1973 a 1978). Apesar da variabilidade interanual, não se observa uma tendência clara crescente ou decrescente. No entanto, o ano de 1973 destaca-se com valores geralmente mais elevados e maior dispersão, compatível com o outlier anteriormente identificado.

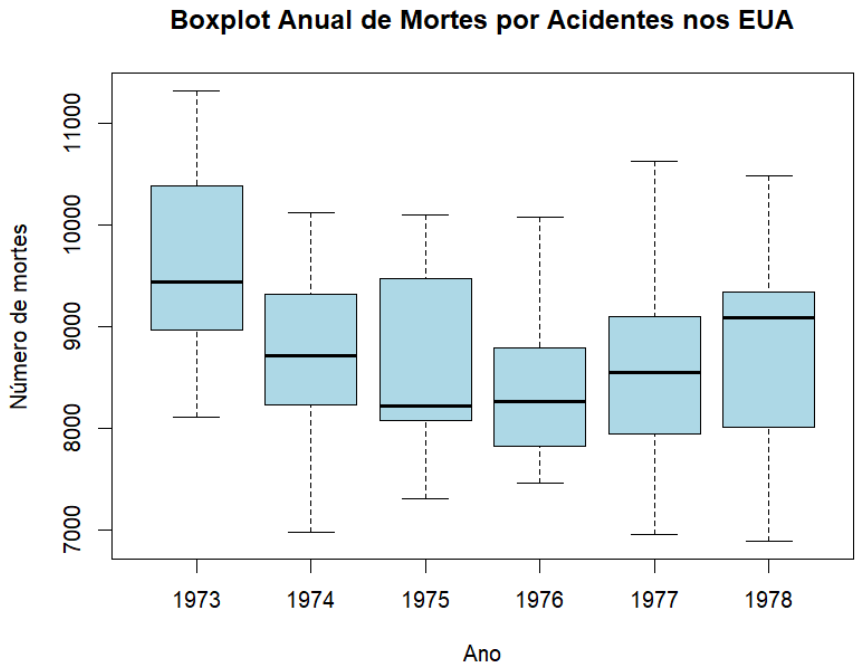

Figura 2.4: Boxplot anual da série `USAccDeaths`. O ano de 1973 destaca-se por ter maior variabilidade.

2.2 Decomposição da Série Temporal

A decomposição de uma série temporal é uma técnica fundamental para identificar e isolar os componentes estruturais da série: tendência, sazonalidade e componente aleatória (ou irregular). No caso da série `USAccDeaths`, foi utilizada a decomposição aditiva, que assume que a série observada pode ser expressa como:

$$Y_t = T_t + S_t + E_t$$

onde Y_t representa o valor observado no tempo t, T_t é o componente de tendência, S_t o componente sazonal, e E_t o componente aleatório (residual).

A Figura 2.5 apresenta os resultados da decomposição aditiva aplicada à série. No painel superior observa-se a série original, que exibe uma flutuação regular ao longo do ano, com picos recorrentes nos meses de verão e valores mais baixos no inverno, refletindo uma forte componente sazonal.

O segundo painel mostra a tendência estimada (T_t), a qual revela um comportamento ligeiramente decrescente no início da série, seguido de um período de estabilização entre 1975 e 1977, e um ligeiro aumento no final do período. Este padrão sugere que, embora exista uma componente de tendência, esta não é dominante.

A componente sazonal (S_t), apresentada no terceiro painel, confirma a existência de um padrão anual bem definido: todos os anos ocorrem picos nos mesmos períodos (particularmente no verão), o que reforça a necessidade de incorporar sazonalidade explícita na modelação.

Por fim, o componente aleatório (E_t), no painel inferior, representa a variação residual após remoção da tendência e da sazonalidade. A sua amplitude é relativamente pequena e não apresenta padrões sistemáticos, o que sugere que os principais padrões da série foram devidamente capturados pelos dois componentes anteriores.

Esta análise reforça a pertinência de considerar modelos sazonais aditivos (como os modelos SARIMA), dado que a estrutura da série contém uma sazonalidade consistente e uma tendência

fraca mas não negligenciável.

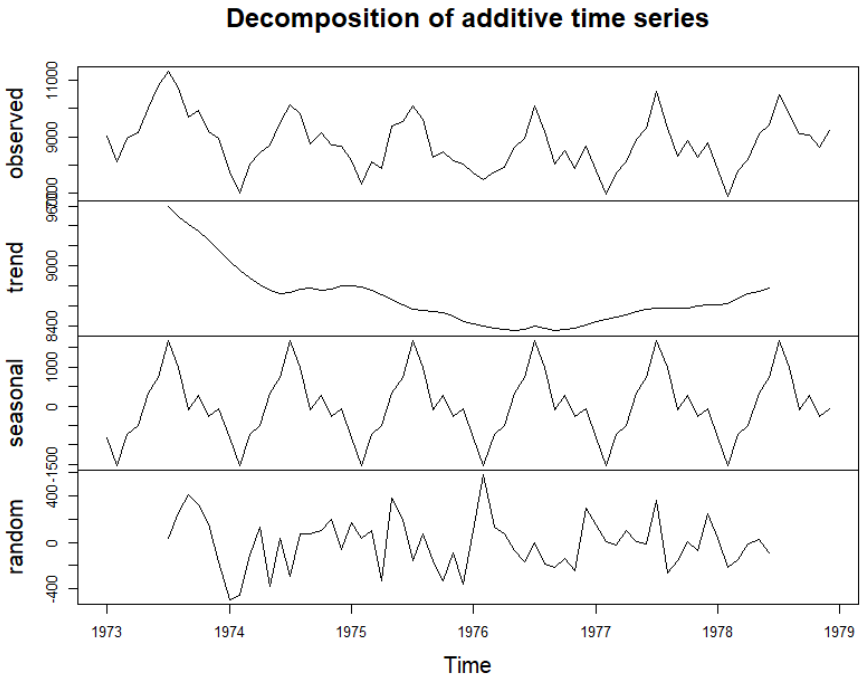

Figura 2.5: Decomposição aditiva da série `USAccDeaths`.

3 Estabilização da Variância

3.1 Transformação de Box-Cox

Antes de prosseguir com a modelação da série temporal, é necessário avaliar se a série apresenta variância constante ao longo do tempo. Muitas vezes, em séries económicas ou demográficas, observa-se um aumento da variabilidade à medida que a média aumenta. Para contornar esse problema, recorre-se frequentemente à transformação de Box-Cox, que permite estabilizar a variância e aproximar a série de uma distribuição normal.

A transformação de Box-Cox é definida por:

$$Y^{(\lambda)} = \begin{cases} \frac{Y^{\lambda}-1}{\lambda}, & \text{se } \lambda \neq 0 \\ \log(Y), & \text{se } \lambda = 0 \end{cases}$$

No caso da série `USAccDeaths`, o valor de λ foi estimado através da maximização da verosimilhança logarítmica, tendo-se obtido $\lambda = -0.65$. A série transformada encontra-se representada na Figura 3.1.

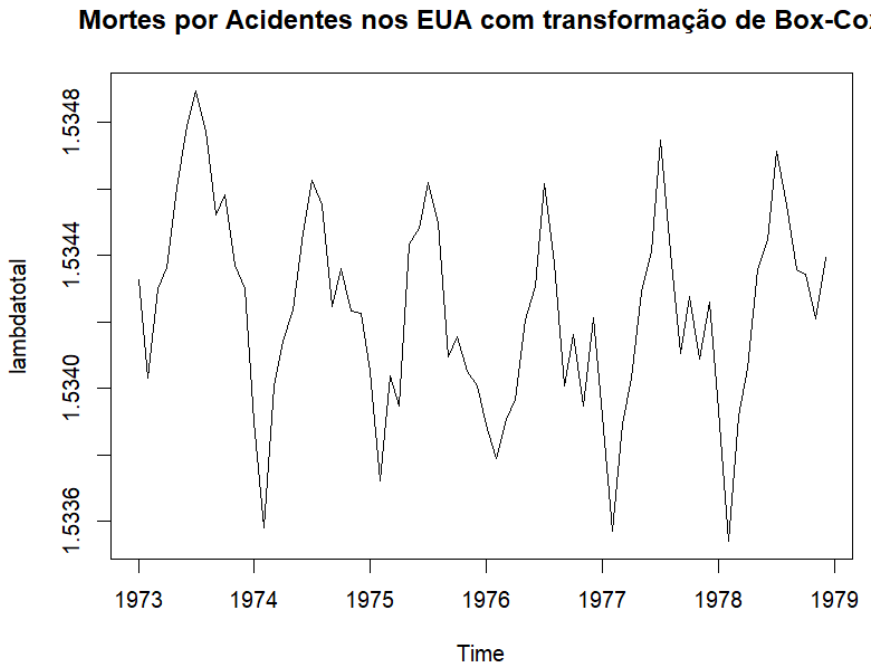

Figura 3.1: Série `USAccDeaths` após transformação de Box-Cox com $\lambda = -0.65$.

3.2 Análise Pós-transformação

Após a transformação de Box-Cox, procedeu-se à análise da estrutura de dependência temporal da série através da Função de Autocorrelação (FAC) e da Função de Autocorrelação Parcial (FACP). Esta análise é fundamental no processo de identificação de modelos ARIMA, uma vez que os padrões observados nestes gráficos permitem sugerir ordens adequadas para os componentes autorregressivos (AR), de médias móveis (MA) e sazonais.

A Figura 3.2 apresenta os gráficos da FAC e da FACP para a série transformada. A FAC (gráfico superior) apresenta um padrão de decréscimo lento e com picos regulares nos múltiplos de 12, indicando a presença de uma componente sazonal anual com período $s = 12$. Além disso, o facto de as autocorrelações diminuírem gradualmente sem corte abrupto é indicativo da presença de termos de tendência ou de integração (diferenciação regular necessária, $d \geq 1$).

A FACP (gráfico inferior), por outro lado, mostra um corte acentuado após a defasagem 1 (lag 1), com valores seguintes a caírem dentro dos limites de confiança. Este padrão é típico de um modelo AR(1), sugerindo a inclusão de um termo autorregressivo de ordem 1 na componente não sazonal.

De acordo com a tabela de identificação de modelos ARIMA (presente na sebenta), podemos caracterizar o comportamento observado da seguinte forma:

- **FAC:** Decréscimo gradual com picos em múltiplos de 12 $\Rightarrow$ necessidade de diferenciação + sazonalidade.
- **FACP:** Corte após lag 1 $\Rightarrow$ componente AR(1) possível.

Adicionalmente, foi utilizada a função `acf2()` do pacote `astsa`, que gera ambos os gráficos lado a lado, facilitando a comparação direta. Esta visualização combinada é especialmente útil na identificação preliminar de possíveis ordens dos modelos ARIMA e SARIMA.

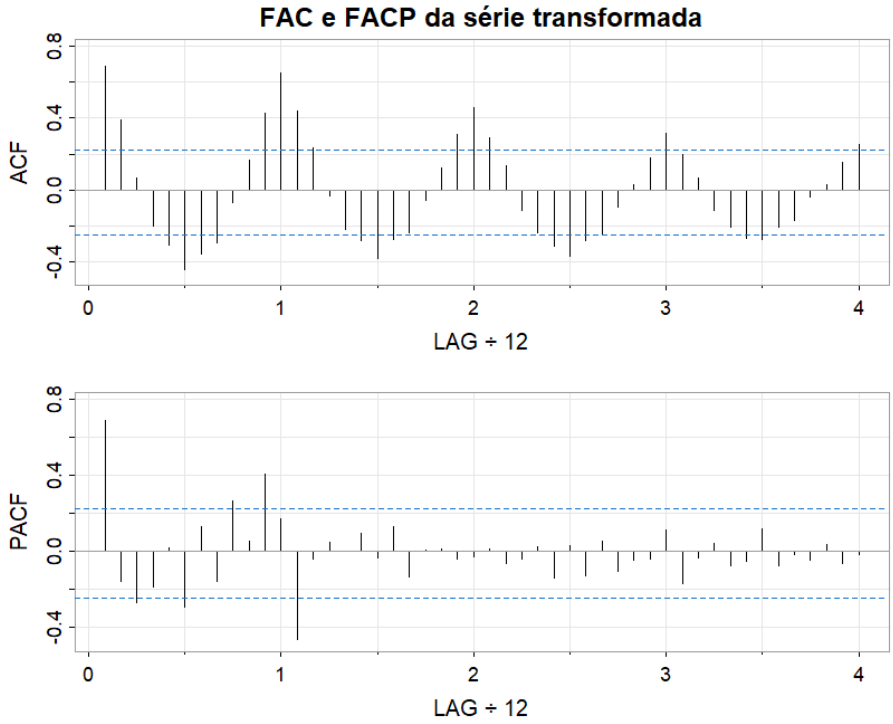

Figura 3.2: FAC e FACP da série transformada com Box-Cox ($\lambda = -0{,}65$). Padrão sazonal claro e evidência de componente AR(1).

—

4 Divisão da Série Temporal

4.1 Série de Treino e Série de Teste

Com a série já transformada, procedeu-se à sua divisão em dois subconjuntos: a **série de treino**, que inclui os primeiros 60 meses (Janeiro de 1973 a Dezembro de 1977), e a **série de teste**, que contém os últimos 12 meses (Janeiro a Dezembro de 1978). Esta partição respeita a boa prática de reservar cerca de 15% das observações para validação, garantindo pelo menos uma observação de cada mês. Deste modo, é assegurada a preservação da componente sazonal no conjunto de teste, permitindo validar previsões em todos os períodos do ciclo anual.

4.2 Visualização da Partição Temporal

A Figura 4.1 mostra a série total transformada com Box-Cox, evidenciando a divisão entre treino (linha preta) e teste (linha vermelha). Observa-se que a transição entre os dois subconjuntos é contínua, respeitando a estrutura temporal da série.

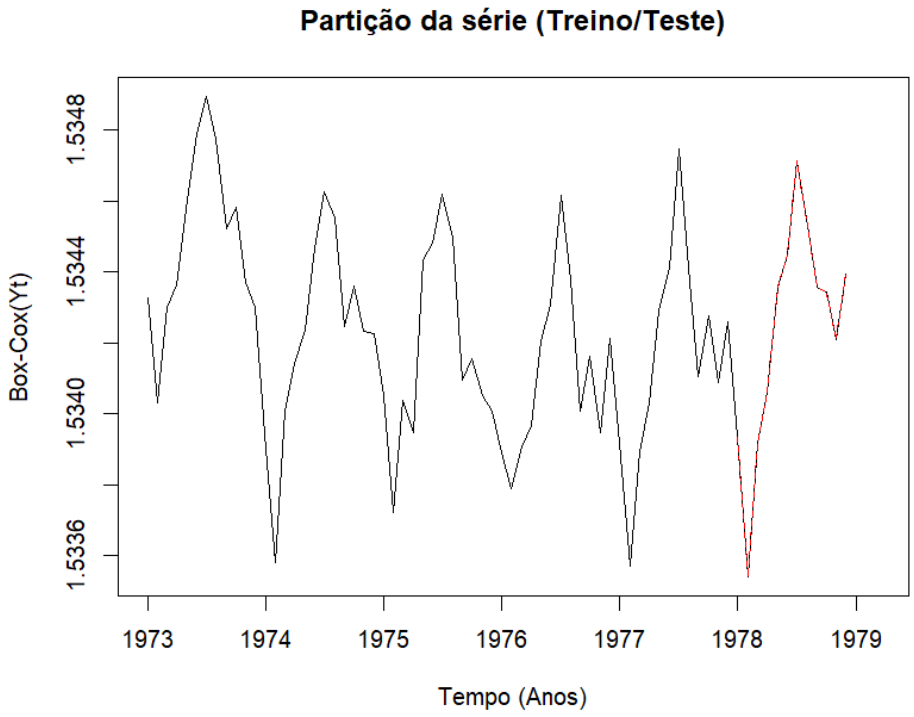

Figura 4.1: Partição da série `USAccDeaths` transformada: treino (linha preta) e teste (linha vermelha).

4.3 Decomposição da Série de Treino

Para estudar a estrutura interna da série de treino, foi novamente aplicada a decomposição aditiva, agora apenas sobre os primeiros 60 meses. A Figura 4.2 mostra os componentes extraídos da série transformada com Box-Cox.

Nesta decomposição, a **componente de tendência** (segundo painel) apresenta uma variação bastante suave, com um ligeiro declínio entre 1973 e 1976, seguido de uma recuperação subtil até ao final de 1977. Esta evolução sugere que a tendência global da série, embora presente, não é fortemente pronunciada — o que pode justificar a inclusão de uma diferenciação regular, mas não necessariamente múltiplos termos autorregressivos.

A **componente sazonal** (terceiro painel) mantém o padrão anual já observado na decomposição da série total: máximos nos meses de verão e mínimos no inverno. A estabilidade desta componente reforça a ideia de que a sazonalidade é persistente e regular, o que justifica a inclusão de uma estrutura sazonal explícita no modelo a ajustar.

Quanto à **componente aleatória** (último painel), esta revela alguma variabilidade, embora sem padrões sistemáticos evidentes. A maior dispersão relativa em comparação com a decomposição da série total pode ser atribuída ao menor número de observações no subconjunto de treino. Ainda assim, a componente aleatória aparenta ter média próxima de zero e variância constante.

Comparando com a decomposição da série completa, verifica-se que:

- A **tendência** na série total apresenta uma forma mais pronunciada e definida, enquanto na série de treino é mais discreta e suave;
- A **sazonalidade** mantém-se praticamente inalterada, o que confirma a sua estabilidade ao longo dos anos;

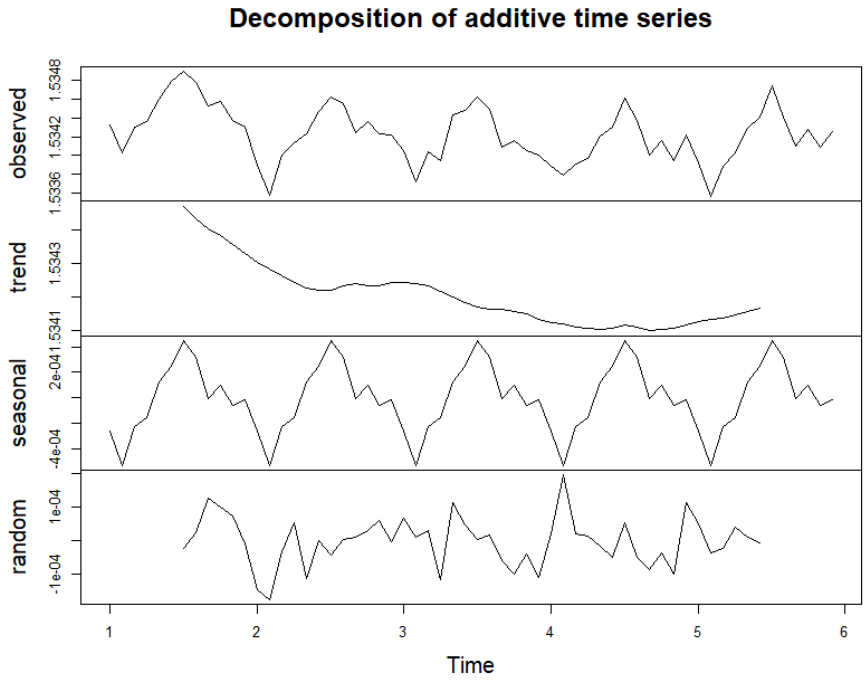

Figura 4.2: Decomposição da série de treino após transformação de Box-Cox.

- A **componente aleatória** tende a ser ligeiramente mais ruidosa na série de treino, mas sem comprometer a identificação clara das restantes componentes.

Esta análise confirma que a série de treino preserva as principais características estruturais da série total e encontra-se apta para ser usada na modelação. A estabilidade da sazonalidade e a presença de uma tendência suave indicam que a aplicação de um modelo do tipo SARIMA é apropriada para a série transformada.

5 Análise da Estacionariedade

A análise da estacionariedade é uma etapa essencial no processo de identificação de modelos ARIMA, uma vez que estes assumem que a série temporal a modelar deve ser estacionária — ou seja, ter média, variância e estrutura de autocorrelação constantes ao longo do tempo.

Após a transformação de Box-Cox, observou-se que a série ainda apresentava sinais de não estacionariedade, nomeadamente variações lentas na média e autocorrelações persistentes, o que foi visível tanto no gráfico da série como na função de autocorrelação (FAC). De modo a confirmar esta suspeita, foram também aplicados os testes estatísticos de raiz unitária (ADF) e de estacionariedade (KPSS), de acordo com os critérios conhecidos. Os testes sugeriram que a série não é estacionária na média, sendo necessária a aplicação de uma diferenciação regular. Além disso, o padrão sazonal detetado na FAC da série transformada indica a necessidade de uma diferenciação sazonal com período 12.

5.1 Análise Gráfica após Diferenciação

Foi aplicada uma diferenciação regular de ordem 1 à série transformada, com o objetivo de estabilizar a média. A nova série, correspondente aos resíduos do modelo ARIMA(0,1,0), foi analisada graficamente através da FAC e da FACP.

A Figura 5.1 apresenta, no painel superior, a série diferenciada, e nos painéis inferiores, os gráficos da FAC e FACP. A série diferenciada aparenta oscilar em torno de uma média constante,

sem tendência evidente, o que indica que a componente de tendência foi efetivamente removida. Ainda assim, é possível observar flutuações de amplitude regular, possivelmente associadas à componente sazonal.

A FAC (gráfico inferior esquerdo) mostra uma autocorrelação bastante elevada no lag 1, o que é típico após uma diferenciação regular, e evidencia também picos mais suaves em múltiplos de 12 períodos, compatíveis com sazonalidade anual. Já a FACP (gráfico inferior direito) apresenta um corte claro após a primeira desfasagem, sugerindo a presença de um termo AR(1) não sazonal, e valores baixos nos desfasamentos seguintes.

É de referir que, apesar do modelo automático sugerido pelo comando `auto.arima()` ter sido SARIMA(0,1,1)(0,1,1)[12], essa proposta foi apenas exploratória e não definitiva. De facto, o comando `ndiffs()` indicou que **não era necessária diferenciação regular** ($d = 0$), o que pode dever-se à tendência não ser monótona — ligeiramente decrescente no início da série e crescente no final — dificultando a sua deteção automática. Por outro lado, `nsdiffs()` sugeriu uma diferenciação sazonal de ordem 1 ($D = 1$), o que se confirma visualmente pela presença de padrão sazonal nas FACs.

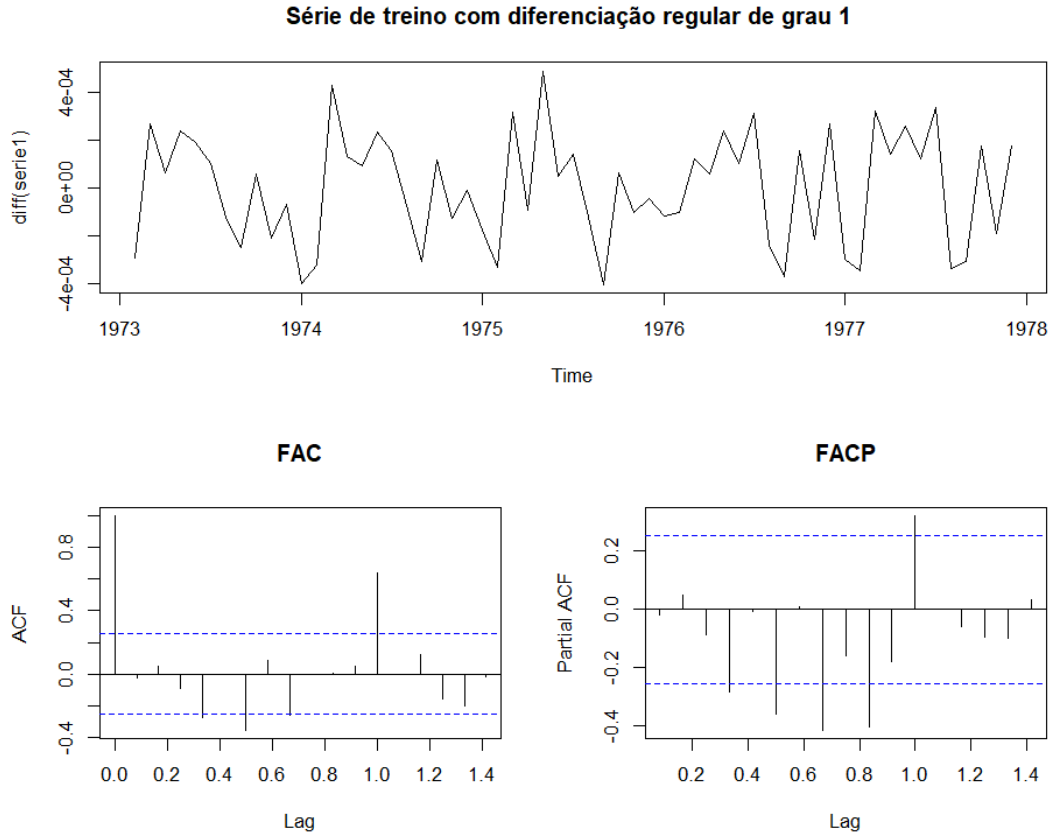

Figura 5.1: Série diferenciada de ordem 1 (cima), e gráficos da FAC e FACP (baixo). A tendência é removida, mas a estrutura sazonal persiste.

Esta análise confirma a necessidade de considerar modelos da classe SARIMA, com diferenciação regular e sazonal, e com termos AR e MA não sazonais e sazonais a serem ajustados nas próximas etapas.

Adicionalmente, foram novamente realizados os testes estatísticos anteriormente descritos — nomeadamente o teste de Dickey-Fuller aumentado (ADF) e o teste KPSS — aplicados agora à série diferenciada. Os resultados confirmaram formalmente a estacionariedade da série, validando a conclusão obtida pela análise gráfica. Assim, verifica-se que a diferenciação regular de ordem 1

foi eficaz na estabilização da média da série transformada.

6 Identificação da Sazonalidade

A identificação de sazonalidade é uma etapa fundamental na modelação de séries temporais, sobretudo em séries mensais ou trimestrais onde há ciclos recorrentes previsíveis ao longo do tempo. A sazonalidade representa variações sistemáticas que ocorrem em intervalos regulares, tipicamente associadas ao calendário (ex: meses do ano). Formalmente, diz-se que uma série apresenta sazonalidade com período s se existe correlação significativa entre observações separadas por múltiplos de s.

No caso da série `USAccDeaths`, a sazonalidade foi inicialmente identificada de forma gráfica na decomposição aditiva e reforçada pela análise das funções de autocorrelação (FAC e FACP). Em particular, a FAC da série transformada revelou picos regulares em lags múltiplos de 12, indicando um padrão sazonal anual bem definido. Este comportamento é típico de séries mensais com sazonalidade de $s = 12$, como também é sugerido pelas oscilações visíveis no gráfico da série diferenciada.

Como complemento, foi também calculado o **periodograma**, uma ferramenta da análise espectral que permite identificar frequências dominantes na série. O valor do período dominante é dado por:

$$\text{período} = \frac{1}{f_{\max}}$$

onde $f_{\max}$ é a frequência associada ao maior pico de energia espectral. No caso da série `USAccDeaths`, o periodograma indicou um valor aproximado de 12 como frequência dominante, o que está em conformidade com a análise da FAC.

Assim, com base tanto na análise de autocorrelações como no periodograma, foi identificado um **período sazonal de** $s = 12$. Este valor será assumido na construção dos modelos SARIMA nas secções seguintes, permitindo incluir termos sazonais autorregressivos e de média móvel adequados ao padrão anual detetado.

7 Modelação com a Metodologia Box-Jenkins

A modelação com base na metodologia de Box-Jenkins segue um processo sistemático e iterativo que inclui três fases principais: identificação do modelo, estimação dos parâmetros e validação dos resíduos. O objetivo é ajustar um modelo que capture a estrutura de dependência temporal da série e permita realizar previsões fiáveis, garantindo simultaneamente que os resíduos do modelo se comportam como ruído branco. Para isso, a série deve estar previamente estacionarizada tanto em média como em variância, como foi assegurado pelas etapas anteriores.

7.1 Identificação dos Parâmetros Sazonais (P, D, Q)

Após a estabilização da variância e da média da série, iniciou-se a fase de identificação da parte sazonal do modelo SARIMA. Começou-se por testar a presença de um termo autorregressivo sazonal de ordem 1 ($P = 1$), através do ajuste de um modelo SARIMA$(0,1,0)(1,0,0)_{12}$. O coeficiente estimado revelou-se estatisticamente significativo, o que justifica a sua inclusão no modelo. Dado o comportamento da FAC e a análise do coeficiente, concluiu-se que não era necessária a diferenciação sazonal, pelo que se considerou $D = 0$.

De seguida, foi testada a inclusão de um termo de média móvel sazonal ($Q = 1$), resultando no modelo SARIMA$(0,1,0)(1,0,1)_{12}$. Este modelo, além de apresentar ambos os coeficientes sazonais significativos, obteve também os melhores resultados em termos de critérios de qualidade (AIC e erro quadrático médio).

Foram ainda considerados modelos alternativos, como por exemplo modelos com $D = 1$, mesmo quando a metodologia indicava que não era necessário testá-los (pois P=1 é significativo). Estes modelos foram incluídos para comparação e confirmação da escolha ótima. A Tabela 7.1 resume os resultados obtidos para os modelos testados.

Tabela 7.1: Modelos testados para escolha de P, D e Q.

Modelo	**P**	**D**	**Q**	**AIC**
Modelo 1	1	0	0	-867.98
Modelo 2 (não necessário)	0	1	1	-708.91
Modelo 3 (escolhido)	1	0	1	**-874.86**
Modelo 4 (não necessário)	0	1	0	-699.58

Dado o desempenho superior do modelo 3 em termos de AIC e raiz do erro quadrático médio (REQM = 0.0002290), este foi selecionado como modelo final sazonal. Verificou-se que ambos os coeficientes estimados eram significativamente diferentes de zero, o que valida a sua inclusão. A escolha de $P = 1$, $D = 0$, $Q = 1$ permite capturar a estrutura sazonal identificada na análise exploratória e na FAC.

A Figura 7.1 apresenta os resíduos do modelo SARIMA$(0,1,0)(1,0,1)_{12}$. Observa-se que os resíduos se mantêm próximos de zero e não exibem padrões sistemáticos. A FAC e a FACP dos resíduos mostram valores dentro dos limites de confiança, o que sugere ausência de autocorrelação significativa — uma condição essencial para validar a adequação do modelo ajustado.

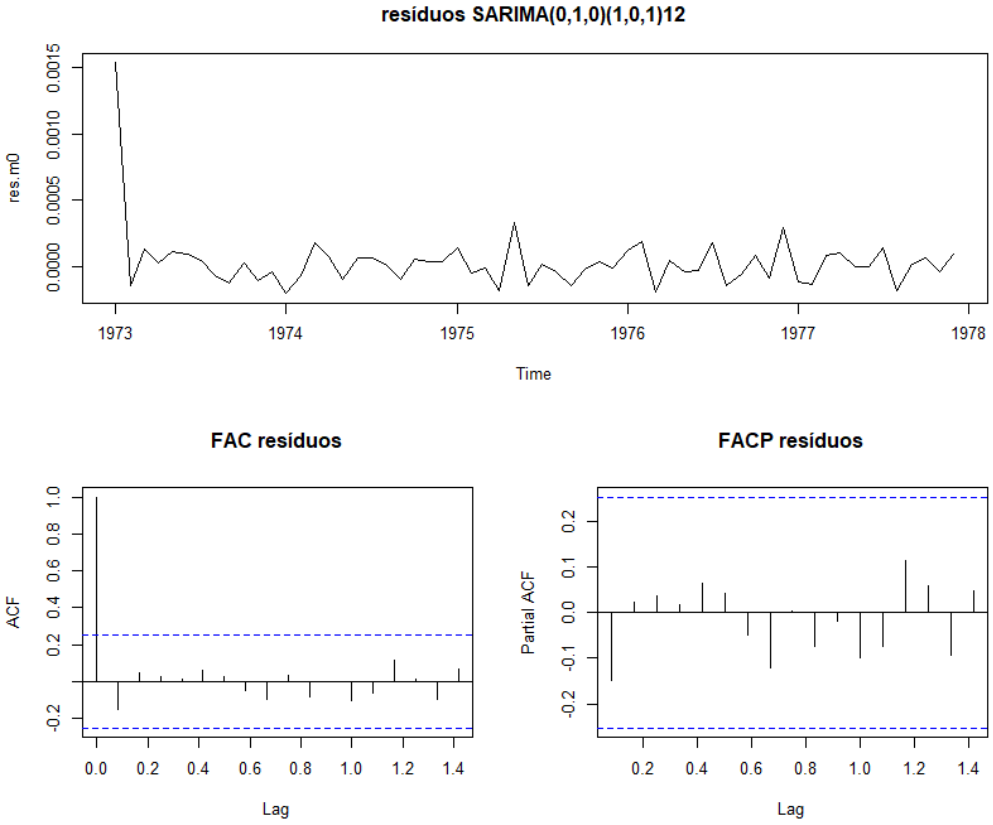

Figura 7.1: Resíduos e funções de autocorrelação (FAC e FACP) do modelo SARIMA$(0,1,0)(1,0,1)_{12}$.

Conclui-se, portanto, que a parte sazonal do modelo foi corretamente especificada, e o modelo

SARIMA$(0,1,0)(1,0,1)_{12}$ será usado como base para determinar os componentes não sazonais nas etapas seguintes.

7.2 Escolha dos Parâmetros Regulares (p, d, q)

Tendo definido previamente a componente sazonal como SARIMA$(\cdot,\cdot,\cdot)(1,0,1)_{12}$, procedeu-se à identificação da parte não sazonal do modelo, isto é, dos parâmetros p, d e q. Confirmada a necessidade de uma diferenciação regular ($d = 1$) com base na análise da estacionariedade e das funções de autocorrelação da série transformada, foram testadas sistematicamente várias combinações possíveis dos termos autorregressivos e de média móvel não sazonais.

Para cada combinação, avaliou-se o desempenho do modelo com base em critérios como o AIC (Critério de Informação de Akaike), a raiz do erro quadrático médio (REQM) e a significância estatística dos coeficientes estimados. O conjunto completo dos **16 modelos** SARIMA testados, bem como os respetivos coeficientes, AIC e REQM, encontra-se resumido na **Tabela A.1** no Anexo.

A Tabela 7.2 apresenta os 5 modelos com melhor desempenho global.

Tabela 7.2: Modelos com melhor desempenho para seleção de (p, d, q), com parte sazonal fixa $(1, 0, 1)_{12}$.

Modelo	**Ordem SARIMA**	**REQM**	**AIC**
Modelo 13	$(1,0,0)(1,0,1)_{12}$	**0.0001121**	**-891.39**
Modelo 16	$(0,1,1)(1,0,1)_{12}$	0.0002247	-878.77
Modelo 12	$(1,1,1)(1,0,1)_{12}$	0.0002249	-878.56
Modelo 15	$(1,1,0)(1,0,1)_{12}$	0.0002258	-877.19
Modelo 14	$(0,1,0)(1,0,1)_{12}$	0.0002290	-874.86

Apesar do modelo 13 apresentar o melhor AIC e o menor REQM, este inclui coeficientes com valores muito elevados (como o AR1 = 0.7804) e todos os parâmetros são altamente significativos, o que pode indicar sobreajustamento (overfitting), especialmente tendo em conta o reduzido tamanho da amostra (60 observações). Além disso, o modelo 13 não inclui diferenciação regular ($d = 0$), que se demonstrou necessária anteriormente pela análise gráfica e pelos testes ADF/KPSS.

O modelo 16, por outro lado, inclui a diferenciação regular $d = 1$, tem um número reduzido de parâmetros (apenas três), e apresenta um AIC próximo do modelo 13. Todos os coeficientes do modelo 16 são estatisticamente significativos (com p-valores inferiores a 0.05), oscilam em valores moderados, e têm interpretação coerente com a estrutura temporal da série.

Modelos como o 12 e o 15 também se destacam com AIC competitivo, mas envolvem mais parâmetros e não trazem uma melhoria substancial no REQM, pelo que foram descartados por não cumprirem o critério de parcimónia.

A Figura 7.2 apresenta os resíduos do modelo SARIMA$(0,1,1)(1,0,1)_{12}$, bem como os respetivos gráficos das funções de autocorrelação (FAC) e autocorrelação parcial (FACP). A análise da série de resíduos revela um comportamento oscilatório em torno de zero, sem tendência sistemática nem evidência clara de heterocedasticidade.

A FAC dos resíduos mostra que todas as autocorrelações estão dentro dos limites de confiança (linhas tracejadas), o que indica ausência de dependência temporal significativa. Por sua vez, a FACP não apresenta picos relevantes, reforçando a ideia de que os resíduos se comportam como ruído branco. O primeiro lag na FAC tem um valor ligeiramente acima de 0,1, mas não ultrapassa os limites críticos, sendo compatível com flutuações aleatórias.

Esta ausência de estrutura remanescente nos resíduos confirma que o modelo ajustado é adequado para a série transformada. Em particular, os resíduos não evidenciam autocorrelação

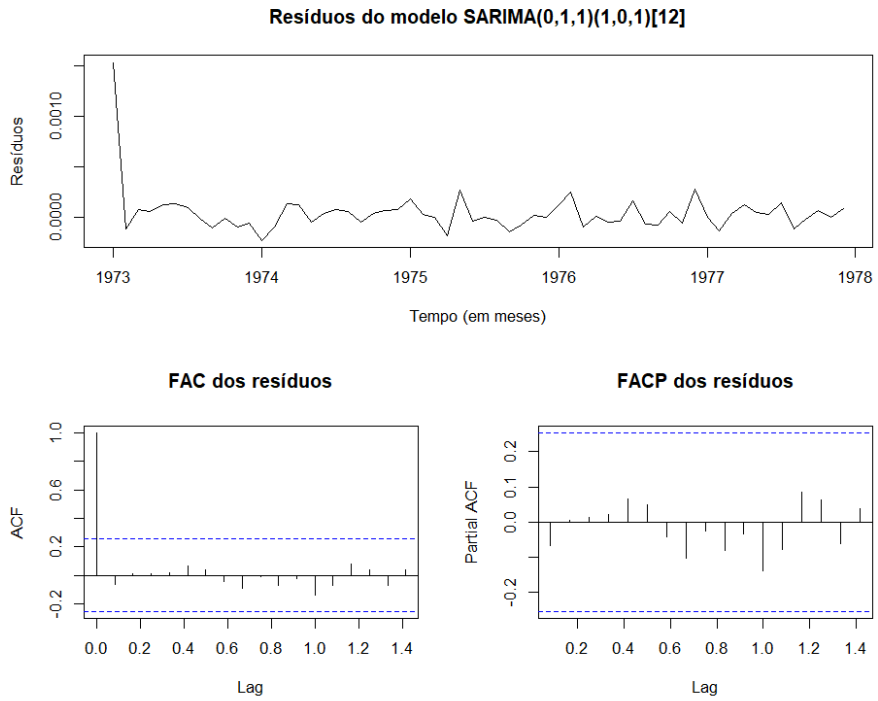

Figura 7.2: Resíduos e funções de autocorrelação do modelo final SARIMA$(0,1,1)(1,0,1)_{12}$.

significativa, validando a capacidade do modelo para capturar a dinâmica da série temporal original.

Representação do Modelo Final

O modelo final ajustado à série transformada foi do tipo SARIMA$(0,1,1)(1,0,1)_{12}$. A sua seleção resultou de um processo estruturado que incluiu:

- Aplicação da transformação de Box-Cox com $\lambda = -0.65$ para estabilizar a variância;
- Verificação da necessidade de uma diferenciação regular de ordem 1 ($d = 1$), mas sem diferenciação sazonal ($D = 0$);
- Análise gráfica das funções de autocorrelação (FAC) e autocorrelação parcial (FACP), que indicaram a presença de componentes sazonais com período $s = 12$;
- Comparação de múltiplos modelos com base nos critérios AIC, REQM e significância estatística dos coeficientes.

O modelo ajustado corresponde a um processo multiplicativo da forma:

$$(1 - \nu_1 B^{12})(1 - B)Y_t = (1 + \phi_1 B)(1 + \eta_1 B^{12})\varepsilon_t$$

Substituindo os coeficientes estimados:

$$(1 - 0.9821B^{12})(1 - B)Y_t = (1 + 0.2719B)(1 + 0.6117B^{12})\varepsilon_t$$

Equivalente a:

$$Y_t = Y_{t-1} + 0.9821Y_{t-12} - 0.9821Y_{t-13} + \varepsilon_t + 0.2719\varepsilon_{t-1} + 0.6117\varepsilon_{t-12} + 0.1665\varepsilon_{t-13}$$

Com:

$$\phi_1 = 0.2719, \quad \nu_1 = 0.9821, \quad \eta_1 = 0.6117$$

onde $\varepsilon_t \sim \mathcal{RB}(0, \sigma^2)$ representa ruído branco com média zero e variância constante.

Antes de avançar para a fase de previsão, procedeu-se à verificação formal da estacionariedade e da invertibilidade do modelo ajustado. Esta análise foi realizada com base nos polinómios característicos associados aos termos autorregressivos e de médias móveis, bem como através da aplicação do teste de estacionariedade KPSS.

A estacionariedade de um processo exige que todas as raízes do polinómio autorregressivo (incluindo o componente sazonal) estejam fora do círculo unitário, isto é, que o módulo de cada raiz seja superior a 1. De igual forma, a invertibilidade requer que as raízes do polinómio de médias móveis também se situem fora do círculo unitário.

Para o modelo SARIMA$(0,1,1)(1,0,1)_{12}$, foram calculadas as raízes dos três polinómios envolvidos:

- A raiz do termo MA(1) regular apresentou módulo > 2.61, confirmando a invertibilidade.
- A raiz do termo AR(1) sazonal apresentou módulo ≈ 1.02, ligeiramente acima de 1, garantindo a estacionariedade sazonal.
- A raiz do termo SMA(1) apresentou módulo > 1.73, o que também confirma a invertibilidade sazonal.

Adicionalmente, aplicou-se o teste de estacionariedade de Kwiatkowski-Phillips-Schmidt-Shin (KPSS) sobre os resíduos do modelo. O valor-p obtido foi superior a 0.1, pelo que não se rejeita a hipótese nula de estacionariedade, o que valida estatisticamente o comportamento estacionário dos resíduos.

Conclui-se assim que o modelo ajustado é estacionário e invertível, satisfazendo todos os pressupostos teóricos para ser considerado adequado do ponto de vista estrutural e estatístico.

Este modelo será agora utilizado para gerar previsões e avaliar o seu desempenho preditivo na série de teste.

8 Análise dos Resíduos

Uma vez ajustado o modelo SARIMA$(0,1,1)(1,0,1)_{12}$, procedeu-se à validação do mesmo através da análise dos resíduos, de forma a verificar se estes se comportam como ruído branco. Foram analisados três aspetos fundamentais: independência temporal, normalidade e média nula.

Independência dos resíduos

Foi aplicado o teste de Ljung-Box, com $h = 24$ desfasagens e $k = 3$ parâmetros estimados no modelo. O valor-p obtido foi superior a 0.99, o que não permite rejeitar a hipótese nula de ausência de autocorrelação nos resíduos. Para confirmação adicional, avaliou-se um conjunto mais alargado de desfasagens ($h = 4$ até 36), onde os valores-p mantiveram-se consistentemente acima de 0.05, o que reforça a conclusão de que os resíduos não apresentam autocorrelação significativa.

Normalidade dos resíduos

A normalidade foi avaliada com dois testes complementares: o teste de Shapiro-Wilk e o teste de Kolmogorov-Smirnov, este último comparando a distribuição empírica com uma normal teórica com média e desvio padrão amostrais. Em ambos os casos, os valores-p elevados (superiores a 0.1) indicam não haver evidência estatística para rejeitar a hipótese de normalidade.

Média dos resíduos

O teste t para a média amostral foi aplicado com hipótese nula de $\mu = 0$. O valor-p não significativo confirma que a média dos resíduos não difere significativamente de zero.

Análise gráfica dos resíduos

A Figura 8.1 apresenta um diagnóstico gráfico completo dos resíduos padronizados do modelo ajustado.

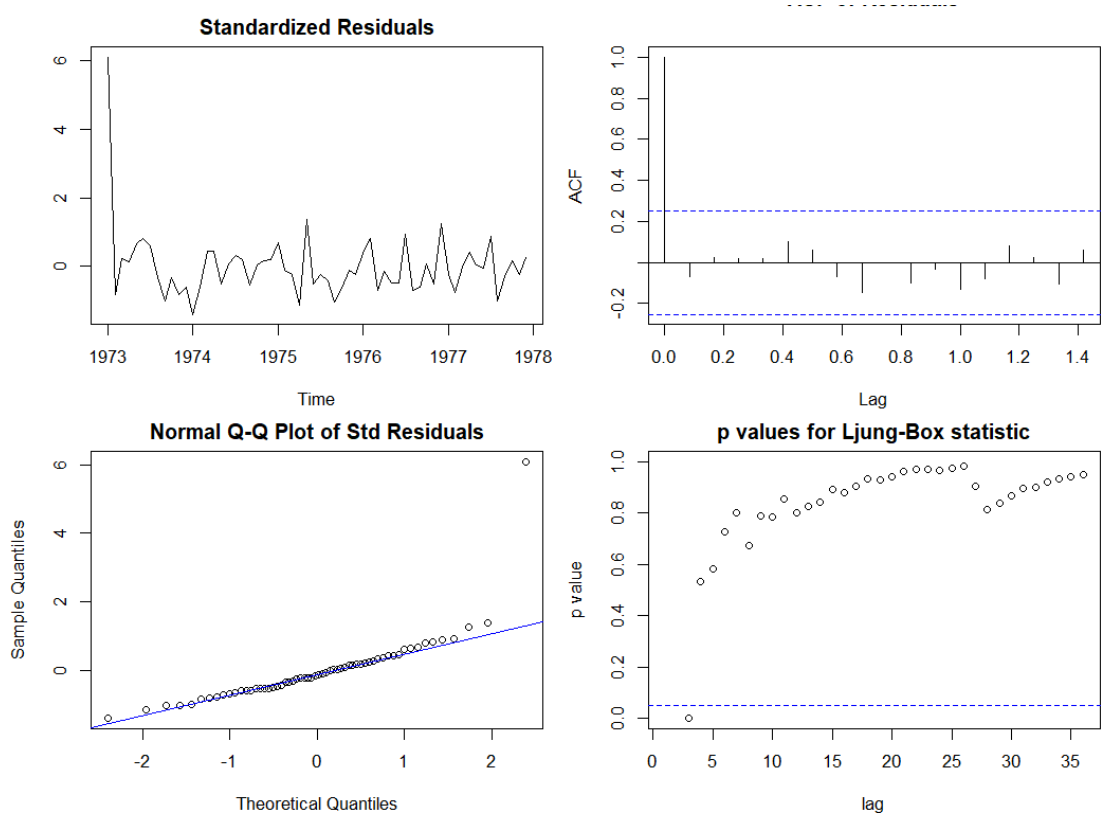

Figura 8.1: Diagnóstico gráfico dos resíduos padronizados do modelo SARIMA$(0,1,1)(1,0,1)_{12}$.

O gráfico superior esquerdo mostra a evolução temporal dos resíduos padronizados, que oscilam de forma irregular em torno de zero e não revelam qualquer padrão sistemático, o que é característico de ruído branco.

No canto superior direito, observa-se a função de autocorrelação (ACF) dos resíduos. Todas as autocorrelações encontram-se dentro dos limites de confiança, o que é consistente com a ausência de dependência temporal.

O gráfico inferior esquerdo apresenta o Q-Q plot dos resíduos padronizados. A maioria dos pontos alinha-se de forma satisfatória com a reta teórica, excetuando ligeiras discrepâncias nas caudas, que não comprometem de forma significativa a normalidade global.

Por fim, o gráfico inferior direito exibe os valores-p do teste de Ljung-Box para lags de 1 a 35. Todos os valores-p permanecem acima do nível de significância de 5%, confirmando graficamente a independência dos resíduos.

Estes resultados indicam que os resíduos do modelo ajustado satisfazem os pressupostos de ruído branco: são independentes, aproximadamente normais e com média nula. Assim, considera-se que o modelo SARIMA$(0,1,1)(1,0,1)_{12}$ é adequado para representar a estrutura da série temporal transformada.

9 Previsão

A previsão é uma das principais finalidades da modelação de séries temporais. Uma vez ajustado e validado o modelo, é possível utilizá-lo para estimar valores futuros da série, bem como quantificar a incerteza associada através de intervalos de previsão. Neste estudo, o modelo SARIMA$(0,1,1)(1,0,1)_{12}$, ajustado à série transformada por Box-Cox com $\lambda = -0.65$, foi utilizado para prever os 12 meses seguintes ao período de treino, correspondentes ao ano de 1978.

9.1 Previsões na Escala Box-Cox

A Figura 9.1 mostra a comparação entre os valores observados (a preto) e os valores ajustados (a vermelho) na fase de treino, utilizando a série transformada por Box-Cox.

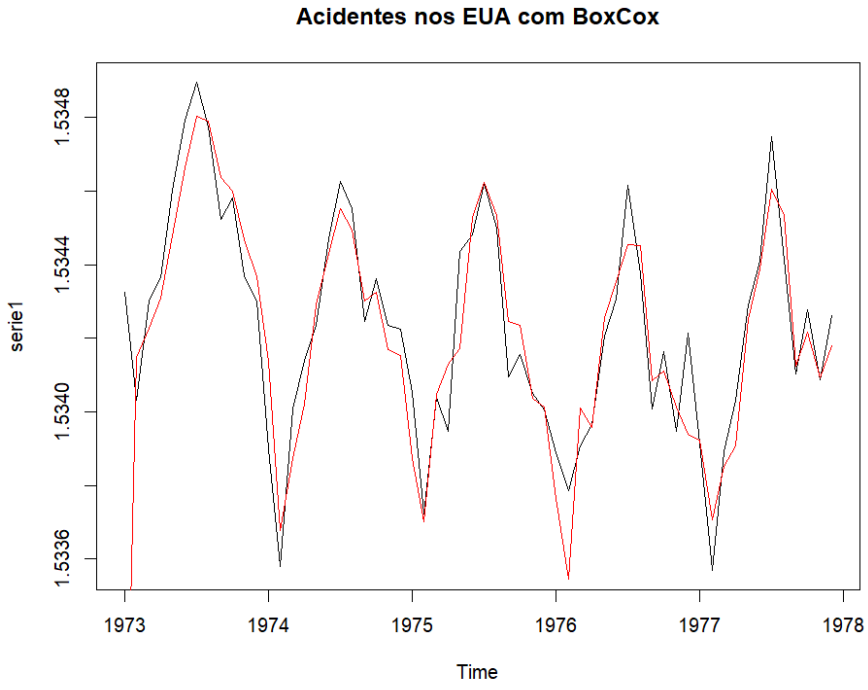

Figura 9.1: Ajustamento na fase de treino na escala Box-Cox.

Observa-se que o modelo acompanha adequadamente a série de treino, replicando os principais picos e vales. A aderência entre as linhas é elevada, com desvios residuais pequenos e sem padrão evidente.

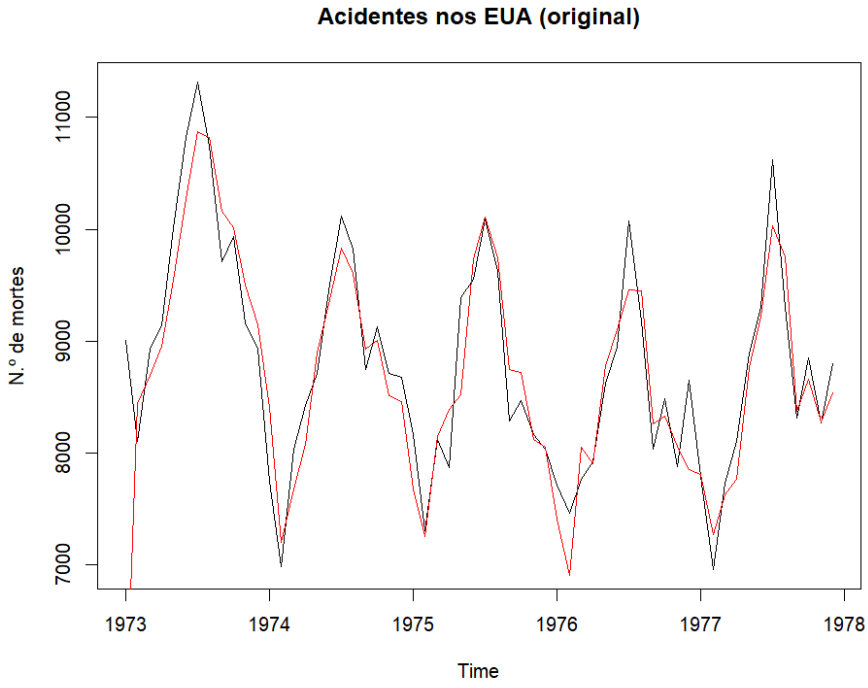

Figura 9.2: Ajustamento na fase de treino na escala original.

Na Figura 9.2, os mesmos resultados são apresentados após a inversão da transformação de Box-Cox. O modelo continua a ajustar bem os dados na escala original, embora se observe uma ligeira suavização dos picos mais acentuados, típica deste tipo de transformação.

9.2 Previsões na Fase de Teste

Para a fase de teste (ano de 1978), a Figura 9.3 apresenta a evolução da série transformada, com os valores observados a preto, os estimados na fase de treino a vermelho, e os valores previstos a azul.

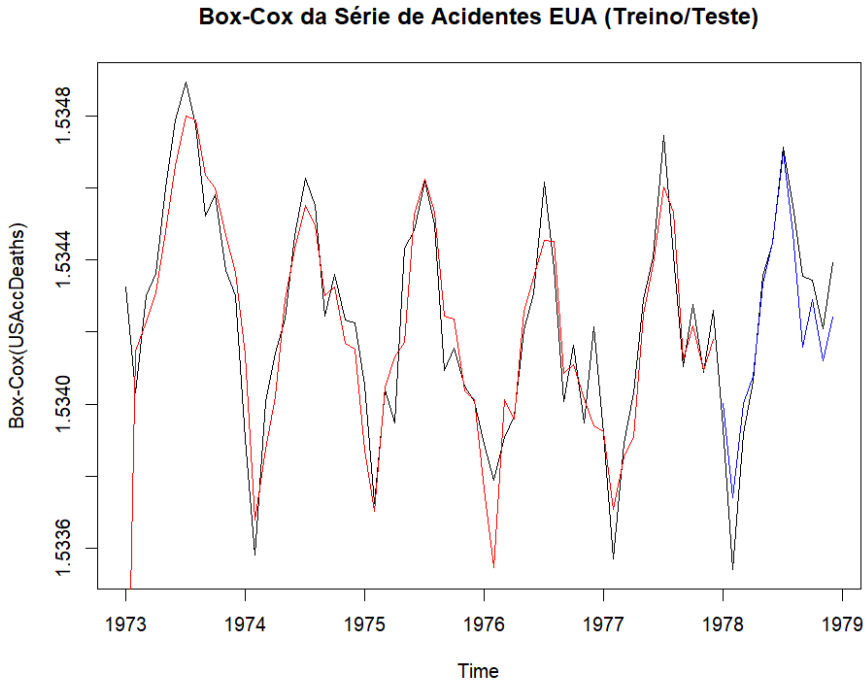

Figura 9.3: Série transformada por Box-Cox: treino, teste e previsões.

O modelo mantém boa capacidade de replicar a sazonalidade da série, mesmo fora da amostra de treino. A evolução geral é captada corretamente, ainda que se verifiquem pequenas discrepâncias nos picos — por exemplo, no verão de 1978, as previsões não atingem o valor máximo observado.

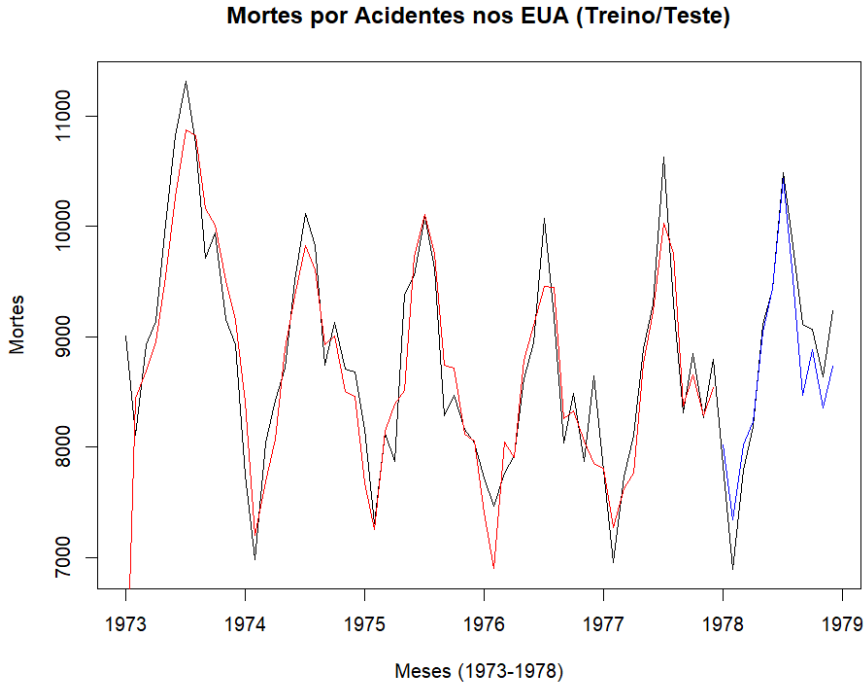

Figura 9.4: Série na escala original com treino e teste (valores reais e previstos).

A Figura 9.4 mostra a mesma informação após inversão da transformação de Box-Cox. Nesta

escala, as previsões continuam a refletir corretamente a estrutura da série, embora com ligeira subestimação nos meses de verão. Nota-se que os meses com valores mais baixos (como fevereiro e novembro) são estimados com maior precisão.

9.3 Intervalos de Confiança

A Figura 9.5 inclui os intervalos de previsão a 95%, desenhados a azul tracejado. Observam-se os valores reais a preto, os estimados (fase de treino) a vermelho, e as previsões (fase de teste) a azul.

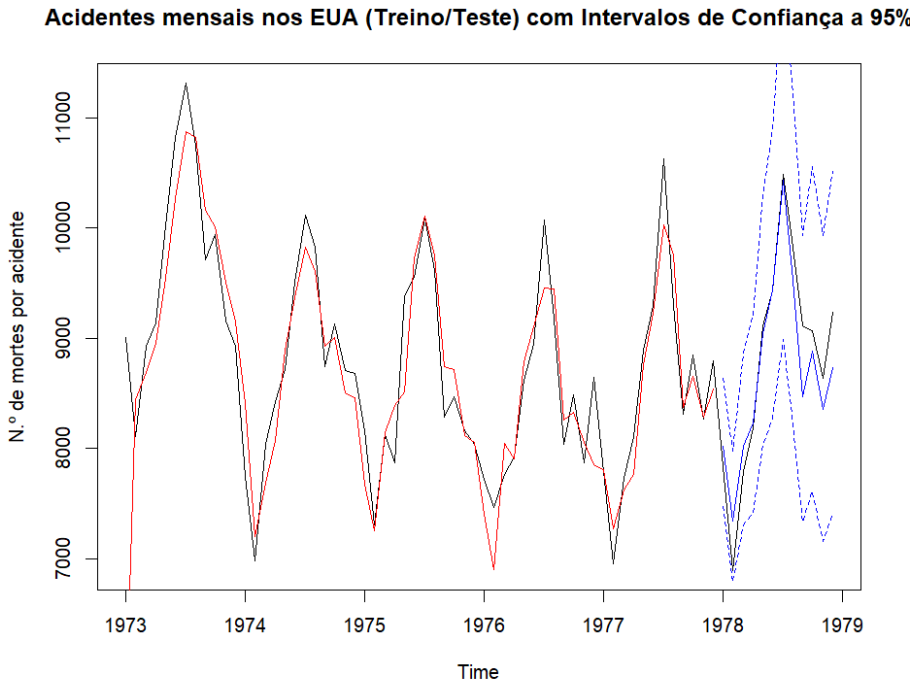

Figura 9.5: Previsões com intervalos de confiança a 95% (escala original).

Os intervalos de previsão alargam-se gradualmente à medida que nos afastamos da fase de treino, refletindo o aumento natural da incerteza. Apesar disso, praticamente todos os valores reais encontram-se dentro do intervalo previsto. Isto evidencia a robustez do modelo não só na replicação da estrutura interna da série, como também na sua capacidade preditiva.

De forma geral, os resultados obtidos demonstram que o modelo SARIMA$(0,1,1)(1,0,1)_{12}$ é apropriado tanto para descrever como para prever a evolução mensal do número de mortes por acidentes nos Estados Unidos no período em análise.

10 Avaliação do Modelo

Após a realização das previsões com o modelo SARIMA$(0,1,1)(1,0,1)_{12}$, é fundamental avaliar a qualidade do ajustamento e da previsão. Para isso, recorreu-se a um conjunto de métricas de erro amplamente utilizadas em análise de séries temporais:

- **EQM — Erro Quadrático Médio:** Mede o erro médio ao quadrado entre os valores previstos e observados. Penaliza fortemente erros grandes.
- **REQM — Raiz do Erro Quadrático Médio:** É a raiz quadrada do EQM, permitindo interpretar o erro médio na unidade original da série.

- **EPAM — Erro Percentual Absoluto Médio:** Expressa o erro médio em termos percentuais relativamente aos valores observados, sendo uma métrica adimensional útil para comparação.

Estas métricas foram calculadas separadamente para a série de treino (valores ajustados) e para a série de teste (valores previstos fora da amostra).

10.1 Métricas de Avaliação na Série de Treino

Tabela 10.1: Métricas de avaliação para o ajustamento na série de treino (1973–1977).

Métrica	Valor	Unidade	Interpretação
EQM	309669.40	mortes^2	Erro médio quadrático durante o treino
REQM	556.48	mortes	Erro médio absoluto na escala da série
EPAM	3.65%	percentual	Erro médio percentual absoluto

Os resultados obtidos indicam que, em média, o modelo comete um erro de cerca de 556 mortes por mês durante o período de treino. Este valor é considerado aceitável, tendo em conta que os valores mensais da série oscilam entre 7000 e 11000 mortes. O erro percentual absoluto médio inferior a 4% reforça a qualidade do ajustamento.

10.2 Métricas de Avaliação na Série de Teste

Tabela 10.2: Métricas de avaliação para a previsão na série de teste (1978).

Métrica	Valor	Unidade	Interpretação
EQM	96513.94	mortes^2	Erro médio quadrático fora da amostra
REQM	310.67	mortes	Erro médio absoluto nas previsões
EPAM	2.90%	percentual	Erro percentual médio absoluto

Na fase de teste, o erro quadrático médio foi inferior ao da fase de treino, o que, neste caso, indica boa generalização do modelo. A previsão média apresenta um desvio de aproximadamente 311 mortes mensais, valor ainda mais reduzido do que o registado na fase de ajustamento. O EPAM inferior a 3% confirma que as previsões são bastante fiáveis, mesmo fora da amostra de treino.

10.3 Discussão dos Resultados

Tabela 10.3: Comparação das métricas de avaliação entre treino e teste.

Métrica	Treino	Teste	Melhor desempenho
EQM	309669.40	96513.94	Teste
REQM	556.48	310.67	Teste
EPAM	3.65%	2.90%	Teste

A comparação direta entre as métricas revela que o modelo apresenta desempenho ligeiramente superior na fase de teste. Este resultado pode parecer contra intuitivo, dado que o modelo foi ajustado com base na série de treino. No entanto, é explicado pelo facto de os valores de treino

incluírem observações com maior variabilidade (ex. picos extremos em 1973–1974), enquanto a série de teste possui oscilações mais regulares e com amplitude mais contida.

Além disso, a utilização da transformação de Box-Cox contribuiu para suavizar a variabilidade e melhorar a capacidade de previsão. O EPAM global inferior a 4% tanto em treino como em teste demonstra que o modelo é estável e confiável para fins de previsão.

Estes resultados reforçam a adequação do modelo SARIMA$(0,1,1)(1,0,1)_{12}$ como solução eficaz para modelar e prever o número mensal de mortes por acidentes nos EUA no período analisado.

11 Conclusão

O presente trabalho teve como principal objetivo a modelação e previsão do número mensal de mortes acidentais nos Estados Unidos, no período entre janeiro de 1973 e dezembro de 1978, com base na série temporal `USAccDeaths`. Para tal, foi seguida a metodologia Box-Jenkins, envolvendo várias etapas fundamentais, desde a análise exploratória à validação do modelo final.

Inicialmente, foi realizada uma transformação de Box-Cox com $\lambda = -0.65$, com o intuito de estabilizar a variância da série. Esta transformação revelou-se apropriada à luz da distribuição assimétrica da série original. Após análise gráfica das funções de autocorrelação (FAC) e autocorrelação parcial (FACP), identificou-se a necessidade de uma diferenciação regular de ordem 1, mas não de diferenciação sazonal. A decomposição revelou uma forte componente sazonal anual, o que orientou a escolha do modelo.

Foram testados múltiplos modelos SARIMA com diferentes combinações de parâmetros, tendo sido selecionado, com base em critérios estatísticos (AIC, REQM, significância dos coeficientes), o modelo $\mathbf{SARIMA(0,1,1)(1,0,1)_{12}}$. Este modelo demonstrou ser parcimonioso, estável e com capacidade de ajustamento adequada. Os resíduos mostraram-se independentes, aproximadamente normais e com média nula, satisfazendo os pressupostos fundamentais da modelação.

Em termos preditivos, o modelo foi avaliado com base em métricas como o erro quadrático médio (EQM), a sua raiz (REQM) e o erro percentual absoluto médio (EPAM). Tanto na fase de treino como na de teste, os resultados revelaram valores baixos destas métricas, com o EPAM inferior a 4%, o que demonstra a fiabilidade do modelo em diferentes contextos. A análise visual das previsões e dos respetivos intervalos de confiança reforçou esta conclusão, dado que os valores observados se mantiveram, em quase todos os casos, dentro dos limites previstos.

Limitações e Perspetivas Futuras

Embora o modelo tenha apresentado bom desempenho, algumas limitações podem ser apontadas. A série considerada é curta (apenas 72 observações), o que condiciona a complexidade dos modelos possíveis e a robustez estatística de alguns testes. Além disso, os dados dizem respeito apenas ao número total de mortes, sem distinção por causas específicas, localizações ou perfis demográficos, o que limita a interpretação de padrões mais profundos.

Como propostas futuras, poderia considerar-se o alargamento do horizonte temporal, a introdução de covariáveis externas (como clima, feriados ou tráfego) ou a utilização de métodos mais avançados, como modelos SARIMAX, modelos estruturais ou abordagens baseadas em redes neuronais. Estes métodos poderiam permitir melhorar ainda mais a qualidade das previsões, sobretudo em contextos com maior variabilidade ou perturbações sazonais irregulares.

Em suma, a metodologia aplicada permitiu desenvolver um modelo estatístico robusto e eficaz, capaz de representar e prever com sucesso a evolução da série `USAccDeaths`, respeitando os princípios da modelação de séries temporais.

Referências Bibliográficas

- Alpuim, T. (1997). *Séries Temporais*. Universidade Nova de Lisboa, Faculdade de Ciências e Tecnologia.
- Ehlers, R. (2009). *Análise de Séries Temporais*. Universidade Federal do Rio Grande do Sul.
- Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and Practice* (3rd ed.). OTexts. Disponível em: `https://otexts.com/fpp3/`
- Vilares, R. (2023). On testing for stationarity of a time series: A comprehensive review. *Journal of Applied Statistics*.
  DOI: `https://doi.org/10.1080/02664763.2023.2238249`

A ANEXO - Resumo dos Modelos Testados

Nesta secção apresenta-se o resumo dos 16 modelos SARIMA testados, com os respetivos parâmetros, valores de AIC e raiz do erro quadrático médio (REQM).

Tabela A.1: Modelos SARIMA testados com respetivos coeficientes estimados, AIC e REQM.

Modelo	**Ordem SARIMA**	**Coeficientes estimados**	**AIC**	**REQM**
Modelo 13	$(1,0,0)(1,0,1)_{12}$	AR1 = 0.7804, SAR1 = 0.9673, SMA1 = 0.5529	**-891.39**	0.00011207
Modelo 16	$(0,1,1)(1,0,1)_{12}$	MA1 = 0.3822, SAR1 = 0.9815, SMA1 = -0.5763	-878.77	0.00022472
Modelo 12	$(1,1,1)(1,0,1)_{12}$	AR1 = 0.5142, MA1 = 0.8420, SAR1 = 0.9772, SMA1 = 0.5792	-878.56	0.00022495
Modelo 15	$(1,1,0)(1,0,1)_{12}$	AR1 = 0.2719, SAR1 = 0.9821, SMA1 = -0.6117	-877.19	0.00022581
Modelo 14	$(0,1,0)(1,0,1)_{12}$	SAR1 = 0.9780, SMA1 = 0.6149	-874.86	0.00022898
Modelo 7	$(2,1,0)(1,0,1)_{12}$	AR1 = -0.3274, AR2 = -0.1679, SAR1 = 0.9799, SMA1 = -0.5617	-876.78	0.00022491
Modelo 11	$(1,1,2)(0,1,1)_{12}$	AR1 = -0.8320, MA1 = 0.4539, MA2 = 0.5399, SMA1 = 0.5519	-713.95	0.00072010
Modelo 3	$(1,1,1)(0,1,1)_{12}$	AR1 = 0.1482, MA1 = 0.6091, SMA1 = 0.5695	-714.20	0.00072056
Modelo 10	$(0,1,2)(0,1,1)_{12}$	MA1 = -0.4532, MA2 = -0.0869, SMA1 = 0.5661	-714.24	0.00072057
Modelo 5	$(3,1,0)(0,1,1)_{12}$	AR1 = -0.4324, AR2 = -0.2545, AR3 = -0.0238, SMA1 = -0.5471	-711.78	0.00072068
Modelo 2	$(2,1,0)(0,1,1)_{12}$	AR1 = -0.4259, AR2 = -0.2436, SMA1 = -0.5506	-713.75	0.00072067
Modelo 1	$(1,1,0)(0,1,1)_{12}$	AR1 = -0.3398, SMA1 = 0.6212	-712.84	0.00072088
Modelo 4	$(2,1,1)(0,1,1)_{12}$	AR1 = 0.0727, AR2 = 0.0570, MA1 = 0.5313, SMA1 = 0.5587	-712.25	0.00072058
Modelo 6	$(2,1,0)(1,1,1)_{12}$	AR1 = -0.4425, AR2 = -0.2703, SAR1 = -0.2266, SMA1 = -0.2868	-711.32	0.00072088
Modelo 8	$(2,1,0)(0,1,2)_{12}$	AR1 = -0.4206, AR2 = -0.2350, SMA1 = -0.5904, SMA2 = -0.0800	-711.80	0.00072042
Modelo 9	$(0,1,1)(0,1,1)_{12}$	MA1 = -0.4982, SMA1 = -0.5655	-715.92	0.00072061