Gestão e Qualidade de Dados Mestre no SAP IBP: Sustentação do Planeamento da Procura

FRANCISCO RIBEIRO SEDAS
(Licenciado)

Estágio de Natureza Profissional para obtenção do grau de Mestre em Matemática Aplicada para a Indústria, na Área de Especialização de Tratamento de Dados

Orientadores:

Doutora Ana Alexandra Antunes Figueiredo Martins
Doutora Iola Maria Silvério Pinto
André Henriqueto Ricardo
Diogo Chambel Reis

Júri:

Presidente: Doutor Luís Manuel Ferreira da Silva
Vogais:

Doutora Paula Cristina Pires Simões
Doutora Ana Alexandra Antunes Figueiredo Martins

Dezembro de 2025

Gestão e Qualidade de Dados Mestre no SAP IBP: Sustentação do Planeamento da Procura

FRANCISCO RIBEIRO SEDAS
(Licenciado)

Estágio de Natureza Profissional para obtenção do grau de Mestre em Matemática Aplicada para a Indústria, na Área de Especialização de Tratamento de Dados

Orientadores:
Doutora Ana Alexandra Antunes Figueiredo Martins, ISEL
Doutora Iola Maria Silvério Pinto, ISEL
André Henriqueto Ricardo, Grupo Nabeiro
Diogo Chambel Reis, Grupo Nabeiro

Júri:
Presidente: Doutor Luís Manuel Ferreira da Silva, ISEL
Vogais:
Doutora Paula Cristina Pires Simões, ISEL
Doutora Ana Alexandra Antunes Figueiredo Martins, ISEL

Dezembro de 2025

Agradecimentos

Após a conclusão deste estágio curricular, muitas são as pessoas a quem devo uma palavra de agradecimento.

Em primeiro lugar, gostaria de agradecer aos meus orientadores externos. Ao André e ao Diogo agradeço por tudo o que me ensinaram, por todas as reuniões onde fui aprendendo mais sobre o universo Delta, mais concretamente, o trabalho desenvolvido no seu departamento, e por toda a disponibilidade em ouvir as minhas dúvidas e me aconselhar e orientar na direção certa.

Uma palavra de apreço também para a Equipa de Planeamento e Dados, mais concretamente ao Rui Pereira, por me ter proporcionado esta experiência que, mais do que uma oportunidade de aprendizagem, foi sem dúvida um primeiro passo importante na minha vida profissional.

Dirijo-me, agora, aos meus orientadores internos no ISEL. À Professora Ana Martins e à Professora Iola Pinto agradeço pelo apoio constante, por toda a orientação dada e, finalmente, por desde o primeiro dia terem acreditado em mim e no meu trabalho.

Deixo um agradecimento, também, a todos os meus colegas que fizeram parte da minha caminhada no ISEL. Por todos os debates sobre este tema que me incentivaram a melhorar o meu projeto, e que espero terem sido uma mais valia para os seus respectivos projetos, o meu sincero obrigado.

Finalmente, como não podia deixar de ser, aos meus pais, à minha família e amigos. Agradeço pelo apoio incondicional, pelas palavras de encorajamento que foram fundamentais nos momentos mais difíceis e pela paciência que tiveram para me ouvir falar do meu trabalho mesmo não tendo tantos conhecimentos na área.

A todos os que aqui referi, obrigado.

Declaração de Integridade

Declaro que este relatório de estágio é o resultado da minha investigação pessoal e independente. O seu conteúdo é original e todas as fontes listadas nas referências bibliográficas foram consultadas e estão devidamente mencionadas no texto. Mais declaro que todas as referências científicas e técnicas relevantes para o desenvolvimento do trabalho estão devidamente citadas e constam das referências bibliográficas.

O autor:

Lisboa, 15 de dezembro de 2025

Resumo

Este relatório descreve o projeto desenvolvido no âmbito do estágio curricular do Mestrado em Matemática Aplicada para a Indústria, centrado na melhoria do planeamento da procura através de metodologias quantitativas. O trabalho enquadra-se na necessidade de reforçar a fundamentação estatística associada à seleção de variáveis explicativas e à construção de modelos preditivos aplicados a séries temporais.

A investigação desenvolvida incidiu sobre a análise de dados históricos de vendas de vários produtos pertencentes a quatro famílias distintas, tendo como objetivo principal identificar variáveis preditoras relevantes e avaliar o impacto da sua inclusão em modelos de previsão. Para tal, foi implementada uma abordagem metodológica estruturada que integrou análise descritiva, três métodos complementares de seleção de variáveis, incluindo regressões lineares e a regressão Lasso, e a comparação entre diferentes modelos de previsão.

O modelo principal estudado foi o SARIMAX (*Seasonal AutoRegressive Integrated Moving Average with eXogenous regressors*), permitindo incorporar variáveis exógenas na previsão da procura. As previsões obtidas foram comparadas com as provenientes de um modelo de Alisamento Exponencial, recorrendo à análise sistemática de métricas de erro e à verificação dos pressupostos através da avaliação dos resíduos.

Os resultados evidenciaram que a integração de variáveis exógenas, aliada a um processo rigoroso de seleção de preditores, melhorou de forma consistente a precisão das previsões. O estudo reforça, assim, a importância de abordagens multivariadas e estatisticamente sustentadas na otimização do planeamento da procura e na gestão eficiente de cadeias de abastecimento.

Palavras-chave: seleção de preditores, SARIMAX, planeamento da procura, modelos de previsão.

Abstract

This report describes the project developed as part of the curricular internship of the Master's in Applied Mathematics for Industry, focused on improving demand planning through quantitative methodologies. The work addresses the need to strengthen the statistical foundations supporting the selection of explanatory variables and the development of forecasting models applied to time series.

The investigation focused on the analysis of historical sales data for several products from four distinct families, with the main objective of identifying relevant predictors and assessing the impact of their inclusion in forecasting models. To achieve this, a structured methodological approach was implemented, integrating descriptive analysis, three complementary variable selection techniques, including linear regressions and the Lasso regression, and the comparison of multiple forecasting models.

The main forecasting model studied was SARIMAX (Seasonal AutoRegressive Integrated Moving Average with eXogenous regressors), which enables the incorporation of exogenous variables into demand forecasting. Its results were compared with those obtained from an Exponential Smoothing model, using systematic error metric evaluation and residual diagnostics to assess model adequacy.

The results showed that integrating exogenous variables, combined with a rigorous predictor selection process, consistently improved forecasting accuracy. The study therefore highlights the importance of multivariate and statistically grounded approaches in optimizing demand planning and enhancing the efficiency of supply chain management.

Keywords: predictor selection, SARIMAX, demand planning, forecasting models.

Índice

Lista de Figuras

Lista de Tabelas

Lista de Símbolos

n	Nº de Observações
min	Mínimo
max	Máximo
R	Amplitude Total
$Q_{0.25}$	Primeiro Quartil
med	Mediana
$Q_{0.75}$	Terceiro Quartil
IQR	Amplitude Inter-Quartil
$\overline{x}$	Média
s^2	Variância
s	Desvio Padrão
CVR	Coeficiente de Variação Resistente
g_B	Coeficiente de *Bowley*
g_K	Coeficiente de Achatamento
T_t	Tendência
S_t	Sazonalidade
R_t	Componente Aleatória
r_s	Coeficiente de correlação de *Spearman*

Lista de Abreviaturas

ADF	*Augmented Dickey–Fuller Test*
AIC	*Akaike Information Criterion*
GAM	*Generalized Additive Model*
GLM	*Generalized Linear Model*
IBP	*Integrated Business Planning*
MAE	*Mean Absolute Error*
MAPE	*Mean Absolute Percentage Error*
MSE	*Mean Squared Error*
RMSE	*Root Mean Squared Error*
SARIMAX	*Seasonal AutoRegressive Integrated Moving Average with eXogenous variables*
SMAPE	*Symmetric Mean Absolute Percentage Error*
STL	*Seasonal and Trend decomposition using Loess*
WMAPE	*Weighted Mean Absolute Percentage Error*

Capítulo 1

Introdução

Esta dissertação foi elaborada no âmbito do estágio curricular do Mestrado em Matemática Aplicada para a Indústria, resultante de uma parceria entre o Instituto Superior de Engenharia de Lisboa e o Grupo Nabeiro, e realizado entre 1 de fevereiro e 15 de julho de 2025.

O Grupo Nabeiro é uma das empresas mais emblemáticas de Portugal, líder de mercado no setor do café. Fundado em 1961 por Rui Nabeiro, na vila de Campo Maior, no Alentejo, o grupo consolidou-se como um exemplo de inovação, sustentabilidade e compromisso social no panorama empresarial português. [1]

Com uma forte presença nacional e internacional, o Grupo Nabeiro opera em mais de 40 países, mantendo um portefólio diversificado que abrange, além do café, setores como o comércio, serviços, tecnologia, vinhos e energias renováveis. A marca Delta Cafés constitui o núcleo do grupo, sendo sinónimo de qualidade e tradição no consumo de café em Portugal. [2]

A aposta em tecnologia e inovação é um dos pilares estratégicos do grupo, visível na atuação da Direção de Sistemas de Informação, da qual a equipa de Planeamento e Dados faz parte. Esta equipa tem como missão apoiar a tomada de decisão através da análise e previsão de dados, bem como da gestão dos ciclos de vida dos produtos.

Durante o estágio curricular, o trabalho desenvolvido incidiu sobretudo na sustentação do planeamento da procura. O objetivo principal consistiu em apoiar a tomada de decisão neste domínio, mediante a análise e modelação de dados históricos relativos às quantidades vendidas de diversos produtos.

1.1 Planeamento da Procura

O planeamento da procura constitui uma componente estratégica essencial na gestão das cadeias de abastecimento, assegurando que a disponibilidade de produtos acompanha, de forma eficiente, as flutuações da procura do mercado. Para além de prever volumes de con-

sumo, este processo visa compreender os fatores que os influenciam, permitindo às empresas antecipar variações e alinhar os seus recursos produtivos, logísticos e comerciais.[3]

A crescente complexidade dos mercados e a volatilidade dos padrões de consumo têm vindo a exigir que o planeamento da procura evolua de abordagens empíricas para metodologias quantitativas rigorosas, suportadas por modelos estatísticos e analíticos. No entanto, a eficácia destas metodologias depende fortemente da seleção das variáveis explicativas que alimentam os modelos de previsão. Quando essa escolha é feita de forma subjetiva ou pouco sistematizada, a qualidade preditiva tende a degradar-se, comprometendo a fiabilidade das decisões subsequentes.

No caso do Grupo Nabeiro, esta limitação torna-se particularmente relevante devido à diversidade de produtos e à influência de múltiplos fatores internos e externos na procura. Embora o sistema *SAP Integrated Business Planning* (SAP IBP) disponibilize ferramentas avançadas de previsão e análise [4], a forma como as variáveis independentes são atualmente definidas baseia-se, em larga medida, em conhecimento empírico e experiência dos analistas. Assim, mais do que um problema de tecnologia, coloca-se um desafio de qualidade e fundamentação estatística das variáveis utilizadas no planeamento.

É neste contexto que se enquadra o presente trabalho, desenvolvido em colaboração com a equipa de Planeamento e Dados do Grupo Nabeiro, e cujo propósito é contribuir para uma abordagem mais estruturada e quantitativamente validada da previsão da procura.

1.2 Enquadramento do Projeto

O trabalho desenvolvido neste estágio curricular surgiu precisamente da necessidade de reforçar a fundamentação estatística do planeamento da procura no Grupo Nabeiro. O desafio central consistiu em reduzir a subjetividade associada à escolha das variáveis explicativas utilizadas nos modelos de previsão, substituindo critérios empíricos por métodos quantitativos e reprodutíveis.

Neste contexto, o projeto teve como principal objetivo avaliar, selecionar e validar variáveis preditoras relevantes para a estimativa da procura, bem como testar modelos de previsão alternativos capazes de melhorar a precisão face às abordagens atualmente empregues no SAP IBP. Assim, a dissertação assume uma dupla vertente: por um lado, metodológica — ao propor um processo estruturado de seleção e avaliação de variáveis — e, por outro, aplicada — ao analisar o impacto dessas variáveis em produtos reais do portefólio da empresa.

Para garantir a comparabilidade dos resultados, o SAP IBP foi utilizado exclusivamente como fonte de extração de dados históricos e previsões de referência, permitindo quantificar de forma objetiva as melhorias obtidas com o modelo desenvolvido.

A implementação de uma pipeline automatizada, associada a técnicas de validação cruzada temporal e à integração de variáveis exógenas, permitiu construir um sistema de previsão mais robusto e transparente. Entre os modelos estudados, destacou-se o SARIMAX, pela sua capacidade de incorporar variáveis externas e pela compatibilidade direta com as metodologias implementadas no SAP IBP, o que possibilitou uma comparação rigorosa entre abordagens.

O projeto enquadra-se, assim, numa perspetiva de melhoria contínua da qualidade dos dados e dos modelos de previsão, contribuindo para a sustentação do planeamento da procura e para uma gestão mais informada e eficiente da cadeia de abastecimento do Grupo Nabeiro.

Com o intuito de especificar e descrever todo o processo desenvolvido, este relatório encontra-se organizado em cinco capítulos.

O Capítulo 1, Introdução apresenta o enquadramento do trabalho, os objetivos do estágio, uma breve caracterização do Grupo Nabeiro e a relevância do planeamento da procura no contexto industrial. O Capítulo 2 inclui a revisão de trabalhos anteriores sobre o tema central deste projeto e evidencia a inovação introduzida no estágio realizado. O Capítulo 3 descreve as metodologias aplicadas na análise e tratamento dos dados. No Capítulo 4, Caso de Estudo, essas metodologias são aplicadas aos dados, apresentando-se os resultados obtidos para um dos produtos analisados. O Capítulo 5, Resultados e Conclusões, compara as previsões obtidas para o produto em estudo com as previsões realizadas pela equipa de Planeamento e Dados, apresentando as principais conclusões do projeto e sugestões de trabalho futuro. Por fim, a Bibliografia reúne os documentos e fontes consultadas, por ordem de citação, e os Anexos apresentam os resultados obtidos para os restantes produtos analisados.

Capítulo 2

Estado da Arte

A previsão de séries temporais constitui uma das áreas mais consolidadas da estatística aplicada e da ciência de dados, tendo adquirido crescente relevância em contextos empresariais e industriais, onde a precisão das estimativas influencia diretamente a eficiência operacional e a sustentabilidade económica. No setor alimentar, esta importância é ainda mais acentuada, devido à forte sazonalidade da procura e à variabilidade introduzida por fatores externos.

Com o intuito de contextualizar o presente estudo, foi realizada uma revisão de trabalhos académicos e de aplicações práticas relacionadas com a previsão em séries temporais. Foram analisadas dissertações e artigos que abordam a previsão da procura em contextos industriais, com especial enfoque na indústria alimentar, considerando diferentes metodologias e critérios de validação.

Este capítulo apresenta os contributos mais relevantes identificados na literatura, destacando as abordagens metodológicas, os contextos de aplicação e as métricas utilizadas na avaliação dos modelos preditivos. Esta revisão não só fundamenta teoricamente as opções metodológicas adotadas neste projeto, como também identifica oportunidades de inovação e melhoria, evidenciando semelhanças e diferenças entre os estudos analisados e a investigação aqui desenvolvida.

2.1 Revisão de Trabalhos Anteriores

O primeiro artigo académico revisto [5] analisa a previsão da procura no setor alimentar, recorrendo a modelos ARIMA e a técnicas de combinação de previsões. Os autores iniciam com uma decomposição das séries temporais, de forma a identificar padrões de sazonalidade e tendência, aplicando posteriormente modelos ARIMA diferenciados para distintas tipologias de produtos. A performance dos modelos é avaliada com recurso a métricas clássicas, como o MAE, o MAPE e o RMSE, comparando os resultados obtidos por modelo e por produto. Destaca-se a tentativa de combinar previsões provenientes de diferentes modelos, com o intuito de aumentar a precisão global. Apesar de constituir um contributo válido para a aplicação de modelos estatísticos a dados empresariais reais, o estudo não contempla a inclusão de

variáveis exógenas nem a utilização de técnicas automáticas de seleção de preditores.

O segundo artigo [6] aborda a previsão da procura para produtos com ciclo de vida curto, um desafio particular em setores como a grande distribuição ou o vestuário. A metodologia proposta combina técnicas de *Data Mining*, nomeadamente árvores de decisão, com regressão penalizada, visando a seleção automática das variáveis mais relevantes e a subsequente estimação de modelos de previsão. Embora considere séries temporais, o enfoque é sobretudo multivariado, privilegiando abordagens de *Machine Learning* em detrimento de modelos clássicos, como o ARIMA ou o SARIMA. As métricas de validação utilizadas incluem o RMSE, o MAPE e o SMAPE. Este estudo é especialmente relevante por evidenciar os ganhos obtidos através da integração de variáveis externas e da automatização da seleção de preditores, antecipando algumas das metodologias exploradas no presente projeto, ainda que com recurso a técnicas distintas.

O terceiro artigo [7] incide sobre a previsão da procura na indústria alimentar, com o objetivo de apoiar o planeamento da produção de uma empresa do setor. Após uma caracterização inicial da procura, incluindo a identificação de padrões sazonais e irregulares, o autor aplica três métodos baseados no alisamento exponencial: o alisamento exponencial simples, o modelo de *Holt* e o modelo de *Holt-Winters*, complementados pela decomposição clássica. A validação é realizada com base em erros históricos, calculados a partir de dados de 2018, utilizando o MAE e o MAPE como métricas principais. Esta dissertação evidencia uma abordagem prática e operacional, adequada a contextos empresariais, mas limitada à utilização de séries univariadas. A ausência de variáveis exógenas e de modelos SARIMA/SARIMAX reduz a sua capacidade preditiva em cenários influenciados por fatores externos.

2.2 Motivação e Comparação com o Trabalho Desenvolvido

O presente projeto constitui uma evolução e um aprofundamento metodológico face aos trabalhos anteriormente revistos, propondo a aplicação de um modelo SARIMAX que, para além de captar a estrutura temporal e sazonal das séries, integra variáveis externas previamente selecionadas através de abordagens rigorosas. Esta integração não só aumenta a precisão das previsões, como também permite interpretar o impacto de fatores contextuais sobre a variável dependente.

Em comparação com os trabalhos de Juliana, Manuel e Ana [5], baseados em modelos ARIMA, e com o estudo de Sebastião Norton [7], centrado em métodos de alisamento exponencial, o projeto aqui desenvolvido distingue-se pela adoção do modelo SARIMAX, pela utilização de critérios de informação como o AIC para seleção de modelos e pela avaliação sistemática dos pressupostos através da análise de resíduos e de testes de autocorrelação. Acresce a aplicação de um conjunto mais diversificado de métricas de erro, incluindo o WMAPE

e o SMAPE, que permitem avaliar de forma mais equilibrada a qualidade preditiva, sobretudo em séries com amplitudes variáveis.

Relativamente ao trabalho de Rie Gaku [6], observam-se semelhanças na preocupação com a seleção de variáveis explicativas e no recurso a algoritmos automáticos. Contudo, enquanto Gaku privilegia técnicas de *Data Mining* e *Machine Learning*, o presente projeto opta por métodos estatísticos clássicos (GLM, regressão Lasso), compatíveis com a interpretação estatística tradicional e mais alinhados com o perfil académico da matemática aplicada. Adicionalmente, introduz um processo estruturado de seleção de preditores, organizado em três abordagens complementares, inexistente nos estudos analisados.

O projeto diferencia-se ainda pela aplicação de métodos robustos de validação cruzada e pela verificação dos pressupostos de estacionariedade dos resíduos, assegurando a validade estatística dos resultados. Por fim, a utilização de três abordagens distintas para a seleção de preditores, seguida da definição do conjunto final como a união das variáveis selecionadas, potencia a capacidade preditiva do modelo e mitiga o risco de omissão de fatores relevantes.

Ao conjugar métodos estatísticos clássicos com técnicas modernas de seleção e validação, este projeto procura colmatar limitações metodológicas identificadas nos estudos anteriores, oferecendo uma solução mais abrangente, flexível e estatisticamente robusta para a previsão de séries temporais multivariadas.

2.3 Súmula de Artigos

A revisão bibliográfica evidenciou a diversidade de metodologias aplicadas à previsão de séries temporais, em particular no setor alimentar, bem como a predominância de modelos clássicos, como o ARIMA, ou de métodos de alisamento exponencial. Constatou-se, no entanto, que a inclusão de variáveis exógenas em modelos preditivos permanece pouco explorada nos estudos analisados, assim como a aplicação sistemática de algoritmos de seleção de preditores.

O presente projeto, ao integrar estas componentes num modelo SARIMAX, validado através de múltiplas métricas de erro e suportado por três abordagens complementares de seleção de variáveis, procura colmatar essas lacunas e contribuir para o desenvolvimento de metodologias mais completas e adaptáveis a contextos empresariais reais.

Capítulo 3

Metodologia

Ao longo deste projeto foram aplicadas diversas metodologias estatísticas e de previsão de séries temporais, com o objetivo de desenvolver modelos preditivos robustos e fiáveis para suporte ao planeamento da procura. Este capítulo descreve, de forma detalhada, todas as etapas seguidas, desde a exploração inicial dos dados até à validação final dos modelos.

Na Secção 3.1, introduz-se a *Pipeline* desenvolvida para sistematizar e automatizar todas as operações necessárias à construção e validação dos modelos de previsão. Esta *pipeline* integra uma técnica de validação cruzada específica para séries temporais, assegurando o respeito pela dependência temporal dos dados e permitindo uma avaliação rigorosa da capacidade preditiva dos modelos.

Na Secção 3.2, é apresentada a Análise Descritiva realizada sobre o conjunto de dados, onde são descritas as diferentes tipologias de variáveis consideradas e as medidas descritivas utilizadas para caraterizar as suas propriedades estatísticas fundamentais.

Segue-se, na Secção 3.3, a Caracterização de uma Série Temporal, onde são definidas as principais componentes que compõem este tipo de dados (tendência, sazonalidade e componente aleatória) e onde se descrevem os métodos exploratórios e de decomposição utilizados para a sua análise.

A Secção 3.4 apresenta os Modelos Univariados de previsão utilizados como referência inicial, nomeadamente as variantes de Alisamento Exponencial: alisamento exponencial simples, modelo de *Holt* e modelo de *Holt-Winters*, adequadas a diferentes padrões de comportamento temporal.

Na Secção 3.5 descrevem-se as três abordagens distintas implementadas para a Seleção de Variáveis Preditoras, combinando critérios estatísticos, métodos não paramétricos e algoritmos automáticos de regularização e seleção.

A Secção 3.6 foca-se nos Modelos de Previsão propriamente ditos, descrevendo o ajustamento e a extensão dos modelos SARIMA e SARIMAX, com especial destaque para a

incorporação de variáveis exógenas selecionadas previamente.

Por fim, na Secção 3.7, é apresentado o processo de Validação dos Modelos de Previsão, onde se aplicam testes estatísticos de análise de resíduos e métricas de erro, de modo a garantir a adequação e fiabilidade dos modelos desenvolvidos.

3.1 *Pipeline*

Uma *pipeline* estatística pode ser definida como uma sequência organizada de procedimentos que estruturam, de forma sistemática e automatizada, as etapas de pré-processamento dos dados, seleção de variáveis, ajustamento de modelos e validação. Cada fase depende das anteriores e contribui cumulativamente para a construção do modelo final, garantindo consistência metodológica ao longo de todo o processo [8].

No contexto da previsão de séries temporais, a implementação de uma *pipeline* estatística assegura a execução coerente e reprodutível de todas as operações necessárias, desde a preparação dos dados até à avaliação final do desempenho dos modelos. Ao integrar de forma sequencial as tarefas de tratamento, seleção, modelação e validação, esta estrutura facilita não apenas a automatização do processo, mas também a sua interpretação e a verificação rigorosa da qualidade dos resultados obtidos.

Em problemas de previsão de séries temporais, a construção de *pipelines* é particularmente importante, dado que os dados apresentam dependência temporal e requerem operações sequenciais específicas que devem respeitar a ordem cronológica das observações.

O esboço da *pipeline* desenvolvida neste projeto, descreve-se na Figura 3.1.

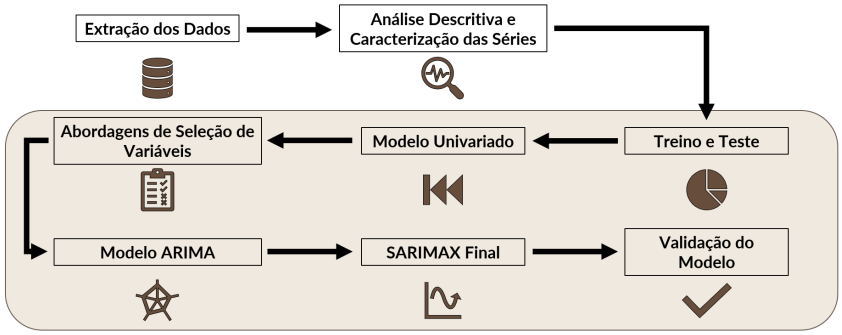

Figura 3.1: Esquema ilustrativo da *pipeline* utilizada neste projeto.

3.1.1 Etapas da *Pipeline*

Revendo a Figura 3.1, identificam-se todas as fases da *pipeline* desenvolvida neste projeto.

As fases de extração de dados, análise descritiva (Secção 3.2) e caracterização das séries (Secção 3.3) apenas serão executadas uma vez.

Após esse processo, procede-se à obtenção dos conjuntos de treino e teste para cada iteração da validação cruzada de acordo com o exposto na Secção 3.1.2. Assim em cada iteração são processadas as fases do modelo univariado (Secção 3.4), seleção de variáveis (Secção 3.5), modelo SARIMA e SARIMAX (Secção 3.6) e a análise de resíduos (Secção 3.7).

3.1.2 Integração com a Técnica de Validação Cruzada

Neste trabalho, a *pipeline* foi intrinsecamente associada a uma técnica de validação cruzada específica para séries temporais, de forma a garantir que os dados de treino e teste respeitam a ordem cronológica e evitam fugas de informação. Para tal, recorreu-se a uma estratégia de validação cruzada com janela deslizante, também designada por *Time Series Cross-Validation*. [9]

A lógica desta técnica consiste em definir, para cada iteração, um conjunto de treino inicial, ao qual se vão adicionando progressivamente novas observações, mantendo-se o tamanho do conjunto de teste fixo. Deste modo, simula-se a evolução do sistema no tempo e garante-se que, em cada iteração, os dados futuros não contaminam a modelação baseada em dados passados.

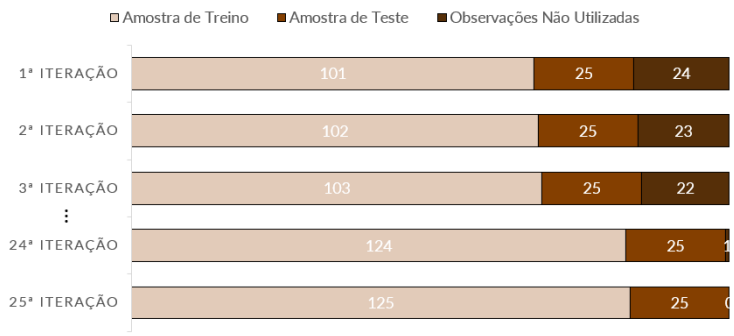

Figura 3.2: Esquema ilustrativo da validação cruzada utilizada neste projeto.

Como é possível verificar na Figura 3.2, a técnica de validação cruzada implementada nesta *pipeline* foi configurada da seguinte forma:

- **Primeira Iteração:** 101 observações para treino e 25 observações para teste.
- **Iterações Seguintes:** a cada nova iteração, adiciona-se uma observação ao conjunto de treino, mantendo as mesmas 25 observações de teste seguintes no tempo.
- **Última Iteração:** 125 observações para treino e 25 observações para teste.

Este processo perfaz um total de 25 iterações, garantindo uma avaliação robusta da capacidade preditiva dos modelos em diferentes períodos temporais.

3.2 Análise Descritiva

De modo a resumir a informação contida num conjunto de dados e a examinar os dados antes da aplicação de outras técnicas estatísticas, procede-se à sua Análise Descritiva. Deste modo consegue-se um entendimento básico dos dados e do tipo de relações existentes entre as variáveis analisadas. Esta abordagem permite ainda identificar padrões, desvios ou anomalias presentes na amostra. Contudo será primeiro importante fazer a distinção entre os vários tipos de dados estudados ao longo do projeto.

3.2.1 Tipos de Dados

Este projeto centra-se na análise e previsão de séries temporais. É, por isso, necessário, em primeiro lugar, definir o que é uma série temporal e, depois, que tipos de valor a série pode assumir.

3.2.1.1 Série Temporal

Uma série temporal é definida como uma sequência de observações de uma ou mais variáveis, registadas ao longo do tempo e, geralmente, em intervalos regulares. A particularidade destas sequências reside no facto das suas observações não serem independentes entre si, uma vez que valores registados num determinado momento podem influenciar os valores subsequentes. [10]

Analiticamente, uma série temporal pode ser representada do seguinte modo:

$$y_1, y_2, \ldots, y_t, \ldots \tag{3.1}$$

onde y_t é o valor observado no instante t.

3.2.1.2 Variáveis Contínuas

As variáveis contínuas são aquelas que podem assumir qualquer valor dentro de um intervalo contínuo, possibilitando a realização de operações aritméticas e análises estatísticas como o cálculo de médias, desvios padrão, tendências e auto-correlações. Em séries temporais, variáveis contínuas são frequentemente objeto de modelação e previsão, dada a sua

variabilidade ao longo do tempo.[9]

3.2.1.3 Variáveis Categóricas

As variáveis categóricas referem-se a atributos qualitativos que classificam as observações em categorias mutuamente exclusivas. Podem subdividir-se em nominais, quando não existe qualquer ordem implícita, e ordinais, quando existe uma hierarquia ou ordenação. [11]

3.2.1.4 Variáveis Binárias

As variáveis binárias constituem um caso particular das variáveis categóricas, assumindo apenas dois valores possíveis, geralmente representados como 0 e 1. No âmbito das séries temporais, estas variáveis são frequentemente incluídas como indicadores para captar o efeito de eventos pontuais ou estruturais que possam interferir no comportamento da variável dependente. [12]

3.2.2 Medidas Descritivas

A obtenção de determinadas medidas descritivas, calculadas a partir dos dados, auxilia neste processo estatístico. As medidas utilizadas no decorrer deste estágio curricular podem ser divididas em: [13]

- Medidas de Localização: este conjunto de medidas localiza o centro da amostra. Estas podem ser:
  - de tendência central, como é o caso da média aritmética, da mediana e da moda.
  - de tendência não central, como são exemplo os extremos e o primeiro e terceiro quartis (Q_p).
- Medidas de Dispersão: este conjunto de medidas mede a variabilidade dos dados. Estas podem ser:
  - de dispersão absoluta. Por exemplo, o desvio padrão, a variância e a amplitude inter-quartis (IQR).
  - de dispersão relativa, das quais foi utilizado o coeficiente de variação. Para variáveis onde não existiam *outliers*, este é dado por $CV = \frac{s}{\overline{x}} \times 100$, onde s é o desvio padrão e $\overline{x}$ a média. Para variáveis com observações classificadas como *outliers*, deve utilizar-se como medida de variabilidade relativa o coeficiente de variação resistente definido por $CVR = \frac{IQR}{Q_{0.50}} \times 100$. Se $CV \leq 15\%$ os dados apresentam uma variabilidade fraca, se $15\% < CV \leq 30\%$ os dados apresentam uma variabilidade média e se $CV > 30\%$ os dados apresentam uma variabilidade elevada. [13]

- Medidas de Assimetria: este conjunto de medidas determina o tipo de assimetria dos dados. O Coeficiente baseado nos momentos foi utilizado para variáveis que não apresentavam *outliers*. Este calcula-se através da seguinte expressão:

$$g_1 = \frac{m_3}{s^3} \tag{3.2}$$

onde m_k é o momento de ordem k em relação à média.

O Coeficiente de *Bowley* (g_B) foi utilizado para variáveis onde existiam *outliers*. Este calcula-se através da seguinte expressão:

$$g_B = \frac{(Q_{0.75} - Q_{0.50}) - (Q_{0.50} - Q_{0.25})}{IQR} \tag{3.3}$$

  - Se $g_i = 0$, a distribuição é simétrica.
  - Se $g_i > 0$, a distribuição é assimétrica positiva, ou seja, existe uma maior concentração dos dados nos valores mais baixos da variável em estudo.
  - Se $g_i < 0$, a distribuição é assimétrica negativa, ou seja, existe uma maior concentração dos dados nos valores mais altos da variável em estudo.
  - i = 1, B.

A Figura 3.3 visa ilustrar graficamente estes três tipos de distribuição e a sua caracterização quanto à assimetria.

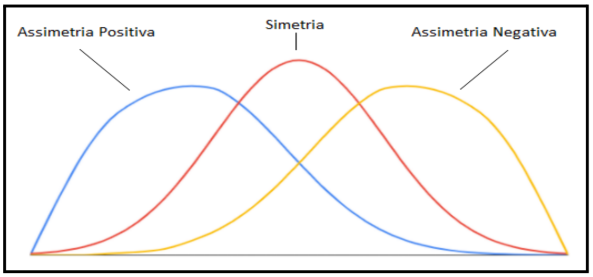

Figura 3.3: Tipos de assimetria que a distribuição dos dados pode assumir.

- Medidas de Achatamento: este conjunto de medidas determina o tipo de achatamento dos dados. O Coeficiente baseado nos momentos foi utilizado para variáveis que não apresentavam *outliers*. Este calcula-se através da seguinte expressão:

$$g_2 = \frac{m_4}{s^4} - 3 \tag{3.4}$$

  - Se $g_2 > 0$, a distribuição tem um achatamento fraco, obtendo-se, graficamente, uma curva leptocúrtica.
  - Se $g_2 = 0$, a distribuição tem um achatamento médio, obtendo-se, graficamente, uma curva mesocúrtica.
  - Se $g_2 < 0$, a distribuição tem um achatamento elevado, obtendo-se, graficamente, uma curva platicúrtica.

O Coeficiente de Achatamento (g_K) foi utilizado para variáveis onde existiam *outliers*. Este calcula-se através da seguinte expressão:

$$g_K = \frac{IQR}{2(Q_{0.9} - Q_{0.1})} \tag{3.5}$$

- Se $g_K < 0.263$, a distribuição tem um achatamento fraco, obtendo-se, graficamente, uma curva leptocúrtica.
- Se $g_K = 0.263$, a distribuição tem um achatamento médio, obtendo-se, graficamente, uma curva mesocúrtica.
- Se $g_K > 0.263$, a distribuição tem um achatamento elevado, obtendo-se, graficamente, uma curva platicúrtica.

É de salientar que todas estas medidas apenas podem ser calculadas para variáveis contínuas. As variáveis categóricas foram caracterizadas pelas tabelas de frequências absolutas e frequências relativas.

3.2.2.1 *Outliers*

Consideram-se *outliers* todas as observações que se afastam significativamente do padrão geral da distribuição dos dados. Estes valores podem resultar de erros de registo, eventos excecionais ou alterações estruturais no comportamento da variável. A sua identificação é fundamental, uma vez que podem distorcer medidas estatísticas clássicas — como média, desvio padrão e coeficientes baseados nos momentos — e, consequentemente, influenciar negativamente os processos de seleção de variáveis e modelação.

Neste estudo, a identificação dos *outliers* foi realizada através do critério baseado nos quartis, amplamente utilizado por métodos de estatística descritiva robusta. Este critério define *outliers* como valores que se situam fora dos limites:

$$x < Q_{0.25} - 1.5 \times IQR \text{ ou } x > Q_{0.75} + 1.5 \times IQR \tag{3.6}$$

Contudo, uma análise mais detalhada permite distinguir entre dois graus de afastamento.

Os *outliers* leves são observações que excedem moderadamente os limites definidos, situando-se entre:

$$Q_{0.25} - 3 \times IQR < x < Q_{0.25} - 1.5 \times IQR$$

ou

$$Q_{0.75} + 1.5 \times IQR < x < Q_{0.75} + 3 \times IQR$$

Os *outliers* severos são valores que se afastam de forma extrema da estrutura da distribuição, localizando-se para além de:

$$x < Q_{0.25} - 3 \times IQR \quad \text{ou} \quad x > Q_{0.75} + 3 \times IQR$$

Estes casos geralmente refletem eventos muito atípicos, anomalias evidentes ou alterações abruptas no sistema, sendo particularmente relevantes para a análise, uma vez que podem influenciar decisivamente a estimação de parâmetros, a deteção de sazonalidades e a qualidade das previsões.

A distinção entre *outliers* leves e severos permitiu orientar a escolha das medidas descritivas utilizadas (coeficientes baseados nos momentos ou coeficientes resistentes), assegurando uma caracterização mais adequada da distribuição e uma modelação estatisticamente robusta.

3.3 Caracterização de uma Série Temporal

A análise de séries temporais baseia-se na compreensão da sua estrutura interna, sendo fundamental identificar as diferentes componentes que contribuem para a formação dos dados observados ao longo do tempo. [10]

3.3.1 Componentes de uma Série Temporal

- **Tendência (T_t):** Refere-se ao movimento de longo prazo na série, refletindo uma evolução global do sentido de crescimento, ou decrescimento.

- **Sazonalidade (S_t):** Componente que reflete variações periódicas e sistemáticas associadas a fatores sazonais, que se repetem a intervalos regulares. O período de variação, ou ciclo sazonal, é composto por s períodos de tempo designados por estações.

- **Componente Aleatória ou Ruído (R_t):** Engloba as variações não explicadas pelas componentes anteriores, correspondendo a flutuações imprevisíveis e sem padrão sistemático.

3.3.2 Cronograma

Antes de aplicar qualquer técnica estatística ou de modelação, é recomendada a análise exploratória da série temporal através do seu cronograma, o gráfico dos valores observados ao longo do tempo. Como é possível verificar na Figura 3.4, esta representação visual permite identificar padrões de tendência, sazonalidade, presença de valores atípicos e possíveis mu-

danças estruturais na série. [10]

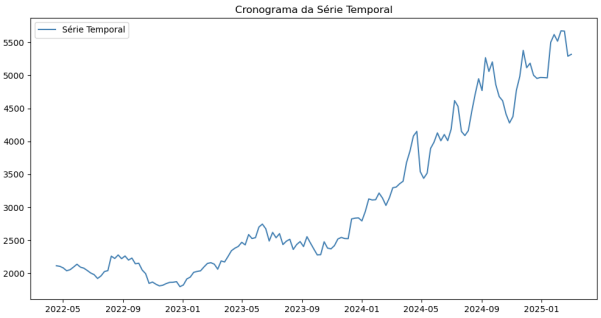

Figura 3.4: Exemplo de um cronograma de uma das variáveis estudadas no trabalho.

3.3.3 Decomposição *Seasonal and Trend decomposition using Loess*

A decomposição de séries temporais permite separar explicitamente as componentes que explicam o seu comportamento: tendência, sazonalidade e irregularidade. Para tal, recorrem-se a métodos de alisamento que reduzem a variabilidade aleatória e evidenciam padrões estruturais.

De forma geral, uma decomposição aditiva pode ser representada por:

$$X_t = f(T_t, S_t, R_t) \tag{3.7}$$

onde T_t representa a tendência, S_t a componente sazonal e R_t o resíduo ou componente irregular.

Neste projeto, foi adotado o método STL — *Seasonal and Trend decomposition using Loess* [14], um procedimento robusto e flexível que combina regressão local com uma iteração estruturada para estimar as componentes determinísticas.

A decomposição STL caracteriza-se por três propriedades fundamentais:

1. Utiliza *Loess* como técnica de alisamento não paramétrico, permitindo capturar relações não lineares e adaptar-se localmente ao comportamento da série.

2. Permite sazonalidade variável, ao contrário de métodos clássicos que impõem sazonalidade estritamente constante ao longo do tempo.

3. É robusta a outliers, através de um esquema iterativo que ajusta a influência de observações discrepantes.

O algoritmo STL aplica *Loess* para extrair a tendência através do alisamento da série original. Em paralelo, ajusta *Loess* em janelas específicas para estimar a sazonalidade, permitindo que esta evolua ao longo do tempo. Finalmente, obtém a componente irregular como diferença

entre a série original e as estimativas conjuntas das componentes tendência e sazonalidade.

Este método foi utilizado na etapa de análise exploratória das séries temporais, permitindo identificar padrões estruturais relevantes e apoiar decisões subsequentes relacionadas com a modelação.

Um exemplo da aplicação da decomposição STL neste estudo é apresentado na Figura 3.5.

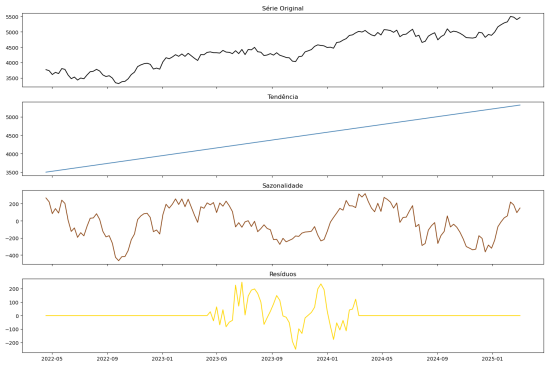

Figura 3.5: Exemplo de uma decomposição STL de uma das variáveis estudadas no trabalho.

3.4 Modelo Univariado

Em modelos para séries temporais designa-se por modelo univariado aquele que permite obter previsões usando apenas a informação histórica de uma variável temporal. A principal vantagem destes modelos reside na sua simplicidade conceptual e facilidade de implementação, sendo frequentemente utilizados como modelos de referência ou quando não se dispõe de variáveis explicativas adicionais.

Entre os modelos univariados mais utilizados na previsão de séries temporais encontram-se os métodos de Alisamento Exponencial. Existem três variantes principais deste método, aplicáveis consoante as características da série temporal. [15]

3.4.1 Alisamento Exponencial Simples

O método de alisamento exponencial simples é adequado para séries temporais que não apresentam tendência nem sazonalidade. A previsão para o próximo período baseia-se numa média ponderada entre o valor observado mais recente e a previsão anterior, com a seguinte expressão:

$$\hat{y}_{T+1|T} = \alpha y_T + \alpha(1-\alpha)y_{T-1} + \alpha(1-\alpha)^2 y_{T-2} + \alpha(1-\alpha)^3 y_{T-3} + \cdots \tag{3.8}$$

onde $0 \leq \alpha \leq 1$ é a constante de amortecimento e quanto maior o valor desta constante

maior o peso relativo atribuído às observações mais recentes.

3.4.2 Modelo de *Holt*

Proposto por *Holt* em 1957, este método expande o alisamento exponencial simples ao incorporar uma componente de tendência linear na previsão. Assim, permite modelar séries temporais com tendência mas sem sazonalidade. O método recorre a duas equações de atualização: uma para o nível e outra para a tendência.

Equação de atualização para o nível:

$$l_t = \alpha y_t + (1-\alpha)(l_{t-1} + \phi b_{t-1}), \tag{3.9}$$

com $0 < \alpha < 1$.
Equação de atualização para a tendência:

$$b_t = \beta^*(l_t - l_{t-1}) + (1-\beta^*)\phi b_{t-1}, \tag{3.10}$$

com $0 < \beta^* < 1$.
O que se traduz na seguinte equação para a previsão:

$$\hat{y}_{t+h|t} = l_t + (\phi + \phi^2 + \cdots + \phi^h) b_t \tag{3.11}$$

3.4.3 Modelo de *Holt-Winters*

O método de *Holt-Winters* estende o modelo de *Holt* para séries temporais que apresentam simultaneamente tendência e sazonalidade. Dependendo da forma como estas componentes se combinam, o método pode assumir versões aditivas ou multiplicativas. A escolha entre versões aditivas ou multiplicativas depende da natureza da sazonalidade: utiliza-se a forma aditiva quando as variações sazonais têm amplitude aproximadamente constante ao longo do tempo, e a forma multiplicativa quando a amplitude da sazonalidade cresce ou diminui proporcionalmente ao nível da série.

A Figura 3.6 sintetiza as possíveis combinações entre os tipos de tendência (nula ou aditiva) e os tipos de sazonalidade (aditiva ou multiplicativa). Cada combinação conduz a um conjunto distinto de equações de alisamento, tendência e sazonalidade, que generalizam o modelo base de Holt, acrescentando-lhe uma componente sazonal ajustada à forma observada na série temporal.

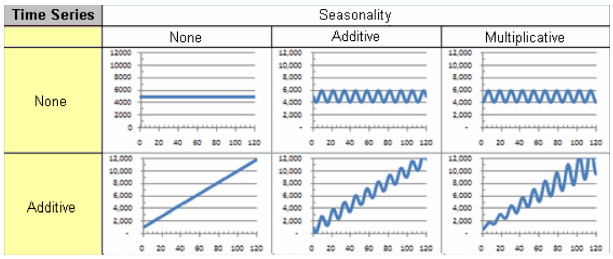

Figura 3.6: Tipos de tendência e sazonalidade.
Fonte: *Zoho Analytics – Working of Forecasting*

De seguida, apresentam-se as quatro formulações possíveis do modelo de *Holt-Winters*, correspondentes às combinações ilustradas na Figura 3.6. Em cada caso, são explicitadas as equações de atualização das componentes do modelo e a respetiva expressão para a previsão.

3.4.3.1 Tendência Nula, Sazonalidade Aditiva

Neste caso, a equação de alisamento é dada por:

$$l_t = \alpha(y_t - s_{t-m}) + (1-\alpha)l_{t-1}, \tag{3.12}$$

com $0 < \alpha < 1$.
A equação de sazonalidade é dada por:

$$s_t = \gamma(y_t - l_{t-1}) + (1-\gamma)s_{t-m} \tag{3.13}$$

com $0 < \gamma < 1$.
E a previsão é calculada através de:

$$\hat{y}_{t+h|t} = l_t + s_{t+h-m(k+1)} \tag{3.14}$$

com $k = \left\lfloor \frac{h-1}{m} \right\rfloor$.

3.4.3.2 Tendência Aditiva, Sazonalidade Aditiva

Neste caso, a equação de alisamento é dada por:

$$l_t = \alpha(y_t - s_{t-m}) + (1-\alpha)(l_{t-1} + b_{t-1}), \tag{3.15}$$

com $0 < \alpha < 1$.
A equação de tendência é dada por:

$$b_t = \beta^*(l_t - l_{t-1}) + (1-\beta^*)b_{t-1}, \tag{3.16}$$

com $0 < \beta^* < 1$.

A equação de sazonalidade é dada por:

$$s_t = \gamma(y_t - l_{t-1} - b_{t-1}) + (1-\gamma)s_{t-m}, \tag{3.17}$$

com $0 < \gamma < 1$.
E a previsão é calculada através de:

$$\hat{y}_{t+h|t} = l_t + hb_t + s_{t+h-m(k+1)} \tag{3.18}$$

com $k = \left\lfloor \frac{h-1}{m} \right\rfloor$.

3.4.3.3 Tendência Nula, Sazonalidade Multiplicativa

Neste caso, a equação de alisamento é dada por:

$$l_t = \alpha(y_t/s_{t-m}) + (1-\alpha)l_{t-1}, \tag{3.19}$$

com $0 < \alpha < 1$.
A equação de sazonalidade é dada por:

$$s_t = \gamma(y_t/l_{t-1}) + (1-\gamma)s_{t-m}, \tag{3.20}$$

com $0 < \gamma < 1$.
E a previsão é calculada através de:

$$\hat{y}_{t+h|t} = l_t s_{t+h-m(k+1)} \tag{3.21}$$

3.4.3.4 Tendência Aditiva, Sazonalidade Multiplicativa

Neste caso, a equação de alisamento é dada por:

$$l_t = \alpha(y_t/s_{t-m}) + (1-\alpha)(l_{t-1} + b_{t-1}), \tag{3.22}$$

com $0 < \alpha < 1$.
A equação de tendência é dada por:

$$b_t = \beta^*(l_t - l_{t-1}) + (1-\beta^*)b_{t-1}, \tag{3.23}$$

com $0 < \beta^* < 1$.
A equação de sazonalidade é dada por:

$$s_t = \gamma(y_t/(l_{t-1} + b_{t-1})) + (1-\gamma)s_{t-m}, \tag{3.24}$$

com $0 < \gamma < 1$.

E a previsão é calculada através de:

$$\hat{y}_{t+h|t} = (l_t + hb_t)s_{t+h-m(k+1)} \tag{3.25}$$

3.5 Algoritmos de Seleção de Variáveis

A seleção adequada de variáveis preditoras é um passo determinante no desenvolvimento de modelos de previsão robustos e interpretáveis, sobretudo em contextos de séries temporais onde existe uma variável dependente e uma ou mais variáveis independentes. Uma má especificação do conjunto de preditores pode não apenas comprometer a capacidade preditiva do modelo, como também induzir problemas de colinearidade, sobreajustamento e interpretações estatísticas enviesadas [16]. Assim, neste projeto foram consideradas três abordagens distintas para a seleção de variáveis, combinando critérios estatísticos, análise gráfica e procedimentos automáticos de seleção de modelos.

Em todas as abordagens, procedeu-se previamente à transformação da variável resposta utilizando a transformação de *Box-Cox* [17], com o objetivo de estabilizar a variância e aproximar a distribuição dos resíduos à normalidade, condição essencial para a validade de vários métodos estatísticos subsequentes. A transformação de *Box-Cox* é definida como:

$$Y_t^{(\lambda)} = \begin{cases} \frac{Y_t^{\lambda}-1}{\lambda}, & \text{se } \lambda \neq 0 \\ ln(Y_t), & \text{se } \lambda = 0 \end{cases} \tag{3.26}$$

onde λ é o parâmetro de transformação estimado pelo método da máxima verosimilhança.

Um exemplo desta transformação pode ser consultado na Figura 3.7:

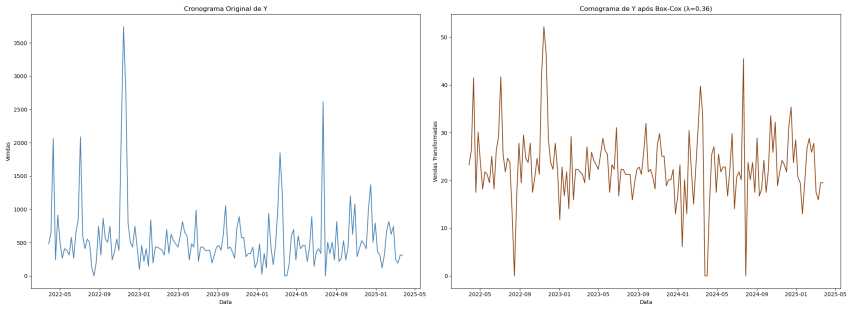

Figura 3.7: Exemplo de uma transformação *Box-Cox*.

Esta transformação implica que todas as previsões efetuadas estarão, também elas, transformadas. Deste modo é importante ter o cuidado de, após fazer qualquer previsão, inverter a sua transformação através de:

$$\hat{Y}_t = \begin{cases} e^{\hat{Y}_t^{(\lambda)}} \left[1 + \hat{\sigma_h^2}\right], & \text{se } \lambda = 0 \\ (\lambda \hat{Y}_t^{(\lambda)} + 1)^{1/\lambda} \left[1 + \frac{\hat{\sigma_h^2}(1-\lambda)}{2((\lambda \hat{Y}_t^{(\lambda)}+1)^2)}\right], & \text{se } \lambda \neq 0 \end{cases} \tag{3.27}$$

3.5.1 Primeira Abordagem

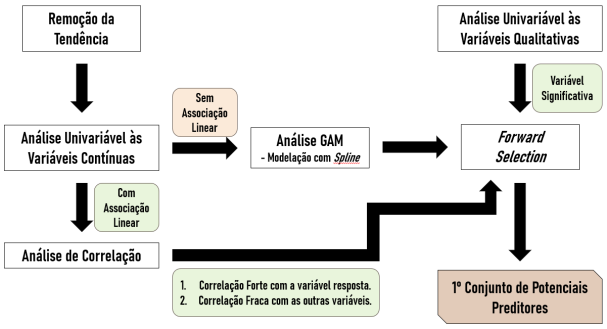

Figura 3.8: Esboço do processo de obtenção de potenciais preditores através da 1ª abordagem de seleção de variáveis.

Etapa 1: Na primeira abordagem, iniciou-se com a remoção da tendência das variáveis explicativas que o justificaram, através da decomposição STL já definida na Secção 3.3.3.

Etapa 2: Posteriormente, realizou-se uma análise univariável para avaliar o efeito individual de cada variável sobre a variável resposta transformada através de *Box-Cox*.

Para essa análise recorreu-se aos Modelos Lineares Generalizados (GLM), definidos por *Nelder* e *Wedderburn* [18] como uma extensão dos modelos lineares clássicos, permitindo modelar variáveis resposta com distribuições pertencentes à família exponencial. Nestes modelos, o valor esperado da variável resposta $\mu_i = \mathbb{E}(Y_i)$ é relacionado linearmente com os preditores por meio de uma função de ligação $g(\cdot)$:

$$g(\mu_i) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} \tag{3.28}$$

onde:

- $i = 1, 2, ..., n$, com n o número total de observações;
- p o número total de variáveis explicativas.

A estimativa dos parâmetros β_j é realizada pelo método da máxima verosimilhança. No presente projeto, considerou-se a função de ligação identidade, uma vez que a variável resposta, após transformação de Box-Cox, apresentou distribuição aproximadamente simétrica e contínua. Com base nos resultados dos modelos lineares generalizados univariáveis, foram selecionadas para a fase seguinte as variáveis para as quais se rejeitou a hipótese de nulidade do respetivo coeficiente de regressão, ou seja, aquelas a que correspondeu um *p-value* inferior ao nível de significância de 0.05.

Etapa 3: A forma funcional entre cada variável explicativa que passou a esta fase e a variável dependente foi analisada através da observação do gráfico da função parcial obtido pelo ajustamento de um modelo aditivo generalizado (GAM) univariável.

Os GAM [19] são uma extensão muito poderosa dos GLM. Enquanto os GLM modelam a relação entre a variável resposta e as variáveis explicativas através de uma combinação linear dos preditores, os GAM relaxam essa suposição linear, permitindo que a relação seja não paramétrica e não necessariamente linear. Nestes, cada preditor X_{ij} tem associado uma função $f_j(\cdot)$, que é uma função suave estimada diretamente dos dados:

$$g(\mu_i) = \beta_0 + f_1(X_{i1}) + \cdots + f_p(X_{ip}) \tag{3.29}$$

Ou seja, a contribuição de cada variável explicativa para a resposta não está restringida a ser linear, mas pode ter formas curvas, ondulações, ou outros formatos complexos.

Os suavizadores $f_j(\cdot)$ são estimados através de métodos flexíveis como *splines*, que são funções feitas por polinómios unidos de forma suave. O ajuste do modelo envolve equilibrar a flexibilidade com a suavidade e, geralmente, é feito usando técnicas iterativas, como *backfitting*.

Os gráficos da função parcial representam graficamente a relação marginal entre cada variável explicativa e a variável resposta, mantendo as restantes variáveis constantes. É exemplo, o gráfico presente na Figura 3.9.

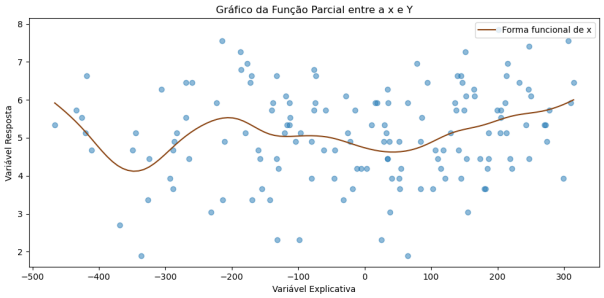

Figura 3.9: Exemplo de um gráfico de uma função parcial estudada neste trabalho.

Etapa 4: No caso das variáveis contínuas que evidenciaram uma associação linear com a variável resposta, analisou-se a matriz de correlação de *Spearman*. Para a seleção, consideraram-se variáveis cuja correlação com a variável resposta fosse superior a 0.3 e cuja correlação mútua com outras variáveis explicativas fosse inferior a 0.3. O valor de 0.3 foi adotado como limiar por ser frequentemente utilizado na literatura como indicador de uma correlação moderada, suficiente para justificar a relevância preditiva sem, contudo, incorrer em problemas de multicolinearidade. O coeficiente de correlação de *Spearman*, que é uma medida não paramétrica de associação baseada nos postos das observações, e que avalia a relação monotónica entre duas variáveis [20], é calculado como:

$$r_s = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \tag{3.30}$$

onde d_i é a diferença entre os postos das observações em cada variável e n o número total de observações.

Etapa 5: Para as variáveis qualitativas, realizaram-se testes não paramétricos de comparação entre os grupos existentes nestas variáveis: [21]

- O teste de *Mann-Whitney* é utilizado para comparar a distribuição de uma variável contínua entre dois grupos independentes, assumindo como hipótese nula a igualdade das distribuições;

- O teste de *Kruskal-Wallis*, uma generalização do anterior para mais de dois grupos, que avalia se existe diferença estatisticamente significativa nas tendências centrais das amostras independentes.

Estes testes não requerem o pressuposto de normalidade e são apropriados para amostras pequenas ou dados assimétricos. Foram consideradas estatisticamente significativas as variáveis cujo *p-value* fosse inferior a 0.05. Significa isto que nestas variáveis existem comportamentos significativamente distintos nos valores da variável resposta para observações pertencentes a grupos diferentes.

Etapa 6: O conjunto de variáveis selecionadas nestas etapas foi posteriormente submetido a um processo de seleção automática utilizando o algoritmo de *Forward Selection*. Este algoritmo de seleção começa com um modelo onde apenas está presente o termo independente e gradualmente adiciona as variáveis independentes, uma a uma, verificando se a qualidade do modelo aumenta. Para isso, avalia-se o AIC, um indicador que é descrito mais à frente na Secção 3.7.1. Caso isso não aconteça, essa variável não é adicionada e segue-se para a próxima iteração.

Este processo culminou na obtenção de um primeiro conjunto de preditores candidatos.

3.5.2 Segunda Abordagem

Etapa 1: Na segunda abordagem, procedeu-se primeiramente à estacionarização de todas as variáveis explicativas, recorrendo ao teste de *Dickey-Fuller* aumentado (ADF). Este teste estatístico é utilizado para verificar a presença de uma raiz unitária numa série temporal, ou seja, para testar a hipótese nula de que a série é não estacionária contra a hipótese alternativa de estacionariedade [22].

A forma geral da regressão estimada no teste ADF é:

$$\Delta Y_t = \alpha + \beta t + \gamma Y_{t-1} + \sum_{i=1}^{p} \delta_i \Delta Y_{t-i} + \epsilon_t \tag{3.31}$$

onde:

- ΔY_t representa a primeira diferença da série temporal;
- α é uma constante;
- β_t representa uma tendência determinística;
- γ é o coeficiente associado ao valor desfasado da série;
- $\sum_{i=1}^{p} \delta_i \Delta Y_{t-i}$ corresponde aos termos de correção para autocorrelação serial nos resíduos;
- ϵ_t é o erro aleatório.

As hipóteses do teste são:

$$\begin{cases} H_0 : \gamma = 0 & \text{existe raiz unitária, logo a série é não estacionária} \\ H_1 : \gamma < 0 & \text{não existe raiz unitária, logo a série é estacionária} \end{cases} \tag{3.32}$$

A rejeição de H_0 ocorre se o *p-value* for inferior a 0.05, indicando que a série é estacionária.

No caso de não rejeição da hipótese nula procede-se à diferenciação da série. Isto é, para cada instante t, o valor z_t da série diferenciada será dado pela diferença entre os valores nesse instante e no seguinte da série original, ou seja:

$$z_t = y_{t+1} - y_t \tag{3.33}$$

Este processo tem como objetivo verificar a presença de raiz unitária na série. Caso esta seja detetada, a série é diferenciada uma vez ou, se necessário, duas vezes, o que permite remover a tendência e torná-la estacionária. A sazonalidade, por sua vez, também pode ser eliminada através de diferenciação, embora este procedimento não decorra diretamente do teste da raiz unitária, mas sim de uma análise complementar da estrutura da série.

As etapas **2**, **3**, **4**, **5** e **6** ocorrem como descritas na primeira abordagem, obtendo-se um segundo conjunto de preditores candidatos.

3.5.3 Terceira Abordagem

Etapa 1: A terceira abordagem começa do mesmo modo que a primeira, com a remoção da tendência nas variáveis que o justifiquem.

Etapa 2: De seguida é aplicada uma técnica automática de seleção de variáveis baseada na regressão Lasso (*Least Absolute Shrinkage and Selection Operator*) [23]. Esta técnica consiste na penalização da soma dos valores absolutos dos coeficientes, promovendo a eliminação de variáveis com contributo reduzido:

$$\min_{\beta_0,\beta} \Big\{ \frac{1}{2n} \sum_{i=1}^{n} (Y_i - \beta_0 - X_i^T \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \Big\} \tag{3.34}$$

onde λ é o parâmetro de regularização ajustado através de um processo de otimização, onde é escolhido o valor do parâmetro que minimiza o erro quadrático médio calculado por validação cruzada. Este procedimento permite realizar, em simultâneo, regularização e seleção de variáveis, obtendo-se um terceiro conjunto de preditores.

O conjunto final de variáveis a integrar o modelo SARIMAX foi definido como a união dos preditores identificados nas três abordagens. Cada abordagem contribuiu de forma distinta para o processo de seleção:

- **Primeira Abordagem:** permitiu avaliar estatisticamente a relevância individual de cada variável explicativa e identificar relações lineares ou não lineares com a variável resposta. A sua principal vantagem reside na interpretabilidade dos resultados e na capacidade de detetar associações específicas. Contudo, pode falhar na exclusão de variáveis colineares ou redundantes.

- **Segunda Abordagem:** garantiu que apenas variáveis estacionárias fossem utilizadas, prevenindo a incorreta especificação do modelo. A vantagem desta abordagem é a robustez estatística na preparação dos dados, mas apresenta como limitação a possibilidade de perda de informação quando a diferenciação é excessiva.

- **Terceira Abordagem:** aplicou um método automático de regularização, adequado a situações de elevada dimensão, reduzindo o risco de sobreajustamento. A grande vantagem é a eficiência computacional e a objetividade da seleção, mas, em contrapartida, existe o risco de eliminar variáveis potencialmente relevantes devido à penalização aplicada aos coeficientes.

A combinação das três abordagens permitiu reunir os pontos fortes de cada uma: a interpretabilidade da análise estatística tradicional, a robustez da verificação de estacionariedade e a eficiência dos métodos automáticos. Assim, assegurou-se uma seleção abrangente, equilibrada e devidamente fundamentada de preditores.

3.6 Modelos de Previsão

A previsão de séries temporais com dependências complexas no tempo exige modelos que consigam capturar padrões de autocorrelação, sazonalidade e tendências estruturais. Um dos modelos mais robustos e amplamente utilizados neste contexto é o modelo ARIMA (*AutoRegressive Integrated Moving Average*), desenvolvido por *Box* e *Jenkins* [10]. Este modelo foi posteriormente generalizado para incorporar padrões sazonais, dando origem ao modelo SARIMA (*Seasonal AutoRegressive Integrated Moving Average*). No presente projeto, o modelo SARIMA foi a base para a construção do modelo final SARIMAX (*Seasonal AutoRegressive Integrated Moving Average with eXogenous regressors*), que integra ainda variáveis exógenas

como fatores explicativos adicionais.

3.6.0.1 Modelo SARIMA

O modelo SARIMA é uma extensão do modelo ARIMA clássico, permitindo modelar séries que apresentam padrões sazonais bem definidos. Formalmente, um modelo SARIMA é representado pela notação:

$$SARIMA(p,d,q)(P,D,Q)_s \tag{3.35}$$

onde:

- p, d, q são, respetivamente, a ordem da parte autorregressiva (AR), a ordem da diferenciação não sazonal e a ordem da média móvel (MA);
- P, D, Q são os equivalentes para a componente sazonal;
- s é a periodicidade sazonal da série.

Já a expressão geral para o modelo SARIMA é dada por:

$$\Phi_P(B^s)\phi_P(B)(1-B)^d(1-B^s)^D Y_t = \Theta_Q(B^s)\theta_q(B)\varepsilon_t \tag{3.36}$$

onde:

- B é o operador do desfasamento $(B^k Y_t = Y_{t-k})$;
- $\phi_p(B)$ e $\theta_q(B)$ são os polinómios de ordem p e q para os termos AR e MA não sazonais, respetivamente, definidos por:

$$\phi_p(B) = 1 - \phi_1(B) - \phi_2(B^2) - \cdots \phi_p(B^p)$$
$$\theta_q(B) = 1 + \theta_1(B) + \theta_2(B^2) + \cdots \theta_q(B^q)$$

- $\Phi_P(B^s)$ e $\Theta_Q(B^s)$ são os polinómios de ordem P e Q para os termos AR e MA sazonais, respetivamente, definidos por:

$$\Phi_P(B^s) = 1 - \Phi_1(B^s) - \Phi_2(B^{2s}) - \cdots \Phi_P(B^{Ps})$$
$$\Theta_Q(B) = 1 + \Theta_1(B^s) + \Theta_2(B^{2s}) + \cdots \Theta_q(B^{Qs})$$

- ε é o ruído branco no instante t, assumido com média zero e variância constante.

O ajuste de um modelo SARIMA permite capturar de forma eficiente a dependência temporal e a sazonalidade da série em estudo, proporcionando previsões de curto e médio prazo de elevada precisão, desde que os pressupostos de estacionariedade e normalidade dos resíduos sejam devidamente assegurados. [11]

3.6.0.2 Modelo SARIMAX

Apesar da eficácia dos modelos SARIMA para séries univariadas, muitas séries temporais são influenciadas por variáveis externas que podem melhorar substancialmente a capacidade preditiva do modelo. Neste sentido, surge o modelo SARIMAX, que integra variáveis exógenas como fatores explicativos adicionais.

O modelo SARIMAX estende a formulação do SARIMA ao incluir um termo de regressão com variáveis exógenas X_t, tendo a seguinte expressão:

$$\Phi_P(B^s)\phi_P(B)(1-B)^d(1-B^s)^D Y_t = \Theta_Q(B^s)\theta_q(B)\varepsilon_t + \beta X_t \tag{3.37}$$

onde:

- X_t representa um vetor de variáveis exógenas no instante t;
- β é o vetor dos coeficientes de regressão associados às variáveis exógenas.

A incorporação de variáveis preditoras permite captar efeitos exógenos que não são explicados pela dinâmica interna da série, como aumentos de preços, eventos específicos ou variáveis económicas. Antes da inclusão das variáveis externas, a estrutura autorregressiva, de média móvel e de diferenciação (sazonal e não sazonal) deve ser previamente identificada e ajustada.

O processo de estimação conjunta dos parâmetros autorregressivos, de média móvel, e dos coeficientes β é habitualmente realizado através do método da máxima verosimilhança, assumindo a normalidade dos erros ε_t. [9]

No presente projeto, procedeu-se inicialmente ao ajuste de um modelo SARIMA à série temporal da variável resposta, identificando a estrutura $(p,d,q)(P,D,Q)_s$ mais adequada através da análise exploratória e otimização de métricas de erro. Posteriormente, e considerando a seleção de variáveis preditoras considerada na Secção 3.5, evoluiu-se para o modelo SARIMAX, incorporando essas variáveis exógenas.

A utilização do modelo SARIMAX revelou-se vantajosa ao permitir não apenas captar a estrutura temporal e sazonal intrínseca da série, mas também quantificar e incorporar o efeito de variáveis explicativas externas sobre a variável dependente, proporcionando previsões mais ajustadas à realidade observada.

3.7 Validação de um Modelo de Previsão

A validação de modelos de previsão de séries temporais constitui uma etapa essencial para garantir a sua fiabilidade e capacidade preditiva em dados não observados. Um modelo apenas se considera adequado quando, para além de um bom ajuste à amostra de treino, evidencia

capacidade para generalizar previsões com qualidade para novos dados. Neste projeto, a validação dos modelos foi realizada de forma sistemática, recorrendo a diferentes abordagens complementares: critérios de informação, análise dos resíduos na amostra de treino e avaliação das métricas de erro na amostra de teste.

3.7.1 Critério de Informação de *Akaike* (AIC)

Durante o processo de seleção da melhor combinação possível de variáveis preditoras utilizou-se, novamente, um algoritmo *stepwise* com o método de *Forward Selection*, onde a principal métrica de comparação entre os diferentes modelos foi o Critério de Informação de *Akaike*. O AIC é um indicador de desempenho do modelo ajustado que penaliza a complexidade do modelo, procurando um equilíbrio entre o ajuste à amostra e a parcimónia, sendo definido por: [24]

$$AIC = -2ln(L) + 2k \tag{3.38}$$

onde:

- L é a função de verosimilhança maximizada do modelo ajustado;
- k é o número total de parâmetros estimados pelo modelo

Este critério permite selecionar o conjunto de preditores que, mantendo um bom ajuste, minimiza o risco de sobreajuste (*overfitting*). Assim, a configuração com menor valor de AIC foi considerada a mais adequada para cada combinação de variáveis exógenas testadas no modelo SARIMAX.

3.7.2 Análise de Resíduos

Um dos métodos fundamentais para validar a adequação de um modelo de previsão a um conjunto de dados é a análise dos resíduos na amostra de treino. Os resíduos de um modelo de séries temporais devem comportar-se como um ruído branco, isto é, uma sequência de variáveis aleatórias independentes e identicamente distribuídas com média zero e variância constante. [10]

Um resíduo do modelo pode ser definido como sendo o erro estimado que resulta do ajustamento deste. É dado pela diferença entre o valor observado da variável resposta e o valor estimado pelo modelo, ou seja:

$$\hat{\epsilon_i} = y_i - \hat{y}_i \tag{3.39}$$

Através da análise destes resíduos devem verificar-se alguns pressupostos, sendo eles:

- Pressuposto da Não Correlação dos resíduos;

- Pressuposto da Homocedasticidade dos resíduos;

Para além destes, é ainda desejável que os resíduos tenham uma distribuição normal e que a sua média esteja muito perto de zero.

A análise de resíduos permite, por isso, saber se há a necessidade de aprimorar o modelo, fazer ajustes nas suposições ou identificar a necessidade de aplicar técnicas de transformação nos dados. Esta análise é uma etapa crucial no processo de modelação estatística, ajudando a garantir a confiabilidade e a validade dos resultados obtidos.

3.7.2.1 Não Correlação dos Resíduos

O primeiro pressuposto na análise de resíduos pode ser verificado através de métodos gráficos, como é o caso do correlograma. No entanto, as conclusões retiradas a partir do gráfico podem ser confirmadas analiticamente com o teste de hipóteses de *Ljung-Box*.

O gráfico da função de autocorrelação dos resíduos, ou correlograma, apresenta um intervalo, cujos limites estão delineados com uma banda azul, e um conjunto de traços verticais representando os *lags*. Caso a totalidade (ou perto disso) dos traços esteja toda contida dentro desse intervalo, verifica-se a independência dos resíduos. Caso contrário, este pressuposto não é verificado. É exemplo de um correlograma o gráfico da Figura 3.10.

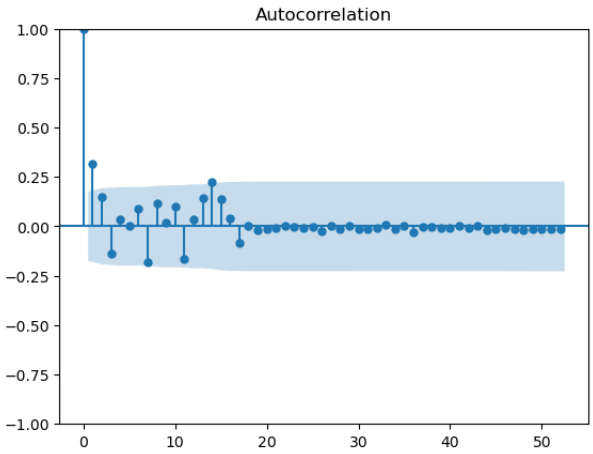

Figura 3.10: Exemplo de um Correlograma de uma das variáveis estudadas neste trabalho.

Em relação ao teste de hipóteses de *Ljung-Box*, as hipóteses são:

$$\begin{cases} H_0 : \rho_1 = \rho_2 = \cdots = \rho_L = 0 \\ H_1 : \exists_j : \beta_j \neq 0 \end{cases} \tag{3.40}$$

onde ρ_j representa a autocorrelação no desfasamento j.

O parâmetro L corresponde ao número de desfasamentos considerados no teste e não deve ser associado diretamente ao comprimento do ciclo sazonal. Na prática, a escolha de

L depende do contexto: valores demasiado pequenos podem não detetar autocorrelações relevantes, enquanto valores demasiado grandes podem reduzir o poder do teste. Assim, recomenda-se selecionar L de forma a abranger os principais desfasamentos de interesse para a análise, sendo comum considerar valores como 10, 20 ou até $\sqrt{n}$, em que n é o número de observações da série.

Dado que a hipótese nula testa se os resíduos são independentes entre si, logo para o pressuposto ser verificado é necessário não rejeitar H_0.

3.7.2.2 Homocedasticidade dos Resíduos

Este segundo pressuposto assume que a variância dos resíduos se mantém constante ao longo do tempo. Esta verificação pode, tal como no pressuposto anterior, ser realizada através de métodos gráficos, recorrendo ao cronograma dos resíduos. Para uma confirmação analítica daquilo que se observa graficamente, é possível recorrer ao teste de heterocedasticidade utilizado pela biblioteca *statsmodels* em *Python*. Este baseia-se na comparação entre dois modelos de séries temporais estruturais estimados através do filtro de *Kalman*: um que assume uma variância constante dos erros e outro que permite que a variância não seja constante [25].

As hipóteses deste teste são definidas da seguinte forma:

$$\begin{cases} H_0 : \text{Não existe mudança estrutural na variância dos resíduos (homocedasticidade).} \\ H_1 : \text{Existe mudança estrutural na variância dos resíduos (heterocedasticidade).} \end{cases} \tag{3.41}$$

A estatística utilizada é uma estatística de razão de verosimilhança, calculada pela diferença entre a log-verosimilhança do modelo estimado sob a hipótese alternativa e a log-verosimilhança do modelo sob a hipótese nula. A expressão da estatística é:

$$H(h) = 2(logL_1 - logL_0) \tag{3.42}$$

onde $logL_1$ representa a log-verosimilhança do modelo estimado admitindo heterocedasticidade, $logL_0$ representa a log-verosimilhança do modelo sob a hipótese nula de homocedasticidade e h corresponde ao número de restrições impostas pelo modelo sob H_0. Sob H_0, a estatística $H(h)$ segue aproximadamente uma distribuição qui-quadrado com h graus de liberdade.

Caso o *p-value* associado ao teste for superior ao nível de significância de 0.05, não se rejeita a hipótese nula. Assim, considera-se que não existe evidência estatística suficiente para concluir que os resíduos apresentam heterocedasticidade.

3.7.3 Avaliação de Métricas de Erro

Para além da validação interna na amostra de treino, é indispensável avaliar a capacidade preditiva do modelo em dados fora da amostra. Para esse efeito, foram calculadas diversas métricas de erro de previsão sobre a amostra de teste, cada uma fornecendo uma perspetiva distinta sobre a performance do modelo.

Para cada métrica de erro apresentada a seguir, considera-se Y_t como o valor observado da série no instante t, $\hat{Y}_t$ como o valor previsto pelo modelo no instante t e n o tamanho da amostra.

3.7.3.1 *Mean Absolute Error* (MAE)

Esta métrica mede o erro médio absoluto entre os valores observados e os previstos, sendo insensível a *outliers*.

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|Y_t - \hat{Y}_t| \tag{3.43}$$

3.7.3.2 *Mean Squared Error* (MSE)

Esta métrica penaliza fortemente os erros elevados devido à elevação ao quadrado.

$$MSE = \frac{1}{n}\sum_{t=1}^{n}(Y_t - \hat{Y}_t)^2 \tag{3.44}$$

3.7.3.3 *Root Mean Squared Error* (RMSE)

Esta métrica é a raíz quadrada do MSE, mantendo a unidade de medida da variável e sendo sensível a grandes erros.

$$RMSE = \sqrt{MSE} \tag{3.45}$$

3.7.3.4 *Mean Absolute Percentage Error* (MAPE)

Esta métrica expressa o erro absoluto médio em termos percentuais. Contudo, é sensível a valores próximos de zero.

$$MAPE = \frac{100}{n}\sum_{t=1}^{n}\left|\frac{Y_t - \hat{Y}_t}{Y_t}\right| \tag{3.46}$$

Para séries que tenham observações com valores nulos, este termo torna-se infinito.

3.7.3.5 *Weighted Mean Absolute Percentage Error* (WMAPE)

Esta métrica é uma versão ponderada do MAE e é muito utilizada em contexto empresarial e logístico. [26]

$$WMAPE = \frac{\sum |Y_t - \hat{Y}_t|}{\sum Y_t} \tag{3.47}$$

3.7.3.6 *Symmetric Mean Absolute Percentage Error* (SMAPE)

Esta métrica corrige algumas limitações do MAPE, nomeadamente a assimetria e a sensibilidade a valores próximos de zero.

$$SMAPE = \frac{100}{n} \sum_{t=1}^{n} \frac{|Y_t - \hat{Y}_t|}{(|Y_t| + |\hat{Y}_t|)/2} \tag{3.48}$$

A combinação destas métricas permite avaliar a performance preditiva de forma abrangente, considerando não só a magnitude absoluta dos erros, mas também a sua proporcionalidade face aos valores reais e previstos.

Capítulo 4

Caso de Estudo

Neste capítulo aplica-se a metodologia desenvolvida ao caso concreto da previsão da procura de produtos comercializados pela empresa, no âmbito do projeto de apoio ao planeamento da procura.

Embora o estudo tenha abrangido diversos produtos pertencentes a diferentes famílias, apresentam-se aqui, em detalhe, apenas os resultados referentes a um produto. Os resultados relativos aos restantes encontram-se disponibilizados nos Anexos.

4.1 Caracterização do Produto

O produto analisado neste estudo encontra-se identificado pelo código interno de gestão **045000** e pertence à família das cervejas. A sua seleção justifica-se por constituir um caso particularmente desafiante no âmbito da previsão da procura, dado o padrão temporal irregular e as alterações estruturais registadas ao longo do período considerado.

Por motivos de confidencialidade, e à semelhança da representação adotada para o produto em análise, todas as variáveis explicativas são designadas genericamente por X_i, com $i = 1, 2, 3, ...$

A Figura 4.1 apresenta o cronograma da procura semanal do produto **045000** na cidade de Lisboa, entre maio de 2022 e maio de 2025.

A análise visual inicial evidenciou a ocorrência de picos pontuais de vendas e instabilidades a partir de agosto de 2023, com particular intensidade a partir de março de 2024. Estes padrões reforçam a necessidade de recorrer a modelos capazes de lidar com irregularidades e alterações estruturais na procura.

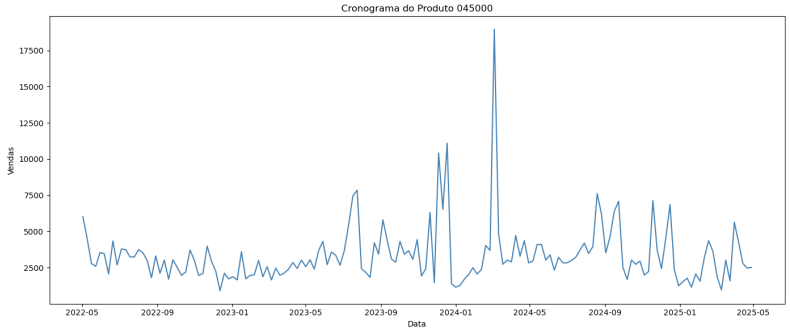

Figura 4.1: Cronograma de **045000**.

4.2 Análise Descritiva de 045000

O primeiro passo da *pipeline* implementada consiste na avaliação de medidas descritivas associadas à série temporal da procura do produto.

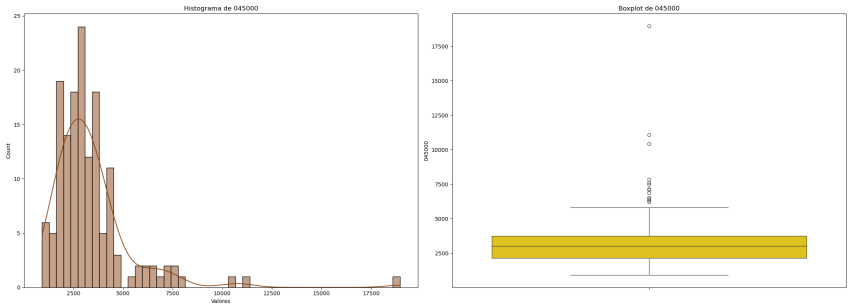

Figura 4.2: Histograma e *Boxplot* de **045000**.

Relativamente à tendência central, a análise revelou um valor médio de 3373.04 unidades e uma mediana de 3000 unidades. Esta diferença sugere a existência de assimetria na distribuição. Com efeito, o histograma apresentado à esquerda da Figura 4.2 confirma que a distribuição dos dados, compreendidos entre 912 e 18960 unidades, é assimétrica positiva. Esta conclusão é corroborada pelo coeficiente de *Bowley*, cujo valor $g_B = 3.63 > 0$ reforça a presença de assimetria.

A utilização deste coeficiente justifica-se pela existência de diversos *outliers*, conforme evidenciado no *boxplot* apresentado à direita da Figura 4.2.

Adicionalmente, a análise da dispersão, com um desvio padrão de $s = 2060.43$, bem como o coeficiente de variação resistente ($CVR = 61.09\% > 30\%$), indicam que a série deverá ser sujeita a uma transformação prévia antes do início do processo de modelação.

4.3 Primeiro *Fold* da Pipeline

Nesta secção apresentam-se os resultados obtidos para todas as etapas da *pipeline* implementada, referentes ao primeiro conjunto treino/teste, ou *fold*, da validação cruzada temporal. De acordo com o procedimento metodológico definido, esta primeira divisão considerou as primeiras 101 observações para o treino do modelo e as 25 observações seguintes para teste.

Assim, a amostra de treino corresponde ao período entre maio de 2022 e abril de 2024, enquanto a amostra de teste abrange as observações compreendidas entre maio e outubro de 2024.

4.3.1 Transformação da Variável Resposta

A série temporal correspondente à variável resposta apresentou variância não constante ao longo do tempo, o que inviabiliza a aplicação direta de modelos estatísticos que assumem homocedasticidade dos resíduos. Para ultrapassar esta limitação, aplicou-se a transformação de *Box-Cox*, cujo parâmetro de transformação assumiu o valor $\lambda = -0.26$, conforme descrito na Secção 3.5.

O efeito da transformação pode ser observado na Figura 4.3.

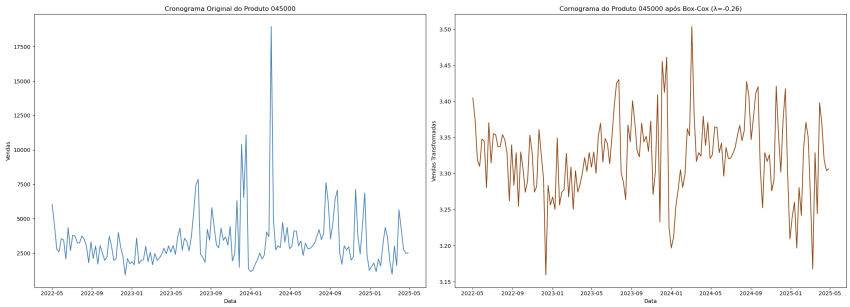

Figura 4.3: Série da procura de **045000** transformada por *Box-Cox* com $\lambda = -0.26$.

Após a aplicação da transformação de *Box-Cox*, a distribuição tornou-se aproximadamente simétrica e contínua, conforme ilustrado no cronograma à direita da Figura 4.3, reduzindo a influência de *outliers* e estabilizando a variância ao longo da série.

Atendendo a esta aproximação à normalidade e à ausência de distorções significativas na distribuição, considerou-se adequada a utilização da função de ligação identidade nos Modelos Lineares Generalizados apresentados mais adiante. Esta escolha assegura a linearidade entre os preditores e a variável resposta transformada, preservando simultaneamente a interpretabilidade estatística dos coeficientes.

4.3.2 Modelo Univariado

Após a decomposição STL da série original da procura semanal do produto, não se identificaram componentes de tendência ou de sazonalidade. Assim, estimou-se um modelo de alisamento exponencial simples e calcularam-se as previsões para a amostra de teste, cujas métricas de erro se encontram apresentadas na Tabela 4.1.

MAE	MSE	RMSE	MAPE	WMAPE	SMAPE
1202.92	2.32×10^6	1523.99	33.89	31.18	15.04

Tabela 4.1: Métricas de erro para as previsões na amostra de teste do Modelo de Alisamento Exponencial Simples do primeiro *fold*.

A análise das métricas de erro obtidas para o modelo univariado de Alisamento Exponencial Simples evidencia limitações significativas na sua capacidade preditiva relativamente à série temporal em estudo:

- O valor do **MAE** indica que, em média, as previsões apresentam um desvio absoluto de aproximadamente 1203 unidades face aos valores observados, o que, considerando a média da série (3373.04 unidades), representa um erro expressivo.
- O **MSE** (2.32×10^6) e o correspondente **RMSE** (1523.99) confirmam esta tendência, dado que o erro quadrático médio penaliza de forma acentuada os desvios mais elevados.
- Em termos percentuais, o **MAPE** de 33.89% revela que, em média, o erro absoluto corresponde a cerca de um terço do valor observado, evidenciando baixa precisão.
- O **WMAPE** (31.18%) reforça esta conclusão ao ponderar o erro absoluto pela procura total, indicando que aproximadamente 31% da procura não foi corretamente antecipada pelo modelo.
- Por fim, o **SMAPE** (15.04%), mais equilibrado na avaliação de erros relativos em séries com variações significativas, revela ainda assim uma margem de erro relevante para o planeamento operacional.

Em síntese, estes resultados demonstram que, embora o modelo univariado constitua um ponto de partida válido, a sua capacidade preditiva é insuficiente para sustentar de forma eficaz o processo de planeamento da procura deste produto, justificando a necessidade de recorrer a modelos multivariados mais robustos.

4.3.3 Seleção de Variáveis Preditoras

Inicia-se, nesta fase, o processo de seleção das variáveis preditoras a integrar nos modelos multivariados.

Após a aplicação da transformação de *Box-Cox*, deu-se início ao processo de seleção de variáveis, conduzido de acordo com as abordagens previamente descritas na Secção 3.5.

4.3.3.1 Primeira Abordagem

Revisitando a metodologia definida na Secção 3.5.1, a **Etapa 1** da primeira abordagem de seleção de variáveis consistiu na identificação e remoção da tendência das variáveis explicativas que o justificassem. Para tal, aplicou-se a decomposição STL a todas as variáveis preditoras, com o objetivo de isolar a componente de tendência de cada série.

A aplicação desta técnica evidenciou a presença de tendência estatisticamente e visualmente clara em apenas duas variáveis: X_4 e X_7. A variável X_4 apresentou uma tendência decrescente ao longo do período analisado, enquanto X_7 revelou uma tendência fortemente crescente. Estes comportamentos encontram-se representados na Figura 4.4.

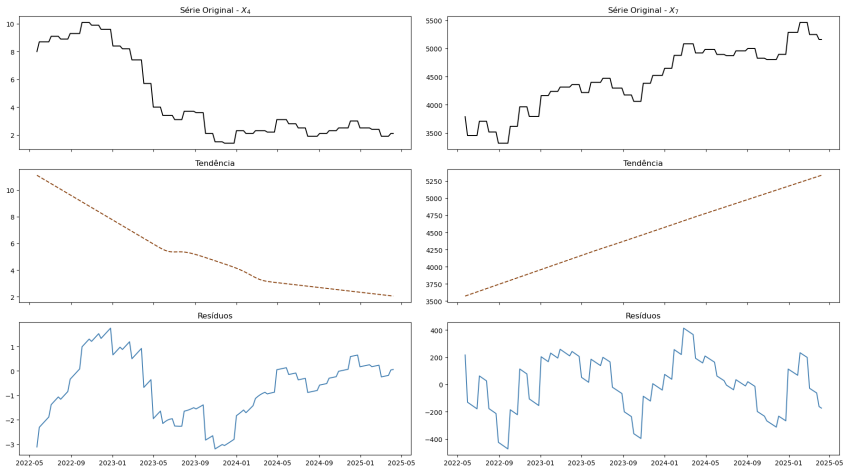

Figura 4.4: Decomposição STL aplicada às variáveis X_4 e X_7, evidenciando as componentes de tendência.

Concluída esta etapa, foram criadas seis variáveis de desfasamento (*lags*) para cada variável explicativa. Especificamente, a partir de cada variável original foram geradas seis novas variáveis, correspondentes aos valores desfasados até três períodos anteriores e até três períodos posteriores.

Deste modo, o conjunto de dados passou de 14 variáveis explicativas para 98. Para facilitar a leitura ao longo do texto, as variáveis desfasadas relativas a X_i serão designadas por $X_i^{(j)}$, com $j \in \{-3, -2, -1, 1, 2, 3\}$.

Prosseguindo com o processo descrito na Secção 3.5.1, na **Etapa 2** desta abordagem identificaram-se as variáveis que apresentavam uma relação linear significativa com a variável resposta, obtendo-se um grupo de 40 variáveis.

Na **Etapa 3**, as restantes 30 variáveis contínuas que não evidenciaram relação linear com a variável resposta foram modeladas recorrendo a *splines*.

Com o objetivo de reduzir ainda mais o número de preditores, na **Etapa 4** analisou-se a correlação entre cada uma das 43 variáveis referidas anteriormente e a variável resposta. Destas, foram consideradas apenas as variáveis cuja correlação mútua fosse inferior a 0.3 e cuja correlação com a variável resposta fosse superior a esse valor. Contudo, apenas uma variável cumpriu simultaneamente estes critérios: X_{13}^{-2}.

Seguidamente, analisou-se o grupo das variáveis qualitativas. Conforme definido na **Etapa 5**, aplicaram-se dois testes não paramétricos a este conjunto, a partir dos quais apenas cinco variáveis se revelaram estatisticamente significativas: três variáveis categóricas (X_5, X_5^{+1} e X_5^{+2}) e duas variáveis categóricas (X_4^{+1} e X_4^{+3}).

Em síntese, avançaram para a **Etapa 6** desta abordagem um total de 36 variáveis. Nesta última etapa, o conjunto foi submetido a um processo de seleção automática com recurso ao algoritmo de *Forward Selection* num modelo GLM, resultando no primeiro conjunto de potenciais preditores:

- X_5 e X_5^{+1}, variáveis categóricas;
- X_{13}^{-2}, variável contínua.

4.3.3.2 Segunda Abordagem

Na segunda abordagem, procedeu-se inicialmente à estacionarização das variáveis explicativas através do teste de *Dickey-Fuller* Aumentado (ADF), replicando-se posteriormente os procedimentos descritos na Secção 3.5.2.

Na **Etapa 1**, nenhuma variável foi considerada estacionária. Assim, aplicou-se a diferenciação a cada uma das séries que, após novo teste de *Dickey-Fuller*, se revelaram estacionárias.

Na **Etapa 2**, identificaram-se apenas duas variáveis com relação linear significativa com a variável resposta: X_2^{-1} e X_3^{-3}.

Na **Etapa 3**, as restantes 68 variáveis contínuas que não apresentaram relação linear com a variável resposta foram modeladas utilizando *splines*.

Na **Etapa 4**, nenhuma das duas variáveis identificadas anteriormente respeitou os critérios de correlação previamente definidos.

Na **Etapa 5**, apenas uma variável foi considerada estatisticamente significativa: X_5^{+3}.

Em síntese, avançaram para a **Etapa 6** desta abordagem um total de 69 variáveis, das quais resultou o segundo conjunto de potenciais preditores:

- X_2^{-3}, variável contínua modelada com uma *spline* com 4 nós e 10 graus de liberdade;
- X_{14}^{-1}, variável contínua modelada com uma *spline* com 4 nós e 10 graus de liberdade;

4.3.3.3 Terceira Abordagem

Por fim, na terceira abordagem aplicou-se a técnica de regressão Lasso, conforme descrito na Secção 3.5.3, da qual resultou o terceiro conjunto de potenciais preditores:

- X_1^{+3}, variável contínua;
- X_5^{-1}, X_5 e X_5^{+1}, variáveis categóricas;
- X_6^{+2} e X_6^{+3}, variáveis categóricas;
- X_8^{-3}, variável contínua.

A união dos preditores identificados nas três abordagens constituiu o conjunto final de potenciais variáveis a considerar na fase de *Forward Selection* do modelo SARIMAX.

O conjunto final de potenciais preditores é composto pelas seguintes variáveis:

- X_1^{+3}, variável contínua;
- X_2^{-3}, variável contínua modelada com uma *spline* com 4 nós e 10 graus de liberdade;
- X_5^{-1}, X_5 e X_5^{+1}, variáveis categóricas;
- X_6^{+2} e X_6^{+3}, variáveis categóricas;
- X_8^{-3}, variável contínua;
- X_{13}^{-2}, variável contínua;
- X_{14}^{-1}, variável contínua modelada com uma *spline* com 4 nós e 10 graus de liberdade.

4.3.4 Parâmetros do Modelo ARIMA

O modelo univariado SARIMA foi ajustado com base na série temporal transformada da variável resposta, tendo sido identificada a estrutura $(3, 0, 0)$ como a mais adequada, por minimizar o valor do AIC (-317.76).

A ausência de termos sazonais justifica-se pelo facto de, conforme referido na Secção 4.3.2, a série não evidenciar a presença de um ciclo sazonal bem definido.

4.3.5 Modelo SARIMAX Final

Através da aplicação do método de *Forward Selection*, utilizando as variáveis previamente selecionadas nas diferentes abordagens e os parâmetros definidos para o modelo SARIMA, obteve-se o modelo SARIMAX final, que apresentou um AIC de -318.20.

O conjunto final de preditores selecionados para este modelo é o seguinte:

- X_5^{-1} e X_5^{+1}, variáveis categóricas;
- X_6^{+2}, variável categórica;
- X_8^{-3}, variável contínua;
- X_{14}^{-1}, variável contínua modelada com uma *spline* com 4 nós e 10 graus de liberdade.

Com as variáveis selecionadas e os parâmetros anteriormente definidos, a equação do modelo pode ser expressa da seguinte forma:

$$\begin{aligned}\hat{Y}_t = 0.5860\, Y_{t-1} + 0.3560\, Y_{t-2} + 0.0581\, Y_{t-3} + 0.0322\, X_5^{-1} + \\ + 0.0395\, X_5^{+1} - 0.1862\, X_6^{+2} - 0.1307\, X_8^{-3} + g(X_{14}^{-1})\end{aligned} \tag{4.1}$$

onde:

- $\hat{Y}_t$ corresponde à previsão da variável resposta transformada através de *Box-Cox*;
- Y_{t-1}, Y_{t-2} e Y_{t-3} representam os termos autorregressivos de ordem 1, 2 e 3, respetivamente;
- $g(X_{14}^{-1})$ denota o efeito não linear da variável X_{14}^{-1}, modelado por meio de uma *spline*.

Após a construção do modelo e a geração das previsões para a amostra de treino, procedeu-se à inversão das mesmas através da equação (3.27), considerando $\lambda = -0.26$. Deste modo, foi possível representar graficamente os resultados do modelo para a amostra de treino e compará-los com os valores observados, conforme ilustrado na Figura 4.5.

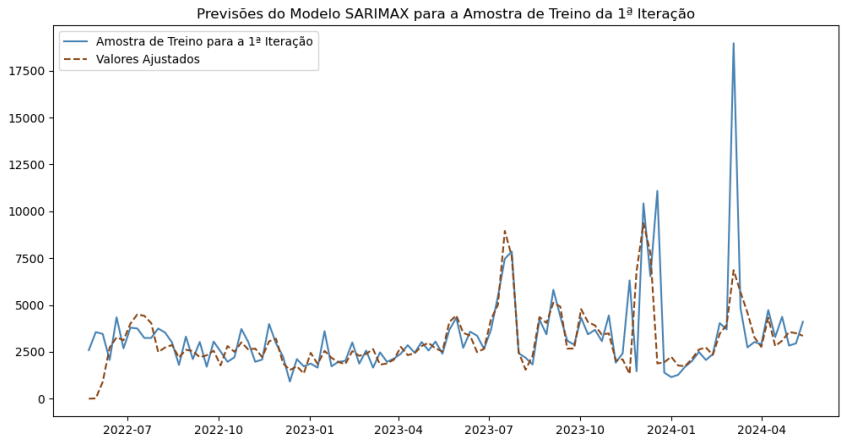

Figura 4.5: Previsões do Modelo SARIMAX para a Amostra de Treino do primeiro *fold*.

Ao analisar a figura, observa-se uma melhoria progressiva na capacidade do modelo de prever os valores reais da série temporal. A previsão das observações referentes ao ano de 2022 revelou-se insuficientemente precisa; contudo, a partir de 2023, o modelo apresentou um desempenho mais adequado. Destacam-se as previsões corretas para os *outliers* ocorridos em julho e dezembro de 2023, enquanto que, em março de 2024, o modelo não conseguiu atingir as 17 500 unidades, apesar de ter previsto um aumento da procura nesse período.

Torna-se, assim, fundamental analisar o comportamento dos resíduos gerados por estas previsões, de modo a validar ou refutar a adequação do modelo. Apenas através desta análise será possível determinar se o modelo é apropriado para a previsão da amostra de teste.

4.3.6 Validação do Modelo

Conforme explicado na Secção 3.7, um resíduo é definido como a diferença entre o valor observado e o valor previsto pelo modelo.

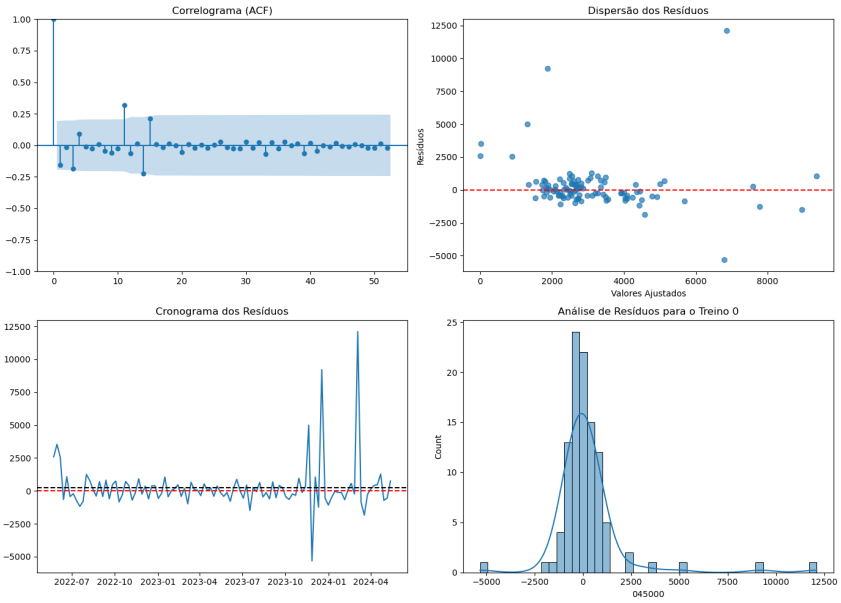

Figura 4.6: Resíduos do Modelo SARIMAX para a Amostra de Treino do primeiro *fold*.

Existem diversas representações gráficas que permitem analisar os resíduos de um modelo. Na Figura 4.6 apresentam-se quatro dessas representações, relativas aos resíduos do modelo estimado na secção anterior.

4.3.6.1 Pressuposto da Independência

Para validar este pressuposto, torna-se necessário analisar o correlograma apresentado na posição superior esquerda da Figura 4.6.

Embora a linha correspondente ao 11º *lag* se situe fora da banda azul, a maioria das demais linhas permanece dentro desta banda. Por conseguinte, considera-se que não existem fundamentos para invalidar este pressuposto.

A confirmação desta conclusão é efetuada através da análise dos resultados do teste de *Ljung-Box*:

Estatística de Teste	***p-value***
36.2436	0.9524

Tabela 4.2: Resultados do teste de *Ljung-Box* para os resíduos do modelo SARIMAX para o primeiro *fold*.

Dado que o *p-value* obtido é superior ao nível de significância de 5%, a hipótese nula do teste não é rejeitada, concluindo-se que os resíduos são independentes entre si.

4.3.6.2 Pressuposto da Homocedasticidade

O pressuposto de que os resíduos apresentam variância constante pode ser avaliado graficamente através de dois gráficos distintos.

No gráfico de dispersão dos resíduos, localizado na posição superior direita da Figura 4.6, embora se verifiquem valores extremos evidentes, não se observa um padrão claro na dispersão dos resíduos.

Adicionalmente, o correlograma situado na posição inferior esquerda indica que a variância dos resíduos se mantém relativamente constante, sendo identificadas apenas anomalias pontuais nos períodos correspondentes aos *outliers*.

Deste modo, recorreu-se ao *output* do teste de heterocedasticidade para confirmar estas suposições:

Estatística de Teste	***p-value***
1.5422	0.2117

Tabela 4.3: Resultados do teste de heterocedasticidade para os resíduos do modelo SARIMAX para o primeiro *fold*.

Como o *p-value* é superior ao nível de significância de 5%, a hipótese nula do teste não é rejeitada, concluindo-se que não ocorre alteração estrutural na variância dos resíduos, ou seja, confirma-se a presença de homocedasticidade.

4.3.6.3 Pressuposto da Nulidade da Média

Ao reexaminar o correlograma, observa-se que a linha tracejada preta, representando a média dos resíduos, encontra-se muito próxima da linha tracejada vermelha correspondente ao valor zero. Assim, considera-se este pressuposto validado.

4.3.6.4 Pressuposto da Normalidade

Por fim, ao observar a distribuição dos resíduos no histograma localizado na posição inferior direita da Figura 4.6, verifica-se que esta se aproxima de uma distribuição normal, com exceção dos *outliers*, que introduzem alguma assimetria.

Após esta análise, todos os pressupostos do modelo foram validados. Apenas o pressuposto relativo à normalidade da distribuição dos resíduos apresenta algumas incertezas. Deste modo, confirma-se a adequada adequação do modelo ao conjunto de dados, permitindo a sua utilização para a realização das previsões sobre a amostra de teste.

4.3.7 Previsões do Modelo para a Amostra de Teste

As previsões do modelo apresentado na Equação 4.1 encontram-se ilustradas na Figura 4.7:

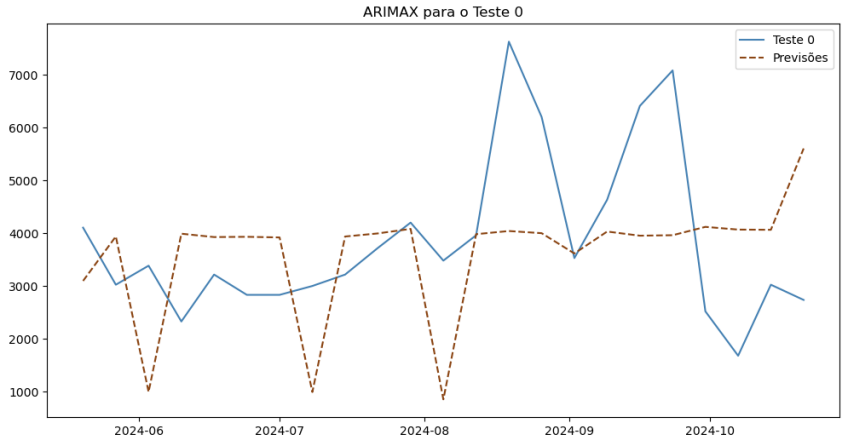

Figura 4.7: Previsões do Modelo SARIMAX para a Amostra de Teste do primeiro *fold*.

Ao analisar as previsões para a amostra de teste, verifica-se que estas não se assemelham de forma consistente aos valores observados. Embora em determinados momentos as previsões coincidam com os valores reais, o seu comportamento geral difere significativamente da realidade.

Para este *fold*, as métricas de erro calculadas foram as seguintes:

MAE	MSE	RMSE	MAPE	WMAPE	SMAPE
1503.35	3.31×10^6	1819.21	43.33	38.97	21.90

Tabela 4.4: Métricas de erro para as previsões na amostra de teste do Modelo SARIMAX do primeiro *fold*.

A avaliação das métricas de erro associadas ao modelo SARIMAX revela, de forma algo surpreendente, um desempenho inferior ao do modelo univariado previamente ajustado:

- O **MAE** de 1503.35 indica um erro absoluto médio superior, representando um acréscimo de cerca de 25% em relação ao valor de 1202.92 obtido no modelo de Alisamento Exponencial Simples.
- De igual modo, tanto o **MSE** (3.31×10^6) como o **RMSE** (1819.21) apresentam aumentos significativos, refletindo previsões com desvios mais acentuados, penalizando particularmente os erros de maior magnitude.
- No que se refere às métricas percentuais, o **MAPE** (43.33%) e o **WMAPE** (38.97%) confirmam esta deterioração, evidenciando erros relativos mais elevados, com o modelo SARIMAX a apresentar um desvio percentual médio superior em aproximadamente 9 pontos percentuais no MAPE e 7.8 pontos percentuais no WMAPE.
- O **SMAPE** (21.90%), embora seja uma métrica menos sensível a *outliers* e mais estável em séries de elevada variabilidade, também regista um aumento de 6,86 pontos percentuais.

Esta redução global do desempenho sugere que, no caso específico desta série e configuração, a inclusão das variáveis exógenas e a estrutura multivariada não contribuíram para melhorar a capacidade preditiva do modelo. Este resultado poderá dever-se ao reduzido tamanho do conjunto de treino ou à instabilidade estrutural da série a partir de março de 2024, já previamente identificada. Assim, estes resultados reforçam a importância de validar criteriosamente os benefícios da inclusão de variáveis externas e de reavaliar a seleção de preditores e a parametrização do modelo em cenários de elevada volatilidade.

Considera-se, por conseguinte, que, para o primeiro *fold* da *pipeline*, os melhores resultados foram obtidos com o Modelo de *Holt-Winters*.

Todo o processo descrito para este primeiro *fold* foi replicado mais 24 vezes e, na secção seguinte, são apresentados os resultados médios obtidos com a aplicação desta *pipeline*.

4.4 Resultados da *Pipeline*

Após a conclusão do processo para os vinte e cinco *folds* da validação cruzada, obteve-se a média dos valores das métricas de erro, tanto para o modelo univariado como para os respetivos modelos multivariados.

É importante salientar que, para cada *fold*, o conjunto de preditores selecionado é independente dos restantes. Aqui, apresenta-se apenas o conjunto de preditores composto pela moda dos mesmos ao longo dos vinte e cinco *folds*. As variáveis explicativas mais frequentemente selecionadas como preditoras são:

- X_5^{+1}, variável categórica, presente em vinte e dois *folds*;
- X_1^{+3}, variável contínua, presente em quinze *folds*;
- X_8^{-3}, variável categórica, presente em quinze *folds*;
- X_5^{-1}, variável categórica, presente em catorze *folds*.

A identificação das variáveis explicativas mais frequentemente selecionadas não implica que um conjunto de preditores formado exclusivamente por estas quatro variáveis constitua uma vantagem para o modelo preditivo. Contudo, permite evidenciar as principais variáveis responsáveis pelo comportamento histórico da procura do produto **045000**.

Na Tabela 4.5 apresentam-se os resultados médios das métricas de erro calculadas para o conjunto de teste ao longo dos vinte e cinco *folds*.

Modelos	MAE	MSE	RMSE	MAPE	WMAPE	SMAPE
Modelo Univariado	1562	8.56×10^6	2329	50.44%	45.00%	20.15%
SARIMAX	1313	2.77×10^6	1659	43.76%	37.06%	18.13%

Tabela 4.5: Média das métricas de erro na amostra de teste utilizando a pipeline para **045000**.

A comparação dos resultados evidencia uma melhoria significativa no desempenho preditivo do modelo SARIMAX em relação ao modelo univariado de referência.

- O **MAE** reduziu-se de 1562 para 1313, correspondendo a uma diminuição de aproximadamente 15.9% no erro absoluto médio. Este decréscimo indica que, em termos médios, as previsões do SARIMAX apresentam um desvio inferior em cerca de 249 unidades face ao modelo univariado, representando uma melhoria operacional significativa para minimizar ruturas ou excessos de *stock*.
- De forma consistente, o **MSE** caiu de 8.56×10^6 para 2.77×10^6, refletindo uma redução expressiva na magnitude dos erros e penalizando menos os desvios de maior amplitude. Esta melhoria repercute-se diretamente no **RMSE**, que diminuiu de 2329 para 1659 (menos 28.7%), evidenciando que o SARIMAX não só reduz a média dos erros como também diminui a sua variabilidade, proporcionando maior estabilidade no desempenho preditivo.

- Nas métricas percentuais, o **MAPE** passou de 50.44% para 43.76%, e o **WMAPE** de 45.00% para 37.06%. A redução do MAPE traduz-se numa diminuição da proporção média do erro em relação ao valor real, enquanto o decréscimo do WMAPE, mais relevante do ponto de vista empresarial, indica que o impacto global do erro na procura total foi reduzido em cerca de 7.94 pontos percentuais, resultando em previsões mais fiéis nos períodos de maior volume de vendas.
- O **SMAPE** registou uma redução mais modesta, de 20.15% para 18.13%, mas esta descida indica um melhor ajuste às flutuações relativas da procura. Por ser simétrico, este indicador evidencia que a melhoria não ocorreu apenas em casos de subestimação ou sobrestimação, mas de forma equilibrada em ambas as situações.

No seu conjunto, estes resultados confirmam que, para esta configuração e horizonte temporal, a inclusão de variáveis exógenas no modelo SARIMAX permitiu captar informação adicional relevante, reduzindo os erros médios, melhorando a estabilidade das previsões e aumentando a precisão relativa. Este desempenho representa um ganho tangível para o planeamento da procura, permitindo decisões mais assertivas na gestão de stocks e na programação logística.

Na Figura 4.8 apresenta-se o comportamento gráfico das previsões geradas para a amostra de teste do produto **045000**.

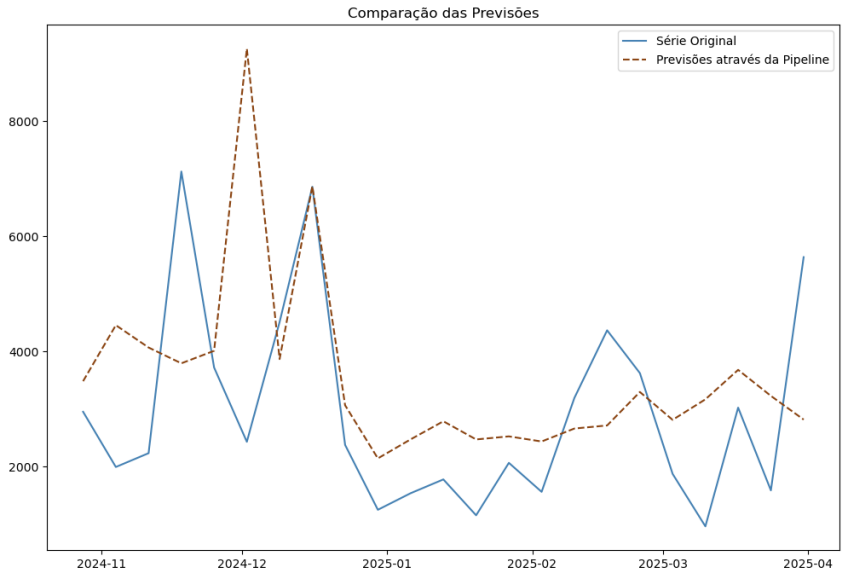

Figura 4.8: Média das previsões geradas pela *Pipeline* na amostra de teste de **045000**.

De forma geral, as previsões acompanham o nível médio da série real, o que está em concordância com as melhorias observadas nas métricas de erro face ao modelo univariado.

Verifica-se que o modelo consegue replicar alguns movimentos ascendentes e descendentes mais amplos, como a subida abrupta em dezembro de 2024 e a descida subsequente, embora tenda a suavizar variações mais acentuadas. Esta característica é comum em modelos SARIMAX quando os preditores não captam integralmente choques pontuais ou eventos extra-

ordinários.

Em determinados pontos, particularmente no primeiro pico de novembro de 2024 e na queda subsequente, registam-se subestimações e sobrestimações consideráveis, explicando os valores ainda elevados do MAPE e do SMAPE.

No início de 2025, quando a procura real apresenta níveis mais baixos, o modelo tende a sobrestimar ligeiramente, evidenciando alguma sensibilidade excessiva ao nível médio estimado. Do ponto de vista operacional, esta suavização das previsões pode ser benéfica para reduzir o risco de ruturas em períodos de elevada procura, mas pode conduzir a excessos de stock em fases de menor procura. Assim, a adequação deste comportamento deve ser avaliada em função da política de inventário e da tolerância ao erro da empresa.

Capítulo 5

Resultados e Conclusões

O presente trabalho teve como principal objetivo apoiar o processo de planeamento da procura no Grupo Nabeiro, através do desenvolvimento de uma *pipeline* de previsão baseada no modelo SARIMAX. O grande foco deste estudo residiu na seleção criteriosa de variáveis explicativas, uma tarefa considerada desafiante pela empresa, sendo crucial para maximizar a capacidade preditiva do modelo. A *pipeline* concebida integra, de forma estruturada, metodologias de seleção de variáveis, modelação e validação rigorosa, permitindo não apenas melhorar a precisão das previsões, mas também assegurar a consistência e a reprodutibilidade dos resultados.

5.1 Comentário aos Resultados Obtidos

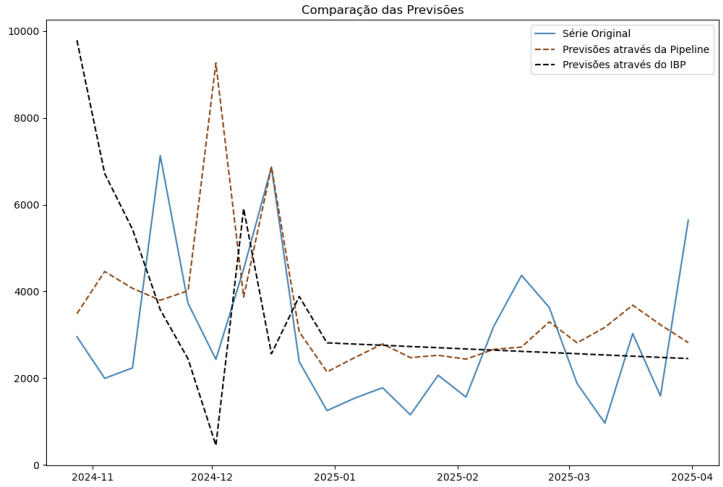

Figura 5.1: Comparação entre previsões obtidas pela *pipeline* e pelo SAP IBP para **045000**.

A comparação entre as previsões geradas pela *pipeline* desenvolvida e aquelas obtidas através do sistema SAP IBP evidencia diferenças significativas. Conforme ilustrado na Figura 5.1, o modelo implementado neste projeto reproduziu de forma mais fidedigna as oscilações da procura, aproximando-se com maior precisão da série original. Em particular, a *pipeline* demonstrou maior eficácia na identificação de picos e quedas súbitas no consumo, enquanto o

SAP IBP apresentou previsões mais suavizadas e tendencialmente conservadoras, revelando limitações na captação da variabilidade da procura.

Este resultado confirma a vantagem da inclusão de variáveis exógenas e da utilização de metodologias estatísticas multivariadas, que permitiram reduzir os erros de previsão e aproximar os valores estimados da realidade observada. Embora algumas discrepâncias ainda se verifiquem em períodos de maior irregularidade, os resultados evidenciam a robustez da abordagem e a sua utilidade prática no apoio à tomada de decisão no planeamento da procura.

5.2 Sugestões de Trabalho Futuro

Apesar dos avanços alcançados, permanecem diversas oportunidades para aprofundar o trabalho desenvolvido, nomeadamente:

1. Explorar modelos de *Machine Learning*, como XGBoost ou Redes Neuronais Recorrentes (LSTM/GRU), capazes de captar relações não lineares mais complexas;
2. Alargar o conjunto de variáveis exógenas, integrando indicadores económicos, tendências de mercado e outros fatores externos que possam influenciar o consumo;
3. Testar previsões com maior granularidade, seja por canal de venda, região geográfica ou família de produtos, de modo a aumentar a aplicabilidade operacional.

5.3 Conclusões Finais

Conclui-se que a metodologia proposta constitui uma mais-valia para o Grupo Nabeiro, ao fornecer previsões mais precisas e alinhadas com a procura efetiva. A *pipeline* desenvolvida combina rigor estatístico, interpretabilidade e flexibilidade, oferecendo um processo replicável e adaptável a diferentes contextos.

Do ponto de vista empresarial, este contributo traduz-se em maior fiabilidade no planeamento da procura, na mitigação do risco de ruturas ou excessos de *stock* e, consequentemente, em ganhos de eficiência operacional e sustentabilidade. A integração de modelos estatísticos robustos no processo de decisão permite um melhor alinhamento entre produção, distribuição e procura real, reforçando a competitividade da empresa num setor em constante transformação.

Em síntese, este trabalho não apenas alcançou os objetivos inicialmente definidos, como também estabeleceu bases sólidas para investigações futuras, conciliando a vertente académica com uma aplicação prática e estratégica no setor industrial.

Bibliografia

[1] Grupo Nabeiro - História. Disponível em: *https://gruponabeiro.com/*. [Consultado em: 17 de abril de 2025].

[2] Delta Cafés. Disponível em: *https://deltacafes.com/*. [Consultado em: 17 de abril de 2025].

[3] CHOPRA, S. & MEINDL, P. *Supply Chain Management: Strategy, Planning, and Operation (7th ed.)*. *Pearson*, 2020.

[4] SAP SE.*SAP Integrated Business Planning for Supply Chain*. Disponível em: *https://www.sap.com/portugal/products/scm/integrated-business-planning.html*. [Consultado em: 17 de abril de 2025].

[5] SILVA, J. C., FIGUEIREDO, M. C., & BRAGA, A. C. *Demand Forecasting: A Case Study in the Food Industry*. Springer, 2019.

[6] GAKU, Rie. *Demand forecasting procedure for short life-cycle products with an actual food processing enterprise*. International Journal of Computational Intelligence Systems, 2014.

[7] NORTON, Sebastião *PREVISÃO DA PROCURA NA INDÚSTRIA ALIMENTAR*. Universidade de Coimbra, 2019.

[8] KUHN, M., & JOHNSON, K. *Feature Engineering and Selection: A Practical Approach for Predictive Models*. CRC Press, 2019.

[9] HYNDMAN, R. J., & ATHANASOPOULOS, G. *Forecasting: Principles and Practice (3rd ed.*. OTexts, 2021.

[10] BOX, G. E. P., JENKINS, G. M., REINSEL, G. C., & LJUNG, G. M. *Time Series Analysis: Forecasting and Control*. Wiley, 2015.

[11] SHUMAY, R. H., & STOFFER, D. S. *Time Series Analysis and Its Applications: With R Examples*. Springer, 2017.

[12] MONTGOMERY, D. C., JENNINGS, C. L., & KULAHCI, M. *Introduction to Time Series Analysis and Forecasting*. Wiley, 2015.

[13] MONTGOMERY, D.C., & RUNGER, G. C. *Applied Statistics and Probability for Engineers*. Wiley, 2018.

[14] CLEVELAND, R. B., CLEVELAND, W. S., MCRAE, J. E., & TERPENNING, I. J. *STL: A seasonal-trend decomposition procedure based on loess*. Journal of Official Statistics, 1990.

[15] MAKRIDAKIS, S., WHEELWRIGHT, S. S., & HYNDMAN, R. J. *Forecasting: methods and applications*. John Wiley & Sons: New York, 1998.

[16] HARRELL, F. E. P. *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Springer, 2015.

[17] BOX, G. E. P., & COX, D. R. *An analysis of transformations*. Journal of the Royal Statistical Society: Series B (Methodological), 1964.

[18] NELDER, J. A., & WEDDERBURN, R. W. M. *Generalized linear models*. Journal of the Royal Statistical Society: Series A (General), 1972.

[19] HASTIE, T., & TIBSHIRANI, R. *Generalized Additive Models*. Statistical Science, 1986.

[20] HAUKE, J., & KOSSOWSKI, T. *Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data*. Quaestiones Geographicae, 2011.

[21] MANN, H., & WHITNEY, D. *On a test of whether one of two random variables is stochastically larger than the other*. *Annals of Mathematical Statistics*, vol. 18, pp. 50–60, 1947.

[22] DICKEY, D., & FULLER, W. *Distribution of the estimators for autoregressive time series with a unit root*. Journal of the American Statistical Association, vol. 74, pp. 427–431, 1979.

[23] TIBSHIRANI, R., & FULLER, W. A. *Regression shrinkage and selection via the lasso*. Journal of the Royal Statistical Society: Series B (Methodological), 1996.

[24] AKAIKE, H. *A new look at the statistical model identification*. IEEE Transactions on Automatic Control, vol. 19, nº 6, pp. 716–723, 1974.

[25] HARVEY, A. C. *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press, 1990.

[26] Boylan, J. E., & SYNTETOS, A. A. *Accuracy and accuracy-implication metrics for intermittent demand*. Foresight: The International Journal of Applied Forecasting, 2006.

Anexo A

Resultados para o produto 150408

O produto selecionado pertence à família dos sumos.

Na Figura A.1 encontra-se representado o cronograma da procura semanal do produto **150408** na cidade de Lisboa, entre maio de 2022 e maio de 2025.

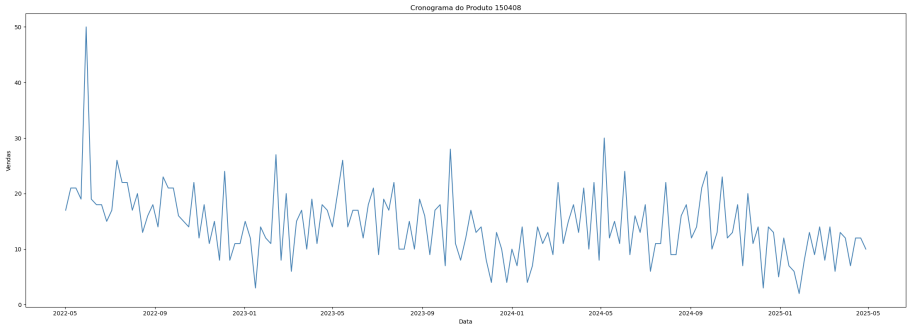

Figura A.1: Cronograma de **150408**.

A.1 Resultados da *Pipeline*

Finalizado o processo para todos os vinte e cinco *folds* da validação cruzada, obteve-se a média dos valores para as métricas de erro, tanto do modelo univariado como para os respectivos modelos multivariados.

As variáveis explicativas que mais vezes são selecionadas como variáveis preditoras são:

- X_1^{-3}, variável contínua, aparece em quatro *folds*;
- $X_1 5^{+2}$, variável contínua modelada por uma *spline*, aparece em vinte e um *folds*;

Na Tabela A.1 podem ser consultados os resultados da média das métricas de erro calculadas para o conjunto de teste dos vinte e cinco *folds*.

Modelos	MAE	MSE	RMSE	MAPE	WMAPE	SMAPE
Modelo Univariado	5.19	36.35	6.03	53.23%	42.38%	22.68%
SARIMAX	4.98	37.00	6.04	60.46%	38.87%	19.60%

Tabela A.1: Média das métricas de erro na amostra de teste utilizando a pipeline para **150408**.

Já a comparação gráfica entre os resultados obtidos pela *pipeline* e as previsões efetuadas pelo IBP podem ser encontradas na Figura A.2.

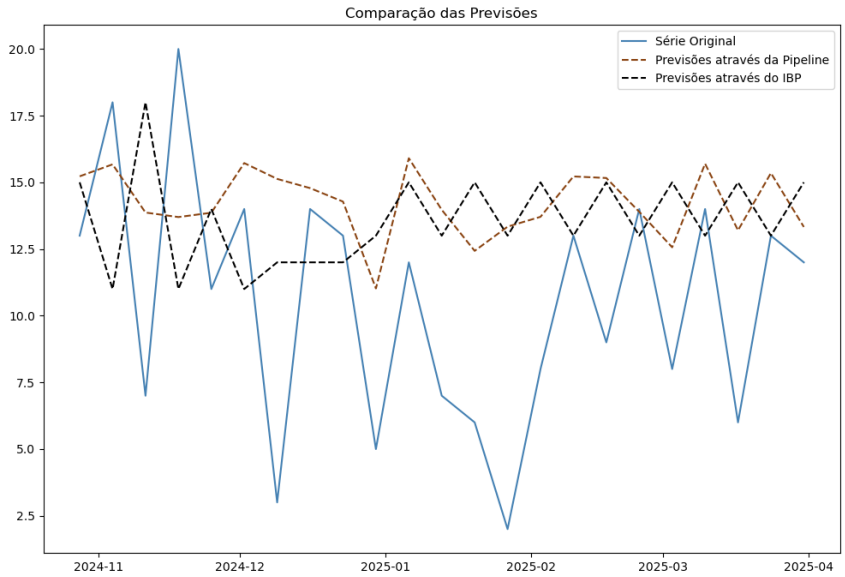

Figura A.2: Comparação entre previsões obtidas pela *pipeline* e pelo SAP IBP para **150408**.

Anexo B

Resultados para o produto 1053096

O produto selecionado pertence à família das cervejas.

Na Figura B.1 encontra-se representado o cronograma da procura semanal do produto **1053096** na cidade de Lisboa, entre maio de 2022 e maio de 2025.

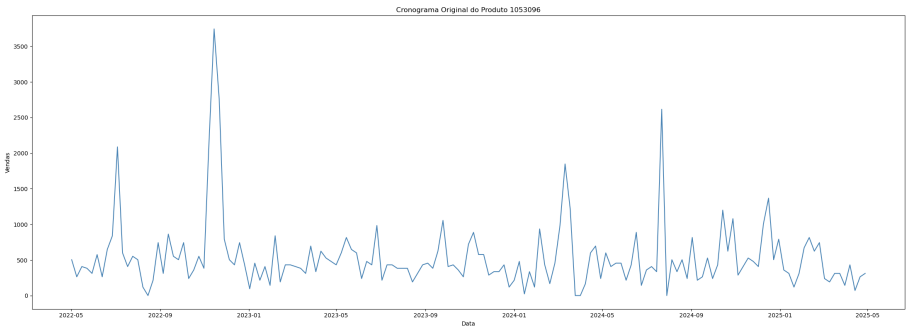

Figura B.1: Cronograma de **1053096**.

B.1 Resultados da *Pipeline*

Finalizado o processo para todos os vinte e cinco *folds* da validação cruzada, obteve-se a média dos valores para as métricas de erro, tanto do modelo univariado como para os respectivos modelos multivariados.

As variáveis explicativas que mais vezes são selecionadas como variáveis preditoras são:

- X_3^{+2}, variável contínua modelada por uma *spline*, aparece em dezassete *folds*;
- X_3^{+1}, variável contínua modelada por uma *spline*, aparece em seis *folds*;

Na Tabela B.1 podem ser consultados os resultados da média das métricas de erro calculadas para o conjunto de teste dos vinte e cinco *folds*.

Modelos	MAE	MSE	RMSE	MAPE	WMAPE	SMAPE
Modelo Univariado	301.40	204713	437.22	89.33%	43.15%	21.66%
SARIMAX	282.60	192173	430.247	101.69%	49.79%	25.69%

Tabela B.1: Média das métricas de erro na amostra de teste utilizando a pipeline para **1053096**.

Já a comparação gráfica entre os resultados obtidos pela *pipeline* e as previsões efetuadas pelo IBP podem ser encontradas na Figura B.2.

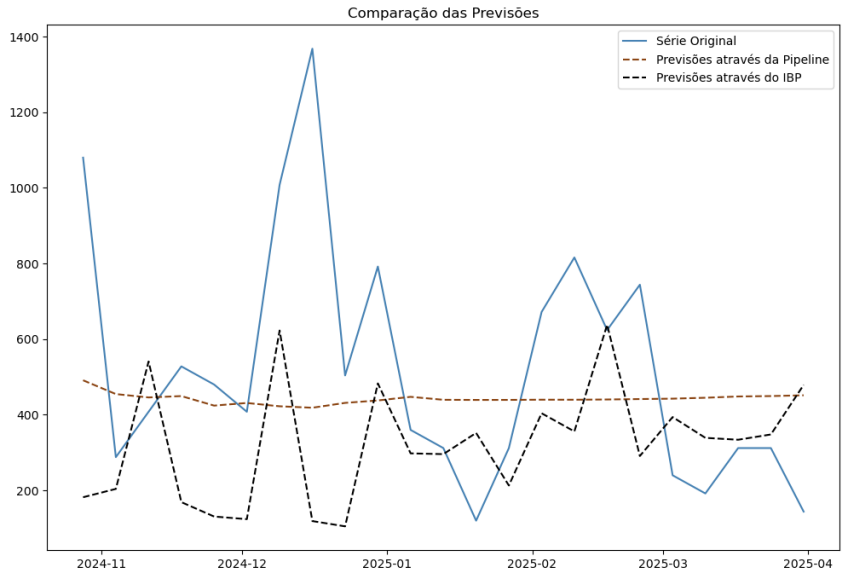

Figura B.2: Comparação entre previsões obtidas pela *pipeline* e pelo SAP IBP para **1053096**.

Anexo C

Resultados para o produto 1056050

O produto selecionado pertence à família dos sumos.

Na Figura C.1 encontra-se representado o cronograma da procura semanal do produto **1056050** na cidade de Lisboa, entre maio de 2022 e maio de 2025.

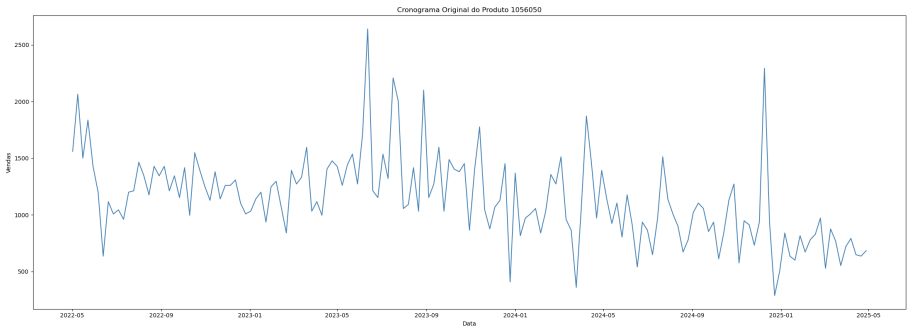

Figura C.1: Cronograma de **1056050**.

C.1 Resultados da *Pipeline*

Finalizado o processo para todos os vinte e cinco *folds* da validação cruzada, obteve-se a média dos valores para as métricas de erro, tanto do modelo univariado como para os respectivos modelos multivariados.

As variáveis explicativas que mais vezes são selecionadas como variáveis preditoras são:

- X_{15}^{+2}, variável contínua modelada por uma *spline*, aparece em vinte e quatro *folds*;
- X_{15}^{-2}, variável contínua modelada por uma *spline*, aparece em dezassete *folds*;

Na Tabela C.1 podem ser consultados os resultados da média das métricas de erro calculadas para o conjunto de teste dos vinte e cinco *folds*.

Modelos	MAE	MSE	RMSE	MAPE	WMAPE	SMAPE
Modelo Univariado	315.89	211273	429.28	41.29%	38.87%	22.62%
SARIMAX	343.36	211860	440.86	40.95%	38.28%	22.28%

Tabela C.1: Média das métricas de erro na amostra de teste utilizando a pipeline para **1056050**.

Já a comparação gráfica entre os resultados obtidos pela *pipeline* e as previsões efetuadas pelo IBP podem ser encontradas na Figura C.2.

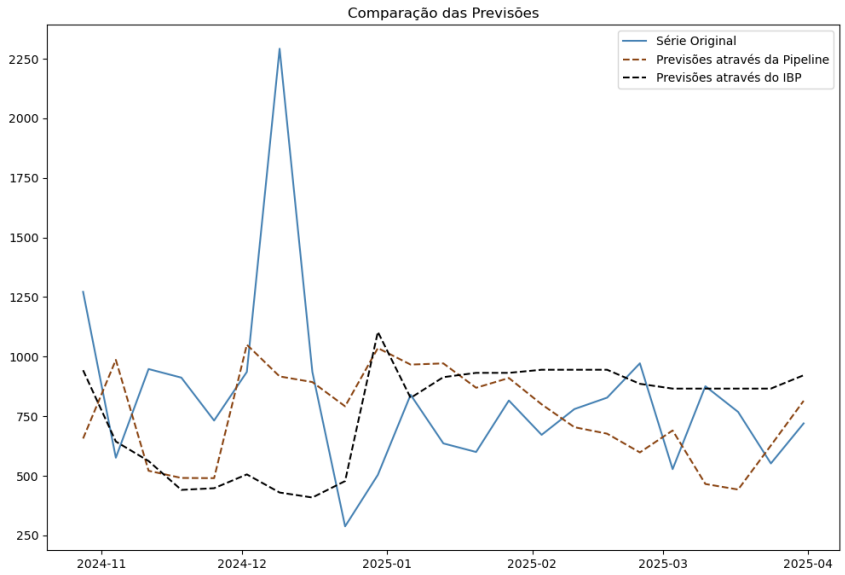

Figura C.2: Comparação entre previsões obtidas pela *pipeline* e pelo SAP IBP para **1056050**.

Anexo D

Resultados para o produto 5017052

O produto selecionado pertence à família dos cafés.

Na Figura D.1 encontra-se representado o cronograma da procura semanal do produto **5017052** na cidade de Lisboa, entre maio de 2022 e maio de 2025.

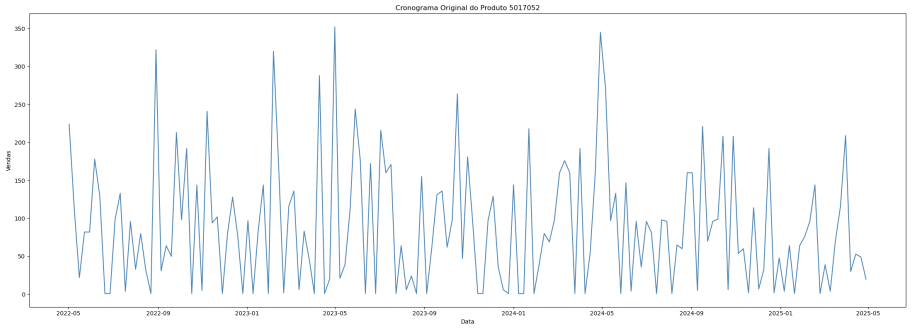

Figura D.1: Cronograma de **5017052**.

D.1 Resultados da *Pipeline*

Finalizado o processo para todos os vinte e cinco *folds* da validação cruzada, obteve-se a média dos valores para as métricas de erro, tanto do modelo univariado como para os respectivos modelos multivariados.

As variáveis explicativas que mais vezes são selecionadas como variáveis preditoras são:

- X_2, variável contínua, aparece em vinte e cinco *folds*;
- X_{14}^{+1}, variável contínua modelada por uma *spline*, aparece em vinte e cinco *folds*;
- X_{13}^{+2}, variável contínua modelada por uma *spline*, aparece em onze *folds*;

Na Tabela D.1 podem ser consultados os resultados da média das métricas de erro calculadas para o conjunto de teste dos vinte e cinco *folds*.

Modelos	MAE	MSE	RMSE	MAPE	WMAPE	SMAPE
Modelo Univariado	76.41	14686	108.54	127.11%	95.25%	46.39%
SARIMAX	57.59	4835	69.50	112.55%	73.15%	41.91%

Tabela D.1: Média das métricas de erro na amostra de teste utilizando a pipeline para **5017052**.

Já a comparação gráfica entre os resultados obtidos pela *pipeline* e as previsões efetuadas pelo IBP podem ser encontradas na Figura D.2.

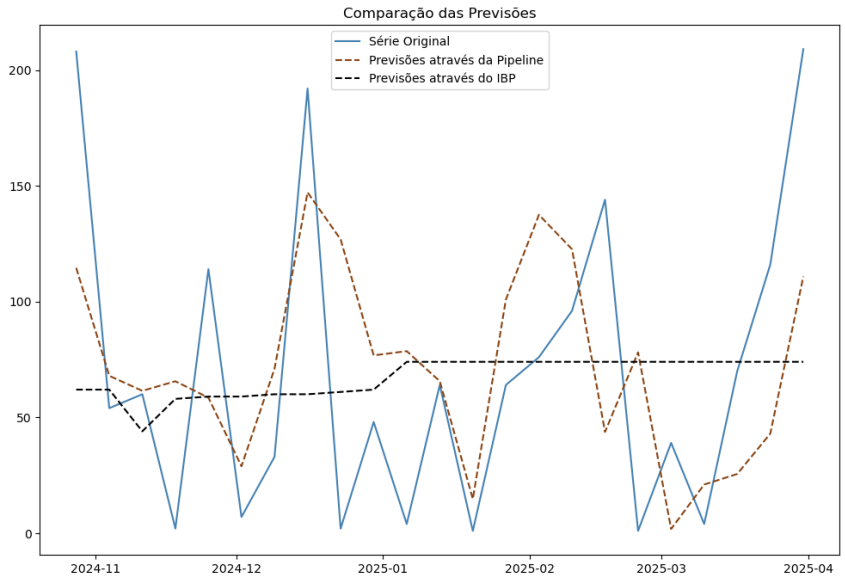

Figura D.2: Comparação entre previsões obtidas pela *pipeline* e pelo SAP IBP para **5017052**.