INSTITUTO UNIVERSITÁRIO DE LISBOA

MODELOS LONGITUDINAIS

Hóspedes Estrangeiros em Estabelecimentos de Alojamento Turístico em Portugal - Análise de Série Temporal

Realizado por:

António Duarte Santos – 123434

Carolina Morgado – 123794

Diogo Nobre – 123441

Gonçalo Henriques – 123422

Inês Silva – 123407

José Alberto – 121959

Turma CD2

2025/2026

1. Tema e Objetivo do Estudo

O presente trabalho tem como objetivo analisar a série temporal referente ao número de hóspedes estrangeiros em estabelecimentos de alojamento turístico em Portugal, entre janeiro de 2013 e julho de 2025.

A escolha deste tema justifica-se pela relevância do turismo em Portugal, a nível económico, visto que é um setor crucial para a criação de receita, emprego e investimento. Assim, ao analisarmos a sua evolução, é possível identificar padrões e antecipar variações importantes para o planeamento económico turístico.

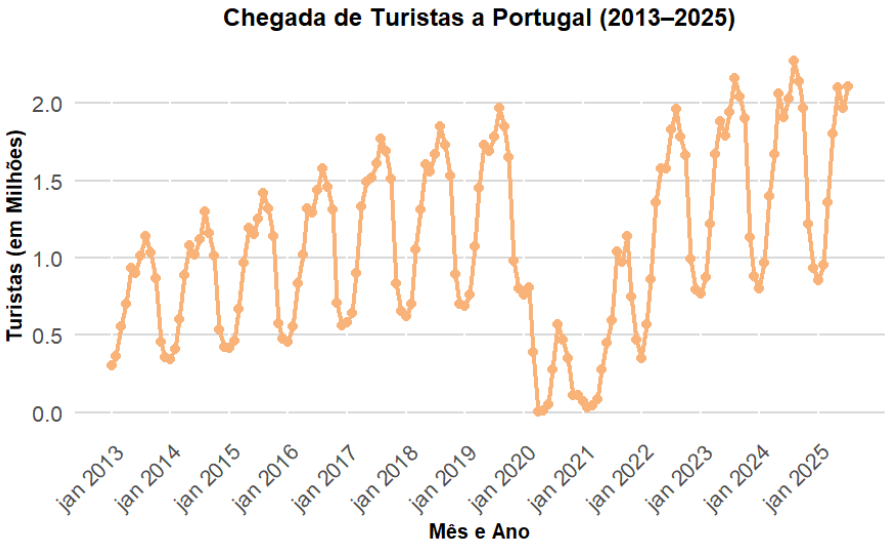
2. Caracterização dos dados

Os dados utilizados foram obtidos manualmente através do *site Trading Economics,* em formato *.csv*, contendo apenas duas variáveis: ***Date***, que representa o mês e o ano da observação, e ***Tourists***, que corresponde ao número de turistas (em milhões) registados nesse período.

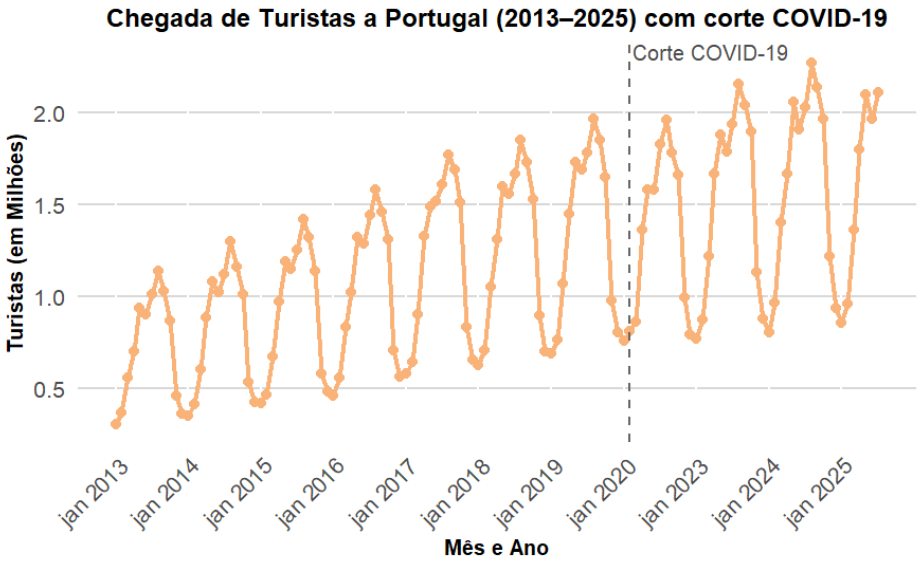
A base de dados contém 151 observações mensais, abrangendo o intervalo de janeiro de 2013 a julho de 2025. Os dados foram importados e tratados em *R*, recorrendo à biblioteca *fpp3*, que integra pacotes como *tsibble, feasts* e *fable* adequados à análise, decomposição e modelação de séries temporais. Após a importação, o campo *Date* foi convertido para o formato *yearmonth* e a base de dados foi transformada num objeto *tsibble*, assegurando a estrutura temporal necessária para os procedimentos seguintes.

3. Análise exploratória

Foi realizada uma primeira representação gráfica da série, que mostrou uma trajetória ascendente das chegadas de turistas desde 2013 até 2019, seguida de uma queda abrupta em 2020, coincidindo com as restrições impostas pela pandemia COVID-19. A partir de 2022, observa-se uma recuperação do turismo, atingindo valores próximos ou superiores aos níveis pré-pandemia.

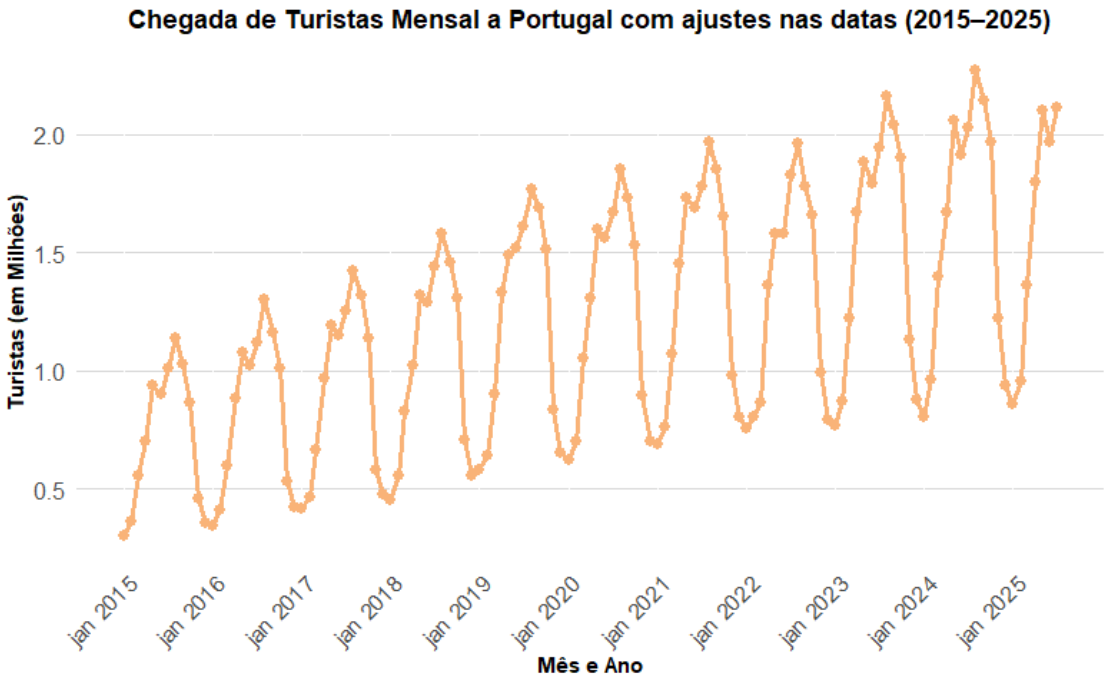


Para reduzir o impacto deste evento atípico, foi criada uma versão alternativa do conjunto de dados, excluindo o período compreendido entre março de 2020 e fevereiro de 2022, permitindo uma análise mais fiel das componentes estruturais da série, como tendência e sazonalidade, sem distorções causadas por fatores externos.



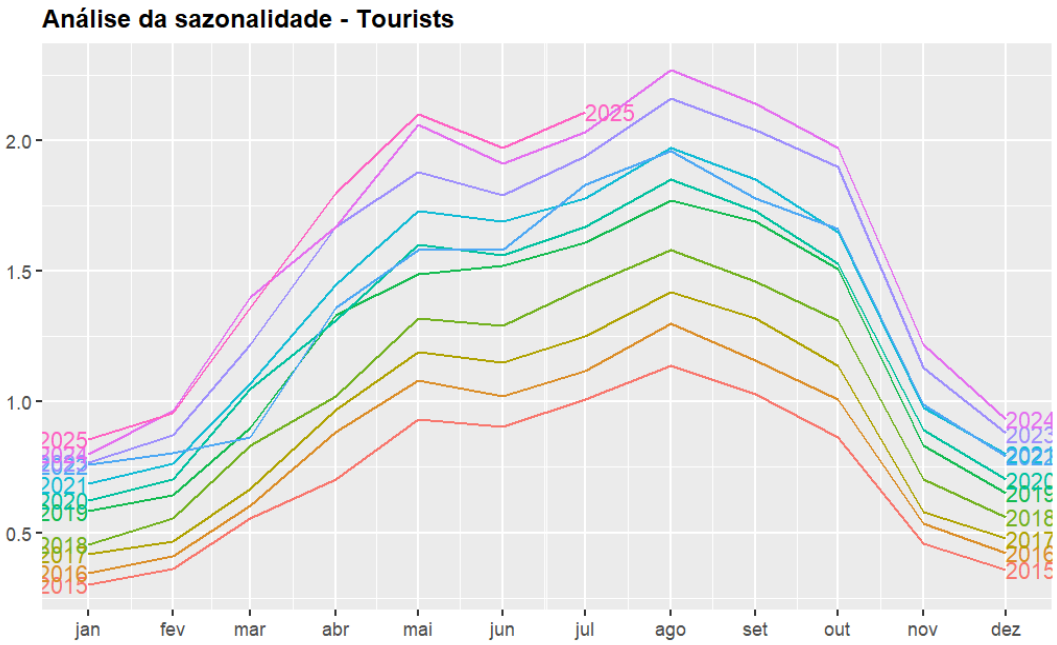
Para possibilitar a análise dos *lags*, autocorrelações e ciclos sazonais, foi criado outro *dataset* com a variável temporal ajustada, de forma a compensar as perturbações causadas pela pandemia de COVID-19. Assim, janeiro de 2013 passou a corresponder a janeiro de 2015, e fevereiro de 2020 a fevereiro de 2022, mantendo-se a restante série temporal, de março de 2022 a julho de 2025, inalterada.

Este ajuste foi necessário porque, após o corte na série temporal, excluindo o período pandémico, o R deixou de reconhecer a descontinuidade temporal e, consequentemente, não permitia prosseguir com a análise.

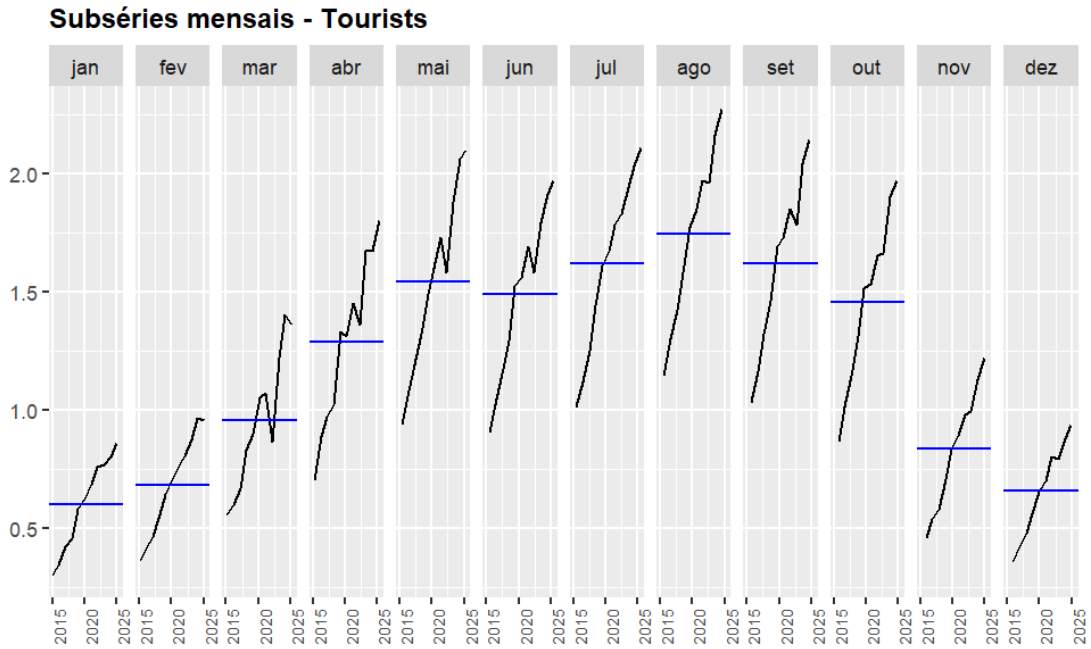


O gráfico abaixo revela uma sazonalidade anual bem definida, característica do setor turístico: os valores mais elevados ocorrem consistentemente entre junho e setembro, enquanto os mais baixos concentram-se entre dezembro e fevereiro.

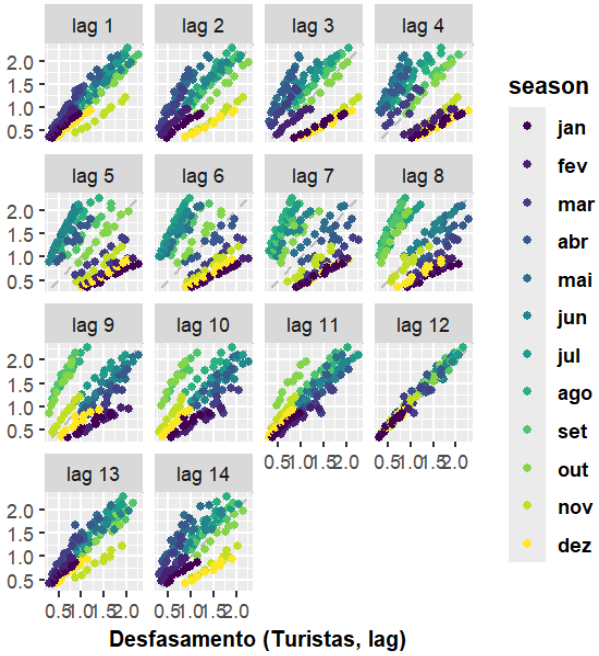
Para além da sazonalidade, observa-se também uma tendência crescente. Estas características são evidenciadas pelos segmentos mensais quase paralelos, que apenas se cruzam pontualmente, e pelas linhas anuais dispostas de forma ascendente, com os anos mais antigos na parte inferior e os mais recentes no topo.



Para complementar a análise, foi utilizado o gráfico de subséries mensais criado pela função *gg_subseries()*. O gráfico evidencia uma tendência global de crescimento no número de turistas, com valores mensais progressivamente mais elevados nos anos recentes. Além disso, nota-se a regularidade do padrão sazonal, em que os meses de verão apresentam mais turistas, enquanto os meses de inverno mantêm os valores mais baixos. Este comportamento confirma a forte sazonalidade e a tendência positiva já identificadas nas análises anteriores.



O gráfico de *lags* criado pela função *gg_lag()*, com desfasamentos de 1 a 14 períodos, permite avaliar a relação entre os valores atuais e os valores passados da série temporal. Observa-se que a linha correspondente ao *lag* 12 se apresenta quase reta, indicando uma forte correlação entre observações separadas por 12 meses, sendo este comportamento consistente com uma sazonalidade anual.

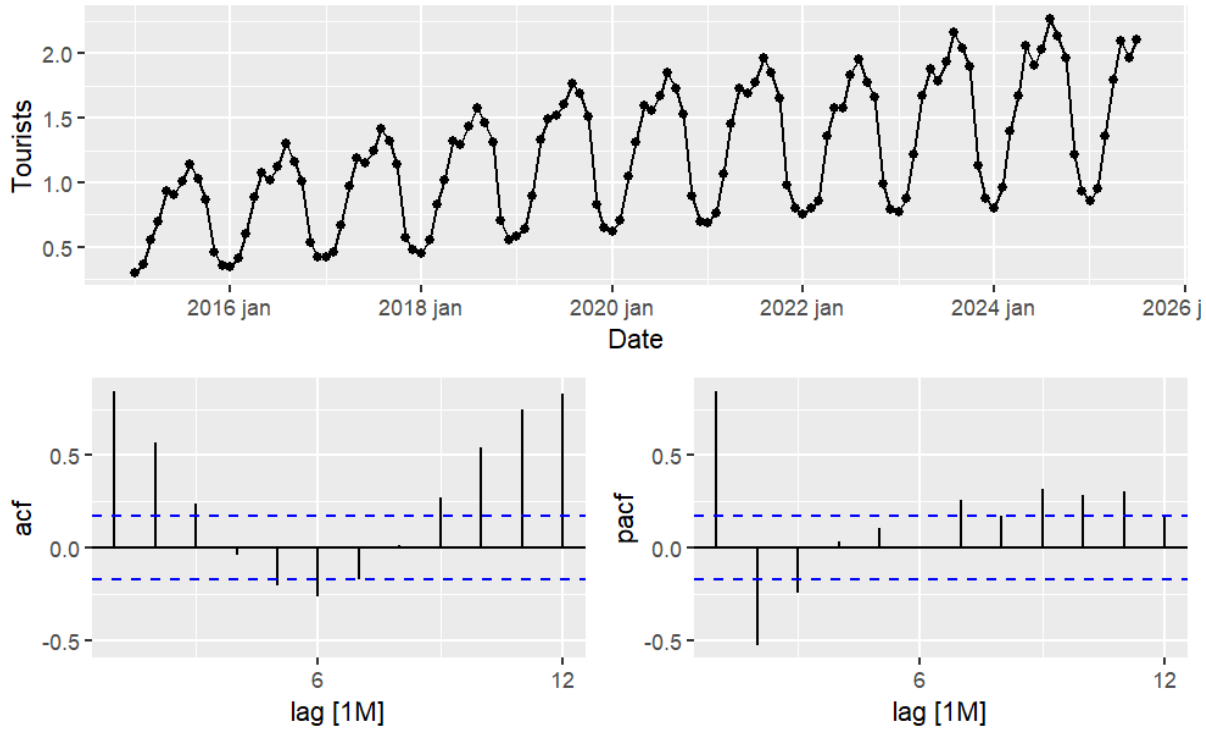


4. Análise de Correlações Temporais

A estrutura de dependência temporal foi estudada através das funções de autocorrelação e autocorrelação parcial, obtidas com *gg_tsdisplay().*

O correlograma da ACF mede a correlação entre os valores da série e os seus próprios valores passados (*lags*). O gráfico da ACF mostra picos significativos particularmente mais altos nos *lags* 1, 2, 11 e 12. Os picos nos *lags* 1 e 2 refletem dependência de curto prazo, típica de séries com memória recente. O pico no *lag* 12 evidencia uma sazonalidade anual clara.

O correlograma da PACF mede a correlação direta entre a série e seus valores passados (*lags*), removendo a influência dos *lags* intermédios. O gráfico da PACF mostra picos significativos no *lag* 1 e 2, indicando que os valores atuais da série dependem, principalmente, dos dois meses anteriores.



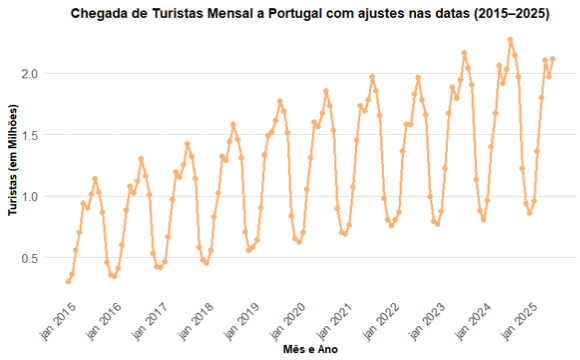
5. Teste de Estacionaridade

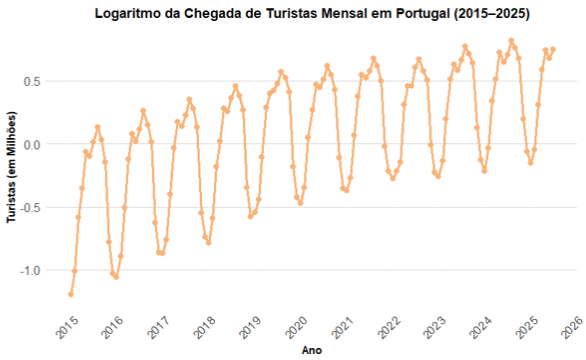
Para avaliar se a série era estacionária, aplicou-se o teste *KPSS*, revelando um valor de estatística de 1.32, com um *p-value* de 0.01, levando à rejeição da hipótese nula de estacionaridade. Conclui-se que a série não é estacionária, ou seja, tem raíz unitária, apresentando tendência e variância não constantes ao longo do tempo, sendo também visível graficamente.

Para determinar o número de diferenciações necessárias à estacionarização, foram aplicadas as funções *unitroot_ndiffs()* e *unitroot_nsdiffs().* Os resultados indicaram a necessidade de uma diferenciação ordinária (*d*=1) e uma diferenciação sazonal (*D*=1). Além disso, a medida de força sazonal (F_S) foi de 0.98, valor superior a 0.64, reforçando a necessidade de aplicar a diferenciação sazonal.

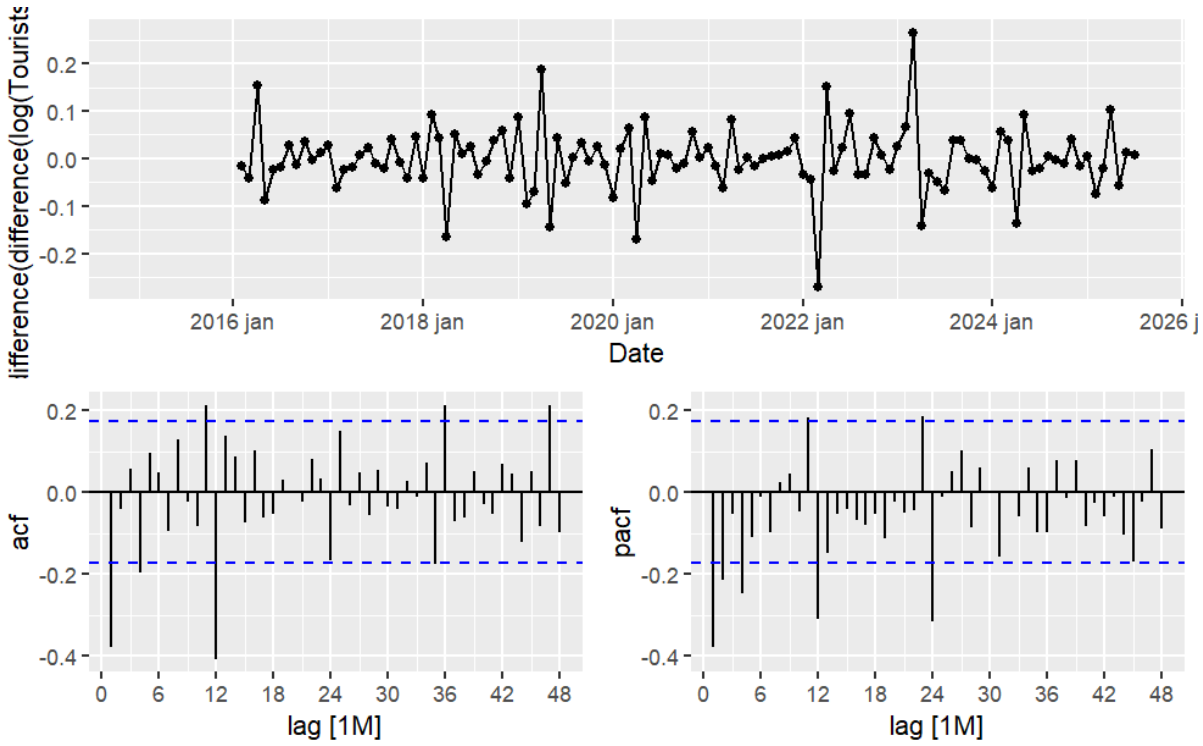
6. Estacionarização da Série Temporal

Para estabilizar a variância da série, foi aplicada a transformação logarítmica aos valores de turistas. O antes e depois da transformação logarítmica podem ser observados abaixo:





Em seguida, realizou-se primeiro a diferenciação sazonal (*D*=1) e posteriormente a diferenciação não sazonal/ordinária (*d*=1), de acordo com os resultados dos testes de diferenciações. Após estas operações, foi novamente realizado o teste *KPSS*, obtendo-se um valor de 0.1, que não rejeitou a hipótese nula da série ser estacionária. Finalmente, foram realizados novos gráficos da ACF e PACF para confirmar que a série estacionarizada apresenta um padrão mais adequado para modelação de séries temporais.



7. Modelação

Na fase de modelação foram propostos 4 modelos SARIMA(p,d,q)(P,D,Q)$_{12}$:

- SARIMA(4,1,0)(2,1,0)$_{12}$
- SARIMA(4,1,0)(0,1,1)$_{12}$
- SARIMA(0,1,4)(2,1,0)$_{12}$
- SARIMA(0,1,4)(0,1,1)$_{12}$

A escolha dos parâmetros baseou-se na análise dos gráficos de autocorrelação e autocorrelação parcial:

- p = 4: último pico significativo na PACF não sazonal;

- q = 4: último pico significativo na ACF não sazonal;

- P = 2: picos sazonais observados nos *lags* 12 e 24 da PACF;

- Q = 1: pico sazonal observado apenas no *lag* 12 da ACF.

Os modelos com zeros nas restantes componentes foram incluídos para avaliar separadamente o impacto de cada modificação e identificar a combinação mais adequada dos parâmetros.

Além destes modelos, foi ajustado um quinto modelo automático, que seleciona os parâmetros de forma otimizada com base na série. O modelo automático selecionado foi SARIMA(1,1,1)(2,1,0)$_{12}$, que reflete um equilíbrio entre simplicidade e capacidade explicativa. A análise automática identificou apenas um termo AR e um termo MA na componente não sazonal (p = 1, q = 1), uma vez que apenas os primeiros *lags* se apresentam como significativamente diferentes de zero. Na componente sazonal, foram mantidos dois termos AR (P = 2), consistentes com a dependência anual observada nos *lags* 12 e 24, enquanto o termo MA sazonal (Q = 0) foi descartado.

Na tabela abaixo apresentam-se os modelos realizados, ordenados por qualidade decrescente, com base nos critérios de informação AIC, AICc e BIC:

	AIC	AICc	BIC
auto	-340.0	-339.5	-326.3
arima014210	-338.8	-337.7	-319.6
arima014011	-337.6	-336.8	-321.2
arima410210	-336.8	-335.7	-317.6
arima410011	-334.2	-333.4	-317.8

O modelo auto foi considerado o com melhor qualidade, uma vez que apresenta os menores valores de AIC, AICc e BIC, refletindo melhor ajuste à série temporal.

O modelo escolhido aplicou uma diferenciação sazonal (*D*) à componente sazonal constituída pelos seguintes parâmetros:

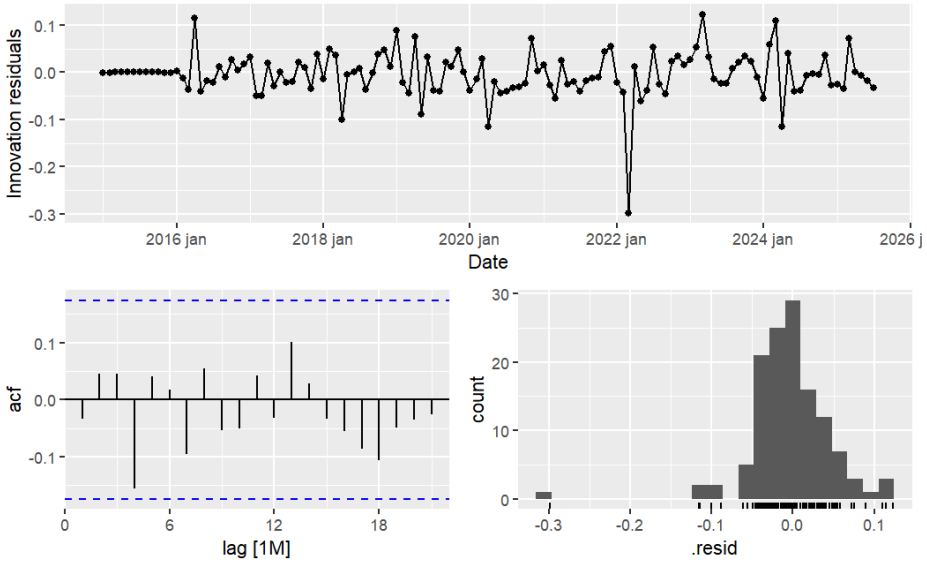
- AR(1) = - 0.5563
- AR(2) = - 0.3367

De seguida aplicou uma diferenciação (*d*) na componente ordinária composta pelos seguintes parâmetros:

- AR(1) = 0.4375
- MA(1) = - 0.8423

A análise dos resíduos do modelo escolhido foi realizada através da função *gg_tsresiduals()*, que cria três gráficos complementares:

- Série temporal dos resíduos – permite avaliar se os resíduos se comportam como ruído branco, isto é, se os valores estão distribuídos aleatoriamente em torno de zero, sem tendência ou padrão visível ao longo do tempo;
- ACF dos resíduos – avalia se existe autocorrelação remanescente. A ausência de picos significativos indica que o modelo capturou toda a estrutura temporal da série;
- Histograma e densidade dos resíduos – verifica se a distribuição dos erros é aproximadamente normal.



Analisando os gráficos dos resíduos, observa-se um comportamento aleatório, sem autocorrelação significativa e com distribuição aproximadamente normal, sugerindo um bom ajustamento do modelo. O teste de *Ljung–Box* (*lag* = 24; *dof* = 4) apresentou *p-value* = 0.77, não rejeitando a hipótese nula de ausência de autocorrelação.

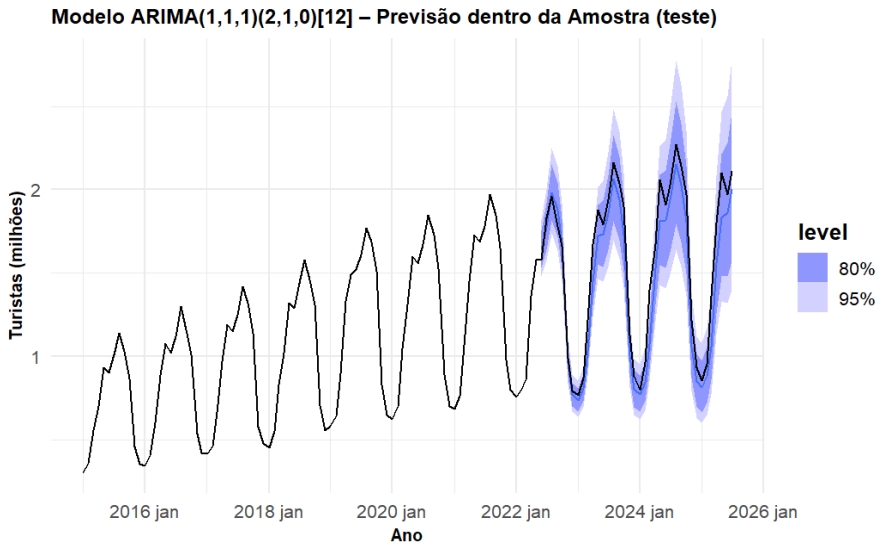
Assim, conclui-se que o modelo SARIMA$(1,1,1)(2,1,0)_{12}$ é estatisticamente adequado e os resíduos comportam-se como ruído branco.

8. Previsão

a. Dentro da amostra

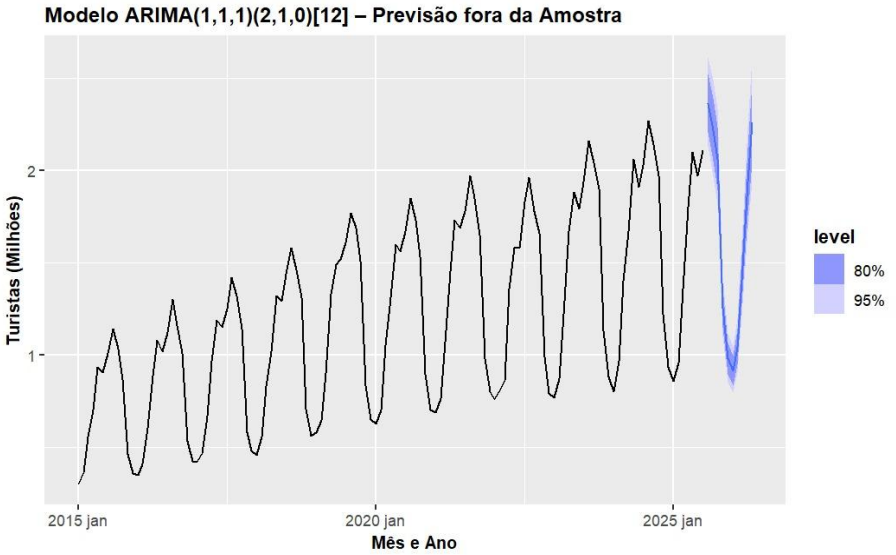
Para avaliar o desempenho do modelo auto, os dados foram divididos em 70% treino e 30% teste. O modelo apresentou bom desempenho preditivo, indicando uma raíz de erro quadrático médio (RMSE) de 142 mil turistas em relação aos valores reais e apresenta um erro percentual médio absoluto (MAPE) de 7.73%, ou seja, em média, as previsões do modelo diferem dos valores reais em aproximadamente 7.73%. Ambas as métricas refletem uma boa precisão preditiva do modelo face à variabilidade da série temporal.

As previsões acompanharam bem a tendência e sazonalidade da série, confirmando a adequação do modelo. O gráfico da previsão é apresentado abaixo.



b. Fora da amostra

Para a previsão fora da amostra, o modelo foi utilizado para estimar os 11 meses seguintes, até junho de 2026. As projeções mantêm o padrão sazonal anual observado, com variações regulares entre os períodos de maior e menor afluência turística.



Na tabela abaixo, a análise das taxas de variação permitiu avaliar a evolução da variável ao longo do período estudado. No primeiro semestre de 2025 (2025-S1), observou-se um aumento homólogo de 2.68% face ao mesmo semestre de 2024. Já para o primeiro semestre de 2026, correspondente ao período previsto, verificou-se

um crescimento homólogo mais expressivo, de 7.87% relativamente ao primeiro semestre de 2025.

Ano-Semestre	Turistas (em milhões)	Taxa de crescimento homóloga
2024-S2	10,565	5,12%
2025-S1	9,043	2,68%
2025-S2	10,985	3,98%
2026-S1	9,755	7,87%

Considerando a variação anual abaixo, estima-se um acréscimo global de 3.39% em 2025 face ao ano anterior, tendo em conta que o segundo semestre desse ano inclui valores projetados.

Ano	Turistas (em milhões)	Taxa de crescimento anual
2023	18,253	14,36%
2024	19,372	6,13%
2025	20,028	3,39%

9. Discussão e conclusão

O modelo apresentou um bom desempenho, tanto dentro como fora da amostra, reproduzindo de forma consistente a tendência e sazonalidade observadas na série. No entanto, as previsões podem apresentar limitações, uma vez que o modelo não considera outras variáveis que podem afetar a chegada de turistas, como fatores económicos, condições climáticas ou eventos imprevistos. Uma possível melhoria seria a utilização de um conjunto de dados mais completo, com variáveis explicativas adicionais (como PIB, preços médios de alojamento, eventos sazonais ou condições meteorológicas), o que permitiria capturar melhor os fatores externos que influenciam a chegada de turistas e, consequentemente, aumentar a precisão das previsões.