Traballo Fin de Máster

Aplicación de modelos de series temporais para a previsión de demanda e prezos nunha empresa de transporte internacional

Diego Airas Fernández

Máster en Técnicas Estatísticas

Curso 2025-2026

Proposta de Traballo Fin de Máster

Título en galego: Aplicación de modelos de series temporais para a previsión de demanda e prezos nunha empresa de transporte internacional

Título en español: Aplicación de modelos de series temporales para la previsión de demanda y precios en una empresa de transporte internacional

English title: Application of time series models for forecasting demand and prices in an international transport company

Modalidad: Modalidad B

Autor/a: Diego Airas Fernández, Universidade de Santiago de Compostela

Director/a: María José Ginzo Villamayor, Universidade de Santiago de Compostela; Jose Ameijeiras Alonso, Universidade de Santiago de Compostela

Tutor/a: Lucía Tajuelo López, Trucksters

Breve resumen del trabajo:

El objetivo principal será desarrollar modelos que permitan predecir, a corto y medio plazo, el volumen de demanda de los principales clientes, el cual está estrechamente relacionado con la estacionalidad del sector al que pertenecen. Además, se prestará especial atención a la predicción de la demanda en días festivos y periodos especiales, basándose en el comportamiento histórico de cada cliente. El trabajo comparará y evaluará distintas metodologías de series temporales, desde enfoques clásicos como ARIMA hasta métodos más recientes como modelos Prophet. Se buscará una solución que combine alta capacidad predictiva con facilidad de interpretación para su uso operativo por parte de los equipos de planificación.

Recomendaciones:

Buen dominio de Python y bibliotecas de análisis de datos como Pandas, Statsmodels, Scikit-learn y/o Prophet. Se valorarán conocimientos previos en modelos de series temporales, descomposición estacional y análisis exploratorio de datos. También será útil familiaridad con técnicas de validación cruzada para datos temporales y representación visual de resultados.

Otras observaciones:

Doña María José Ginzo Villamayor, Profesora Axudante Doutora de la Universidade de Santiago de Compostela, don Jose Ameijeiras Alonso, Titular de Universidade de la Universidade de Santiago de Compostela, doña Lucía Tajuelo López de Trucksters, informan que el Trabajo Fin de Máster titulado

Aplicación de modelos de series temporais para a previsión de demanda e prezos nunha empresa de transporte internacional

fue realizado bajo su dirección por don Diego Airas Fernández para el Máster en Técnicas Estadísticas. Estimando que el trabajo está terminado, dan su conformidad para su presentación y defensa ante un tribunal. Además, Doña María José Ginzo Villamayor, don Jose Ameijeiras Alonso y don Diego Airas Fernández

√ si □ no

autorizan a la publicación de la memoria en el repositorio de acceso público asociado al Máster en Técnicas Estadísticas.

En Santiago de Compostela, a 20 de enero de 2026.

La directora:
Doña María José Ginzo Villamayor
Firmado por GINZO VILLAMAYOR MARIA JOSE - ***4013** el día 22/01/2026

El director:
Don Jose Ameijeiras Alonso
Firmado por AMEIJEIRAS ALONSO JOSE - ***5023** el

La tutora:
Doña Lucía Tajuelo López
Signed by: 03F2DA7D09284E9...

El autor:
Don Diego Airas Fernández

Declaración responsable. Para dar cumplimiento a la Ley 3/2022, de 24 de febrero, de convivencia universitaria, referente al plagio en el Trabajo Fin de Máster (Artículo 11, Disposición 2978 del BOE núm. 48 de 2022), **el autor declara** que el Trabajo Fin de Máster presentado es un documento original en el que se han tenido en cuenta las siguientes consideraciones relativas al uso de material de apoyo desarrollado por otros/as autores/as:

- Todas las fuentes usadas para la elaboración de este trabajo han sido citadas convenientemente (libros, artículos, apuntes de profesorado, páginas web, programas,...)
- Cualquier contenido copiado o traducido textualmente se ha puesto entre comillas, citando su procedencia.
- Se ha hecho constar explícitamente cuando un capítulo, sección, demostración,... sea una adaptación casi literal de alguna fuente existente.

Y, acepta que, si se demostrara lo contrario, se le apliquen las medidas disciplinarias que correspondan.

Índice xeral

Resumo

Resumo en galego

Este Traballo de Fin de Máster presenta unha metodoloxía integral para a predición da demanda loxística na empresa Trucksters, aplicando técnicas de modelización de series temporais. O obxectivo principal é anticipar o volume de envíos nun mercado caracterizado pola súa alta volatilidade e dependencia de eventos estacionais.

A investigación desenvólvese baixo unha abordaxe comparativa entre a metodoloxía clásica ARIMA e os modelos *Prophet*, baseados nos modelos aditivos xeneralizados, co obxectivo de determinar que arquitectura captura con maior precisión a complexidade da demanda. Para tal fin, a estratexia de modelización articúlase mediante unha metodoloxía dual: por unha banda, contrástase a utilidade da segmentación estratéxica de clientes a través de clústeres e, por outra, avalíase o rendemento dos modelos sobre a base de datos orixinal fronte a unha serie depurada de atípicos. Esta análise permite demostrar que o tratamento previo da información e a modelización desagregada logran reducir o ruído estatístico e mitigar o impacto das anomalías. Como resultado, obténse un marco preditivo máis estable e robusto que optimiza a planificación operativa e a rendibilidade comercial de Trucksters.

English abstract

This Master's Thesis presents a comprehensive methodology for logistic demand forecasting at Trucksters, using advanced time-series modeling techniques. The central objective is to anticipate shipment volumes in a market defined by high volatility and a strong dependence on seasonal patterns.

The research is developed under a comparative approach between the classic ARIMA methodology and Prophet generalized additive models, in order to determine which architecture most accurately captures the complexity of demand. To this end, the modeling strategy is articulated through a dual methodology: on the one hand, the utility of strategic customer segmentation through clusters is contrasted, and on the other, the performance of the models is evaluated on the original database versus a series cleared of outliers. This analysis demonstrates that the prior treatment of information and disaggregated modeling reduce statistical noise and mitigate the impact of anomalies. As a result, a more stable and robust predictive framework is obtained that optimizes Trucksters operational planning and commercial profitability.

Prefacio

O sector do transporte e loxística constitúe a columna vertebral da economía actual, actuando como indicador da mobilidade de mercadorías e da integración dos mercados globais. Nun contexto económico cada vez máis competitivo, o desempeño eficiente das cadeas de subministración é determinante non só para a rendibilidade das empresas, senón tamén para a estabilidade do comercio internacional.

Non obstante, a xestión deste activo enfróntase a unha complexidade crecente. A demanda de servizos de transporte non é nin lineal nin estática; senón que se ve influenciada por unha multiplicidade de factores que abarcan desde a actividade macroeconómica e os patróns de comercio internacional ata a estacionalidade climática e os cambios no comportamento do consumidor final. Tal e como sinalan Basavaraju e Valilai (2025), a precisión na planificación da demanda impacta directamente nos niveis de inventario e no rendemento global dunha empresa. A dificultade radica en que, frecuentemente, as compañías carecen de información completa ou a tempo real sobre estes factores exóxenos, o que xera escenarios de incerteza que dificultan a toma de decisións e a optimización da cadea de subministración.

Estreitamente ligado á evolución da demanda, o comportamento dos prezos engade unha capa adicional de complexidade. As tarifas no sector loxístico presentan flutuacións significativas debido á dispoñibilidade de capacidade, os custos operativos (combustible, persoal) e a presión competitiva. Boin et al. (2020), destacan que unha estratexia de prezos dinámica e baseada en datos pode incrementar a marxe operativa dunha empresa loxística entre un 30 % e un 60 % . Ademais, a globalización introduciu novas variables de estacionalidade; por exemplo, informes recentes do sector destacan a necesidade de incorporar calendarios internacionais, como o Ano Novo chinés, para anticipar correctamente os picos de carga en Europa (Maersk , 2025), evidenciando que a previsión de prezos e demanda constitúe un desafío cada vez máis sofisticado.

Neste escenario global sitúase o caso de estudo deste traballo: a empresa **Trucksters**. Esta empresa emerxente irrompeu no sector cun modelo de negocio innovador baseado nun sistema de relevos de camións dirixido mediante intelixencia artificial. Un dos seus principais obxectivos consiste en reducir os tempos de tránsito de longa distancia e mellorar as condicións laborais dos condutores.

Para unha compañía desta natureza, a volatilidade non é só unha estatística, senón que constitúe un reto operativo diario. A planificación de rutas, a asignación de camións e condutores nos puntos de relevo e a xestión da capacidade dependen de forma crítica dunha anticipación precisa da carga de traballo. En Trucksters, errar na previsión implica ineficiencias directas: camións parados ou incapacidade para cubrir o servizo comprometido. Polo tanto, a modelización da demanda e dos prezos transcende o ámbito académico para converterse nunha necesidade operativa de primeira orde.

Outra característica de Trucksters radica na complexidade da demanda dos seus clientes. Dende unha perspectiva de negocio, pódese intuír que esta demanda depende de múltiples factores, como o sector de actividade do cliente ou o mes do ano (existencia de *peak seasons*). En consecuencia, unha das necesidades clave da empresa consiste en dispoñer dun método científico e reproducible que permita anticipar a demanda dos seus clientes con maior precisión.

Neste contexto, a literatura académica ofrece un amplo abanico de ferramentas para abordar este

problema, dende os modelos estatísticos clásicos ata enfoques máis recentes baseados en algoritmos de aprendizaxe automática. Se ben a utilidade teórica destas técnicas está amplamente demostrada, a súa aplicación práctica en contextos reais non sempre é trivial, tal e como demostran Revelo e Mafla (2025).

Este Traballo Fin de Máster sitúase na intersección entre a literatura académica e a práctica empresarial. O obxectivo principal consiste en aplicar, avaliar e comparar distintas metodoloxías de previsión, dende enfoques clásicos como os ARIMA ata métodos máis recentes como *Prophet* sobre os datos reais de Trucksters. Este tipo de metodoloxías son amplamente utilizadas na práctica, aplicándose a problemas similares ao abordado neste proxecto, como son os estudos de Ertürk (2024) e Wu et al. (2024). Búscase daquela, non só achegar evidencia empírica sobre o rendemento destes modelos, senón tamén xerar implicacións prácticas que sirvan de soporte para a toma de decisións estratéxicas na compañía.

Estrutura da memoria

A memoria organízase en dous bloques principais: o desenvolvemento metodolóxico e a aplicación práctica.

No bloque metodolóxico, o Capítulo 1 define en detalle o problema de negocio e o contexto do mercado. O Capítulo 2 establece o marco teórico, fundamentando as técnicas de análise multivariante, detección de atípicos e modelización de series temporais Box-Jenkins e *Prophet*, así como as ferramentas computacionais empregadas.

Na parte práctica, o Capítulo 3 aborda a análise exploratoria, a segmentación de clientes e a limpeza de datos. O Capítulo 4 constitúe o núcleo da investigación, desenvolvendo os modelos preditivos e analizando o impacto da eliminación de atípicos. Finalmente, o Capítulo 5 sintetiza as conclusións acadadas e propón liñas futuras de investigación para continuar mellorando a intelixencia loxística da empresa.

Parte I

Metodoloxía

Capítulo 1

O problema da demanda no sector do transporte

Este capítulo establece o marco operativo e de negocio sobre o que se fundamenta o presente Traballo Fin de Máster. O obxectivo consiste en trasladar a problemática xeral da predición de demanda á realidade concreta do transporte de mercadorías da empresa Trucksters.

Durante a última década, a loxística transcendeu a súa función tradicional de movemento de mercadorías para converterse nun ecosistema complexo gobernado pola información. A dixitalización do sector e a integración dos mercados globais incrementaron exponencialmente o volume de datos dispoñibles, pero tamén a volatilidade e a incerteza operativa. Neste escenario, a vantaxe competitiva das empresas xa non reside unicamente na capacidade de transporte, senón na capacidade analítica para anticipar as flutuacións do mercado.

A previsión de demanda constitúe, polo tanto, o desafío central. Non se trata unicamente dun exercicio estatístico, senón dunha necesidade crítica para paliar o impacto de factores exóxenos e garantir a eficiencia operativa.

Para abordar esta cuestión, o capítulo estrutúrase en tres bloques diferenciados. Na Sección 1.1 analizarase o contexto macroeconómico do transporte en Europa, achegando datos sobre a situación actual do mercado. Posteriormente, na Sección 1.2, presentarase o caso de estudo específico de Trucksters, detallando como o seu modelo de negocio baseado en relevos acentúa a necesidade dunha planificación precisa. Finalmente, na Sección 1.3 formalizaranse os obxectivos xerais e específicos que guiarán o desenvolvemento metodolóxico do traballo.

1.1. Contexto actual do sector

O transporte de mercadorías garante a fluidez das cadeas de subministración industriais e o abastecemento do consumo nunha economía altamente interconectada. Non obstante, a última década, e moi especialmente no período post-pandemia, produciuse un punto de inflexión na dinámica operativa do sector.

Historicamente, a loxística xestionábase baixo criterios de estabilidade e optimización de custos. Porén, tal e como analizan Boin et al. (2020), o sector experimentou unha transformación radical: a loxística deixou de ser un centro de custos para converterse nunha vantaxe competitiva e nun elemento de diferenciación estratéxica.

Tras a recuperación do sector posterior á pandemia sufrida no 2020, a realidade actual do mercado

europeo caracterízase pola escasez de condutores e o crecente número de empresas que quebran, o que aumenta os colos de botellas no propio mercado. Isto pode contrastarse mediante o barómetro de transporte de Timocom (2025), un indicador do mercado que mostra a situación actual de ofertas de cargas e camións no transporte por estrada europeo. Este barómetro é empregado polas empresas para averiguar os prezos do mercado e empregar dita información nas negociacións.

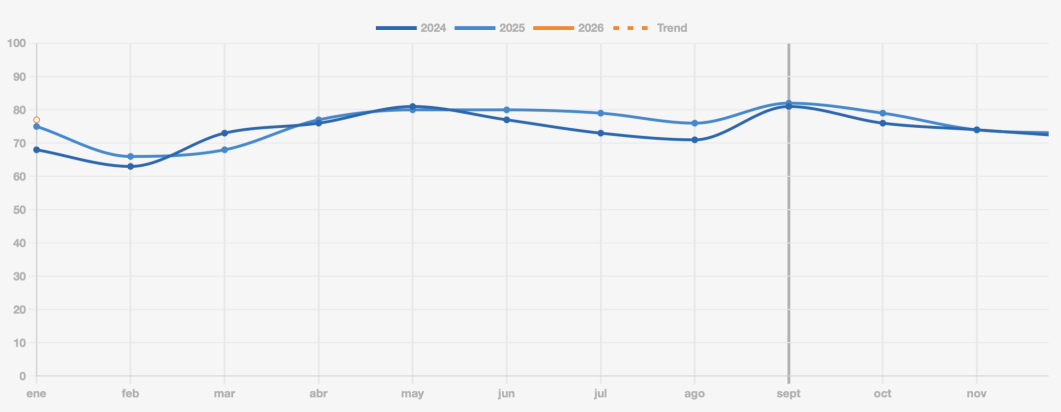


Figura 1.1: Evolución do ratio de ofertas de carga fronte ao espazo de camións no mercado europeo. A gráfica ilustra a proporción de flota ocupada durante os meses de 2024 (azul escuro) e 2025 (azul claro).

Fonte: Barómetro de Transporte de Timocom (2025).

A Figura 1.1 mostra a proporción de camións ocupados no mercado europeo durante os últimos dous anos. Esta gráfica representa como, a pesar de estar o mercado certamente estabilizado (apenas difire o comportamento entre 2024 e 2025), a demanda é extremadamente elevada, superándose apenas un 30 % da flota dispoñible nos meses de menor carga. Isto deriva nuns prezos cada vez máis caros, provocando que cada decisión a nivel loxístico sexa determinante no futuro dunha empresa.

1.2. Trucksters: Un modelo de negocio innovador

No contexto de mercado presentado sitúase Trucksters, empresa para a cal se realizou este estudo. Fundada como unha compañía tecnolóxica de transporte de mercadorías, Trucksters irrompeu no sector loxístico europeo cunha proposta de valor innovadora: a optimización do transporte de longa distancia mediante un sistema de relevos de condutores construído mediante intelixencia artificial.

A diferenza do modelo tradicional, onde un único condutor realiza a ruta completa (véndose obrigado a deter o camión para cumprir cos descansos regulamentarios), o modelo de Trucksters baséase na substitución do condutor en puntos estratéxicos ao longo do corredor. Isto permite que a mercadoría se manteña en movemento continuo, reducindo os tempos de tránsito. Simultaneamente, este sistema permite aos condutores regresar aos seus domicilios con maior frecuencia, mellorando substancialmente a súa calidade de vida e conciliación (véxase Trucksters , 2026).

Este modelo de negocio, experimentou un crecemento exponencial nos últimos anos. Porén, desde o punto de vista da análise de datos e da previsión de demanda, a natureza emerxente da empresa introduce unha serie de particularidades e desafíos que condicionan a metodoloxía deste traballo:

- Histórico temporal limitado: Ao tratarse dunha empresa de recente creación, as series temporais dispoñibles son curtas en comparación con operadores históricos. Isto limita a capacidade de capturar ciclos de longo prazo e require o uso de modelos eficientes en contextos de mostras reducidas.

- Cambios estruturais na tendencia: A serie histórica reflicte distintas etapas de madurez empresarial. Durante a súa curta historia, a empresa presenta fases de crecemento agresivo (captación de cota de mercado) seguidas de fases de consolidación e optimización de marxes (Capítulo 3). Estas variacións na estratexia comercial, que inclúen cambios na política de mercado para gañar competitividade, introducen dificultades estruturais na serie de demanda que os modelos estatísticos deben ser capaces de captar.
- Alta sensibilidade operativa: A natureza do sistema de relevos require unha sincronización case perfecta. Un erro na previsión de demanda non só implica perdas económicas, senón que pode desaxustar a rede de relevos, deixando condutores sen camión ou camións sen condutor en puntos intermedios da ruta.

Ademais destes factores internos, Trucksters opera cunha carteira de clientes extraordinariamente heteroxénea. Esta diversidade implica que o comportamento da demanda non é uniforme. Axustar un único modelo preditivo para o agregado total da compañía enmascararía as dinámicas particulares de cada sector, mentres que modelar cliente a cliente resultaría inviable ao ruído estatístico.

Unha vez coñecidos os contextos económicos do mercado, así como a natureza e política de negocio de Trucksters, o capítulo conclúe coa presentación formal dos obxectivos do proxecto.

1.3. Obxectivos do traballo

A meta fundamental do traballo consiste na identificación do modelo de previsión máis adecuado para Trucksters. Para iso, selecciónanse dous enfoques de modelización de series temporais: os modelos clásicos ARIMA e o modelo *Prophet*. A decisión de empregar estas metodoloxías xustifícase no seu amplo uso en estudos recentes, como son Ertürk (2024) ou Wu et al. (2024).

Estes modelos avaliaranse sobre datos reais da empresa, comparando o seu rendemento mediante métricas do erro axeitadas (Sección 2.4). Adicionalmente, ante a heteroxeneidade dos clientes cos que traballa Trucksters, no traballo analizarase se a aplicación dun procedemento previo de segmentación dos clientes permite mellorar a capacidade preditiva dos modelos. Deste xeito, o traballo non aporta só unha evidencia do desempeño de distintas técnicas de previsión, senón que tamén ofrece a Trucksters recomendacións prácticas sobre o procesamento de datos, facilitando así a toma de decisións baseada en analítica avanzada.

Polo tanto, no vindeiro capítulo abordaranse diversos conceptos de análise multivariante, así como de series temporais, asentando as bases teóricas do marco metodolóxico a seguir durante todo o proxecto.

Capítulo 2

Cuestións e fundamentos teóricos

Como xa se introduciu no capítulo anterior, os datos que son obxectivo neste proxecto presentan dependencia temporal. No ámbito da loxística e o transporte internacional, a análise de observacións illadas carece de valor estratéxico; a vantaxe competitiva reside na capacidade de modelar a dinámica temporal da demanda para anticipar temporadas de alto volume.

Neste capítulo desenvólvense os fundamentos teóricos e metodolóxicos necesarios para abordar esta problemática dende unha perspectiva da ciencia de datos. A estratexia proposta non se limita á modelización directa, senón que percorre un fluxo completo dende o preprocesado avanzado ata a avaliación preditiva.

A estrutura do capítulo organízase como segue:

Para resolver o problema da heteroxeneidade de clientes exposto no Capítulo 1, comezaranse abordando técnicas de análise multivariante na Sección 2.1. En particular, centrarase o estudo na análise de correspondencias (CA) e o *clustering*.

Na Sección 2.2 presentarase o algoritmo *Isolation Forest*. Esta ferramenta de *Machine Learning* empregarase no Capítulo 4 para a detección de *outliers*, permitindo un estudo dual durante o traballo: analizar os resultados sobre os datos orixinais fronte aos acadados nos datos depurados.

A continuación, a Sección 2.3 constitúe o núcleo da modelización temporal. Tras revisar os conceptos previos necesarios, estudaranse os modelos de Box-Jenkins (ARIMA), que permiten capturar a estrutura de autocorrelación lineal dos datos, e o modelo *Prophet*, baseado en modelos aditivos xeneralizados (GAM), caracterizados pola súa robustez ante estacionalidades múltiples e eventos exóxenos.

Finalmente, as Seccións 2.4 e 2.5 detallan, respectivamente, as métricas estatísticas para avaliar a precisión das predicións e o entorno computacional de Python empregado para a implementación práctica de todo o fluxo de traballo.

2.1. Técnicas de Análise Multivariante

A modelización de series temporais en contextos empresariais complexos enfróntase frecuentemente ao desafío da heteroxeneidade latente. Cando se analiza a demanda agregada dunha compañía, é habitual que a serie global sexa o resultado da superposición de múltiples compoñentes con dinámicas estocásticas moi distintas (diferentes estacionalidades, sensibilidade a prezos ou tendencias de crecemento). En tales escenarios, a construción dun modelo único para o total pode resultar ineficiente, xa que tende a suavizar os patróns específicos de cada subgrupo, perdendo precisión preditiva.

No contexto do presente traballo, a carteira de clientes de Trucksters componse de empresas pertencentes a sectores diversos (dende alimentación perecedeira ata *retail*) e suxeitas a tipoloxías contractuais distintas. Co obxectivo de mellorar a capacidade de xeneralización dos modelos de predición, faise necesaria unha estratexia de segmentación que permita agrupar os clientes en clústeres homoxéneos, proceso que se levará a cabo no Capítulo 4. A hipótese subxacente é que, ao modelar por separado grupos con comportamentos similares, reducirase a varianza intragrupo e maximizarase a precisión dos futuros modelos de series temporais considerados (Sección 2.3).

Para levar a cabo esta segmentación de forma rigorosa e non arbitraria, empregaranse técnicas de análise multivariante. Tendo presente a información dispoñible sobre os distintos envíos no Apéndice A, as variables fundamentais para caracterizar aos clientes son de natureza cualitativa ou categórica (tipo de contrato e sector da empresa). En consecuencia, as técnicas clásicas de agrupamento baseadas en distancias euclídeas non resultan aplicables de forma directa.

Por conseguinte, defínese unha metodoloxía en dúas etapas, secuencialmente dependentes, cuxos fundamentos teóricos se detallan na presente sección:

1. **Análise de Correspondencias (CA):** Técnica de redución de dimensións para transformar a información categórica nun espazo vectorial continuo.
2. **Técnicas de formación de grupos (*Clustering*)**: Técnica de clasificación non supervisada para identificar os grupos naturais de clientes sobre o espazo transformado.

A referencia fundamental seguida nesta sección é Pateiro e Sánchez (2024).

2.1.1. Análise de Correspondencias

O primeiro paso da análise multivariante que se aplicará na práctica (Sección 3.2) reside no estudo da relación entre dúas variables discretas (tipo de contrato e sector) mediante a súa distribución conxunta de frecuencias.

As táboas de frecuencias, tamén coñecidas como táboas de continxencia, reflicten a distribución conxunta do par formado polas dúas variables discretas, que se denotarán por (c, s) durante a presente sección. Partindo das frecuencias relativas, a táboa de continxencia pode expresarse como segue:

$c \setminus s$	s_1	s_2	$\cdots$	s_J	**Marginal** C
c_1	f_{11}	f_{12}	$\cdots$	f_{1J}	$f_{1\cdot}$
c_2	f_{21}	f_{22}	$\cdots$	f_{2J}	$f_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
c_I	f_{I1}	f_{I2}	$\cdots$	f_{IJ}	$f_{I\cdot}$
Marxinal S	$f_{\cdot 1}$	$f_{\cdot 2}$	$\cdots$	$f_{\cdot J}$	1

Cadro 2.1: Estrutura dunha táboa de continxencia para as frecuencias relativas conxuntas e marxinais de dúas variables categóricas c e s.

No Cadro 2.1 f_{ij} representa a frecuencia relativa do par (c_i, s_j), $(f_{1\cdot}, \dots, f_{I\cdot})$ a distribución marxinal da variable c e $(f_{\cdot 1}, \dots, f_{\cdot J})$ a correspondente distribución marxinal da variable s.

Test de independencia

Como xa se sinalou, a análise de correspondencias pretende estudar a relación entre dúas variables, sendo o caso extremo a independencia entre elas. Polo tanto, un paso previo ao CA consiste en verificar a

existencia dunha relación estatisticamente significativa entre as variables de interese. Con este obxectivo constrúese o estatístico chi-cadrado:

$$\chi^2 = n \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(f_{ij} - f_{i.} f_{.j})^2}{f_{i.} f_{.j}}, \tag{2.1}$$

sendo n o tamaño mostral. Baixo a hipótese nula de independencia, o estatístico χ^2 constrúese así e segue unha distribución chi-cadrado con $(I-1) \times (J-1)$ graos de liberdade. Así, fixado un nivel de significación do 5 %, rexeitariase a hipótese de independencia se

$$\chi^2 > \chi^2_{0.95,(I-1)\times(J-1)},$$

é dicir, se o estatístico calculado como se indicou na ecuación (2.1) supera o cuantil 0.95 da distribución chi-cadrado con $(I-1) \times (J-1)$ graos de liberdade. No caso de que non se rexeite a independencia das variables, a análise de correspondencias carecería de sentido.

Análise de compoñentes principais dos perfís estandarizados

Unha vez contrastada a existencia de relación entre as variables de interese, procédese a elaborar a CA. A idea deste procedemento reside en aplicar unha análise de compoñentes principais (PCA) sobre os perfís de fila estandarizados (ou columnas estandarizadas), os cales se definen a continuación.

Definición 2.1 *Dadas dúas variables categóricas* (c, s) *e as súas correspondentes distribucións marxinais* $(f_{1.}, \dots, f_{I.})$ *e* $(f_{.1}, \dots, f_{.J})$*, defínese o* ***perfil de fila i-ésimo*** *como*

$$r_i = \left(\frac{f_{i1}}{f_{i.}}, \cdots, \frac{f_{iJ}}{f_{i.}} \right).$$

Analogamente, defínese o ***perfil da columna j-ésima*** *como*

$$p_j = \left(\frac{f_{1j}}{f_{.j}}, \cdots, \frac{f_{Ij}}{f_{.j}} \right).$$

Así, o resultado fundamental da análise de correspondencias reside na obtención das coordenadas principais (ou *scores*) tanto para as filas como para as columnas. Estas puntuacións obtéñense proxectando os perfís orixinais sobre os autovectores da matriz de covarianzas. A continuación explicarase brevemente o precedemento da análise de compoñentes principais sobre os perfís de fila estandarizados, sendo o caso dos perfís de columna estandarizados análogo e podéndose consultar na referencia Pateiro e Sánchez (2024).

Sexa C unha matriz cuxas filas son os perfís de fila estandarizados. A análise de compoñentes principais (PCA) dos perfís de fila estandarizados consiste na diagonalización da matriz de covarianzas Σ_r dos perfís de fila estandarizados. Tendo en conta que ditos perfís suman 1 por columna, o número máximo de dimensións que aportan información será: mín$\{I-1, J-1\}$.

En particular, o estatístico χ^2 (ec. 2.1) dividido polo tamaño da mostra n denomínase **inercia** e presenta unha interpretación moi interesante. En efecto, denotando por λ_1 o autovalor nulo (pola perda dunha dimensión), cúmprese que

$$\frac{\chi^2}{n} = \text{traza}(\Sigma_r) = \lambda_2 + \ldots + \lambda_J, \tag{2.2}$$

sendo λ_k o k-ésimo autovalor da matriz de covarianzas Σ_r. Así, denotando por V a matriz cuxas columnas son os autovectores asociados aos autovalores calculados, as coordenadas das filas respecto das compoñentes principais virán dadas polas columnas da matriz $A = CV$.

Deste xeito, a análise de correspondencias (CA) actúa como unha técnica de cuantificación de datos categóricos. As puntuacións obtidas nos eixos que recollen a maior parte da inercia servirán como variables numéricas de entrada para o algoritmo de *clustering* que se describirá na Sección 2.1.2. Isto permite aplicar distancias métricas sobre unha información que, orixinalmente, carecía dunha estrutura de espazo vectorial, superando así a limitación de traballar con variables puramente cualitativas no Capítulo 3.

2.1.2. Análise Clúster

A análise clúster é unha técnica de clasificación non supervisada que busca formar grupos dentro dunha mostra en base a certas características da mesma. Dentro dos procedementos de formación de grupos distínguense varios tipos de métodos, centrándose esta sección nos métodos xerárquicos.

Este tipo de algoritmos parten dunha matriz de distancias entre os individuos, agrupándoos en base a estas distancias en distintos niveis. Daquela, para o problema do presente traballo, empregaranse as coordenadas calculadas mediante a análise de correspondencia (CA) explicada na sección previa como *inputs* para formar os grupos.

En particular, a construción dos clústeres levarase a cabo mediante un algoritmo aglomerativo. Estes tipos de algoritmos presentan os seguintes pasos:

1. Defínense os clústeres $C_1, \dots, C_n$ formados cada un por un único individuo.
2. Búscanse os grupos C_q e C_l que estén máis próximos, xuntándose e reducíndose, en consecuencia, o número de grupos.
3. Recalcúlanse as distancias de todos os demais grupos ao correspondente clúster formado.
4. O algoritmo detense cando só existe un único grupo, repetíndose en caso contrario os pasos 2 e 3.

Á vista da estrutura do algoritmo, un paso crucial reside na selección da distancia considerada entre os grupos. Neste proxecto, decidiuse traballar co método do máximo, que define a distancia entre dous grupos C_q e C_l como:

$$d(C_q, C_l) = \max_{i \in C_q, j \in C_l} d_{ij},$$

sendo d_{ij} a distancia entre os individuos i e j.

Cómpre sinalar que existen outras formas de definir as distancias entre clústeres, as cales se poden consultar na referencia Pateiro e Sánchez (2024).

O resultado do algoritmo xerárquico aglomerativo reflíctese visualmente a través dun dendrograma, como o que se presenta a modo de exemplo na Figura 2.1. Esta árbore de clasificación permite observar de forma intuitiva as sucesivas unións ou fusións realizadas polo algoritmo: no eixo horizontal dispóñense os individuos (ou categorías da análise de correspondencias), mentres que o eixo vertical representa a distancia á que se produce cada unión.

A selección do número óptimo de grupos (K) é un paso crítico que require equilibrar a parsimonia do modelo coa homoxeneidade interna dos grupos. Ao contrario que nos métodos de partición, no *clustering* xerárquico o número de grupos decídese a posteriori analizando a estrutura da árbore.

Na correspondente análise clúster que se realizará no Capítulo 3, na decisión do número de grupos a considerar seguirase un criterio que busca un equilibrio entre a calidade matemática e a interpretación de negocio. A finalidade é que cada clúster represente un segmento de mercado con necesidades e previsión de demanda diferenciadas. Daquela, por unha banda buscarase de forma visual un corte no correspondente dendrograma nun nivel de distancia onde exista un salto significativo entre unións. Un espazo vertical longo sen fusións indica que os grupos existentes nese nivel están ben diferenciados. Por

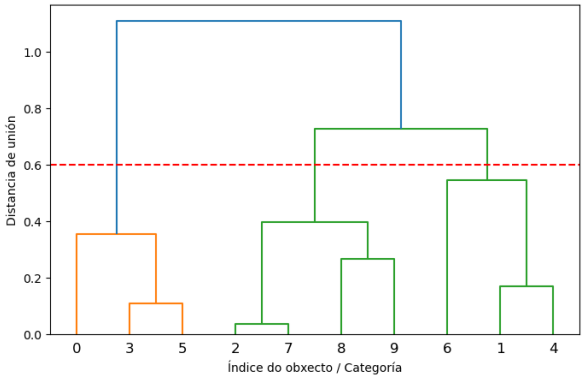


Figura 2.1: Dendrograma xerado mediante un exemplo de simulación de clasificación xerárquica aglomerativa empregando o método do máximo. O eixo horizontal identifica as categorías, mentres que o eixo vertical representa a distancia de disimilitude entre grupos. A liña descontinua marca un hipotético nivel de corte para a definición de $K = 3$ clústeres.

outra banda, buscarase que os grupos formados teñan certa coherencia a nivel de mercado, facilitando así a aplicación de estratexias de planificación diferenciadas en Trucksters.

Unha vez determinado o valor de K, aplícase a función de corte para asignar a cada categoría un identificador de clúster único. Estas etiquetas constitúen a base para a segmentación das series temporais de pedidos que se analizarán no Capítulo 4.

2.2. Detección de outliers: *Isolation Forest*

Unha vez segmentada a carteira de clientes, o seguinte paso crítico no preprocesado de datos é a identificación e tratamento de observacións atípicas ou *outliers*. Estas observacións representan datos con características estatisticamente diferentes do conxunto xeral de datos. No contexto deste traballo, o interese da análise dos *outliers* radica no estudo do seu impacto sobre os modelos axustados no Capítulo 4.

A literatura estatística clásica ofrece métodos baseados na distribución dos datos, como o *Z-score* ou o Rango Intercuartílico (IQR). Porén, estas técnicas presentan limitacións severas ante datos con alta asimetría, colas pesadas ou comportamentos multimodais, características que presentan as variables do presente proxecto (véxase Sección 3.3). Ante esta problemática, optouse pola aplicación dun algoritmo non paramétrico baseado en árbores de decisión: o ***Isolation Forest*** (IF), introducido por Liu et al. (2012).

Fundamento teórico do algoritmo

A diferenza da maioría de métodos de detección de anomalías que intentan modelar os puntos "normais" para identificar aqueles que se saen do patrón, o IF baséase na suposición de que os atípicos son poucos e diferentes.

A idea subxacente a este algortimo resulta bastante sinxela: separar as observacións mediante árbores de decisión e medir a profundidade media precisa para dividir cada variable. De forma máis explícita:

1. Contrúese un conxunto de árbores de decisión aleatorias (denominados *iTrees*). En cada nodo, a variable e o valor de corte considerados para dividir os datos elíxense de forma aleatoria.

2. Este procedemento repítese en cada árbore ata que todas as observacións queden illadas nun nodo da árbore cada unha.

3. Calcúlase a profundidade media para cada observación, entendéndose a profundidade como a distancia á raíz da árbore (a profundidade da raíz sería 0).

4. Para decidir se unha observación é un atípico, o algoritmo normaliza a profundidade media dende a raíz ata a folla nunha puntuación de anomalía s. Valores próximos a 1 indican unha alta probabilidade de ser un *outlier*, mentres que valores inferiores a 0.5 suxiren observacións normais.

A intuición matemática do modelo reside en que as observacións anómalas requiren de moi poucas particións para ser illadas, resultando nunha menor profundidade. Pola contra, os puntos normais, requiren un maior número de cortes e presentan profundidades significativamente máis longas.

Deste xeito, o *Isolation Forest* permite unha detección de *outliers* robusta, eficiente computacionalmente e independente da distribución estatística das variables consideradas.

2.3. Modelización de series temporais

Nesta sección explicaranse os conceptos correspondentes aos modelos de series temporais que se empregarán no Capítulo 4.

As referencias clave nesta sección serán: Aneiros (2024), Cryer e Chan (2008), Box et al. (2015) e Taylor e Letham (2018).

2.3.1. Conceptos previos

Antes de definir en que consiste unha serie temporal, é preciso introducir o concepto de proceso estocástico.

Definición 2.2 *Un* ***proceso estocástico*** *consiste nun conxunto de variables aleatorias, $\{Y_t\}_{t\in P}$, definidas sobre o mesmo espazo de probabilidade. O conxunto P denomínase* ***dominio temporal*** *e representa os intres nos que se observa o fenómeno.*

En particular, o traballo centrarase no proceso estocástico $\{Y_t\}_{t\in\mathbb{Z}}$, sendo $\mathbb{Z}$ o conxunto dos números enteiros. Así, o subíndice t indica o instante de tempo na que se observa.

A partires dun proceso estocástico, é posible definir as series de tempo.

Definición 2.3 *Unha* ***serie de tempo*** *trátase dunha realización dun proceso estocástico e denotarase por*

$$\{y_1, \dots, y_T\},\ T \in \mathbb{Z}.$$

Na práctica, pártese do coñecemento dunha serie temporal e trátase de atopar o proceso estocástico que a xerou. Porén, isto non é sempre posible. Por iso, a forma de actuar é construír un proceso estocástico "básico" que poida xerar unha serie de tempo das características da serie obxecto de estudo.

Tal e como se explica no Capítulo 2 de Cryer e Chan (2008), para poder facer inferencia sobre a estrutura dun proceso estocástico, é preciso facer algunhas suposicións sobre a serie temporal. A máis importante é o concepto de estacionariedade. A idea básica deste concepto reside en que as leis de probabilidade que dominan o comportamento da serie non cambian ao longo do tempo. Dunha forma máis formal, introdúcense a continuación os conceptos de estacionariedade estrita e estacionariedade débil.

Definición 2.4 *Defínese un proceso estocástico* $\{Y_t\}_{t\in\mathbb{Z}}$ *como* ***estritamente estacionario*** *se, dados* $t_1, \dots, t_n$ *momentos e un retardo* k*, a distribución conxunta de* $\{Y_{t_1}, \dots, Y_{t_n}\}$ *é a mesma que a distribución de* $\{Y_{t_1-k}, \dots, Y_{t_n-k}\}$*.*

Como esta suposición é difícil de verificar, defínese unha condición similar pero menos restritiva.

Definición 2.5 *Dirase que un proceso estocástico* $\{Y_t\}_{t\in\mathbb{Z}}$ *é* ***debilmente estacionario*** *(ou* ***estacionario de segunda orde****) se se cumpren as seguintes condicións:*

1. $\mathbb{E}(Y_t) = \mu_t = \mu, \ \forall\ t \in \mathbb{Z}$.
2. $Cov(Y_t, Y_{t+k}) = \gamma(t, t+k) = \gamma_k, \ \forall\ t, k \in \mathbb{Z}$.

A condición de estacionariedade débil consiste en que a media sexa constante ao longo do tempo e a autocovarianza dependa exclusivamente do retardo k. En particular, como consecuencia da segunda propiedade, a varianza de cada variable Y_t non depende do intre t.

Durante o presente traballo, os termos serie estacionaria e proceso estacionario farán referencia a esta segunda definición.

A partires das defincións previas, introdúcese agora un exemplo de proceso estacionario que xoga un papel fundamental na construción do resto de procesos e modelos que se estudarán.

Definición 2.6 *Sexa unha secuencia de variables aleatorias* $\{a_t\}_{t\in\mathbb{Z}}$ *incorreladas, con media 0 e varianza constante finita* σ^2*. Dirase que o proceso formado por estas variables aleatorias é de* ***ruído branco****.*

Observación 2.1 *Na definición anterior, se o proceso segue ademais unha distribución gaussiana, entón as variables aleatorias que o conforman son independentes e identicamente distribuídas (i.i.d.).*

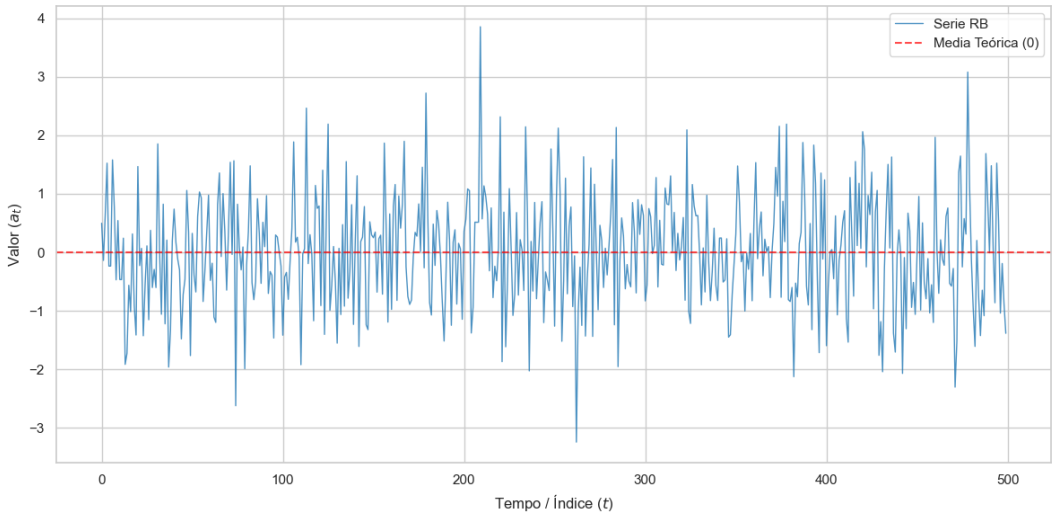


Figura 2.2: Representación dun proceso de ruído branco gaussiano con media nula e varianza unitaria ($\mu = 0$ e $\sigma^2 = 1$).

Dende unha perspectiva visual, unha realización de ruído branco caracterízase por unha oscilación puramente aleatoria arredor de cero, sen presentar ningunha tendencia nin ciclo recoñecible (ver Figura 2.2). Esta ausencia de patróns implica que o valor pasado do proceso non aporta ningunha información para predicir o valor futuro, sendo este o obxectivo ideal para os residuos de calquera modelo de predición que se axuste no Capítulo 4.

Introdúcense a continuación un par de conceptos que serán claves na identificación do modelo: as funcións de autocorrelacións.

Definición 2.7 *Defínese a* ***función de autocorrelacións simples*** *(fas) como segue*

$$\rho(s,t) = \frac{\gamma(s,t)}{\sigma_s \sigma_t},$$

onde $\gamma(s,t) = Cov(Y_s, Y_t)$ *é a covarianza entre as observacións nos momentos* s *e* t*.*

É importante sinalar que a *fas* é unha medida da dependencia lineal existente entre Y_s e Y_t e toma valores no intervalo $[-1, 1]$. Na práctica (Capítulo 4), o seu gráfico ou **correlograma** permitirá identificar a necesidade de aplicar diferenzas para acadar a estacionariedade e determinar a orde de compoñentes de media móbil (MA).

Definición 2.8 *Defínese a* ***función de autocorrelacións parciais*** *(fap) da seguinte forma*

$$\alpha(s,t) = \frac{Cov\left(Y_s - \hat{Y}_s^{(s,t)}, Y_t - \hat{Y}_t^{(s,t)}\right)}{\sqrt{Var\left(Y_s - \hat{Y}_s^{(s,t)}\right) Var\left(Y_t - \hat{Y}_t^{(s,t)}\right)}},$$

onde $\hat{Y}_j^{(s,t)}$ *é o mellor preditor lineal de* Y_j *construído a partires das variables medidas nos intres entre* s *e* t *sen incluír os extremos.*

Ao igual que a *fas*, a *fap* toma valores no intervalo $[-1, 1]$. A súa utilidade reside en que permite illar a dependencia directa entre dous instantes, o que resultará fundamental no presente traballo para identificar a orde dos procesos autorregresivos (AR).

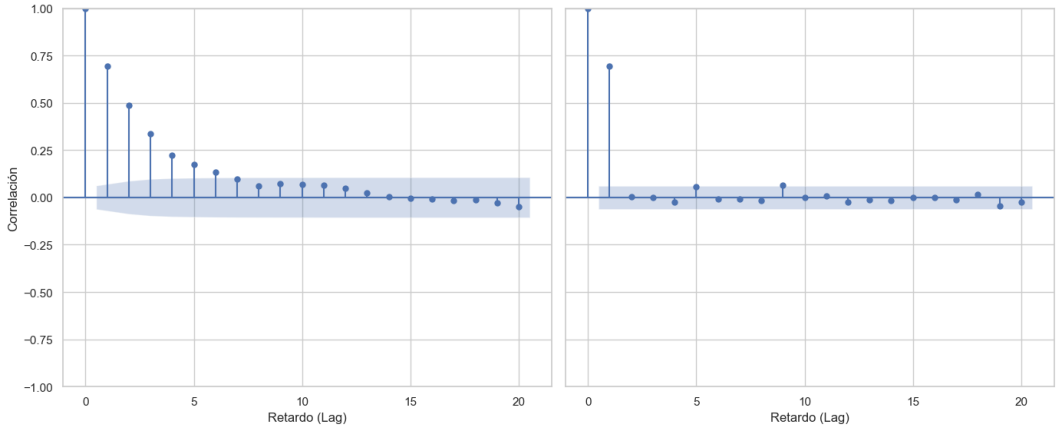


Figura 2.3: Correlogramas *fas* (esquerda) e *fap* (dereita) para un proceso autorregresivo AR(1) con parámetro $\phi_1 = 0.7$. Na *fas* obsérvase un decaemento positivo, característico dun proceso con forte persistencia ou memoria de curto prazo. Pola contra, a *fap* mostra un único pico significativo no primeiro retardo e un corte brusco cara a cero nos restantes, o que permite identificar a orde $p = 1$ do modelo autorregresivo.

Na Figura 2.3 represéntanse a modo de exemplo os correlogramas da *fas* e *fap* mostrais para un proceso autorregresivo de orde 1 (ver Sección 2.3.2).

Para rematar esta sección, preséntanse unha serie de propiedades de interese dun proceso estocástico.

Definición 2.9 *Dirase que un proceso estocástico, $\{Y_t\}_{t\in\mathbb{Z}}$, é un proceso* ***lineal****, se admite unha representación do tipo*

$$Y_t = c + \sum_{i=-\infty}^{+\infty} \psi_i a_{t-i}, \text{ onde } \sum_{i=-\infty}^{+\infty} |\psi_i| < \infty,$$

sendo $\{a_t\}_{t\in\mathbb{Z}}$ un proceso de ruído branco e os coeficientes ψ_i determinan a influencia de cada innovación pasada e presente sobre o valor actual do proceso.

Sinálese que todo proceso lineal é estacionario, onde a súa función de autocovarianzas será:

$$\gamma_k = \sigma_a^2 \sum_{i=-\infty}^{+\infty} \psi_i \psi_{i+k}.$$

Definición 2.10 *Un proceso estocástico $\{Y_t\}_{t\in\mathbb{Z}}$ dise* ***causal*** *(MA(∞)) se admite unha representación do tipo*

$$Y_t = c + \sum_{i=0}^{+\infty} \psi_i a_{t-i}, \text{ con} \sum_{i=0}^{+\infty} |\psi_i| < \infty,$$

con $\{a_t\}_{t\in\mathbb{Z}}$ ruído branco e os coeficientes ψ_i determinan a influencia de cada innovación pasada e presente sobre o valor actual do proceso.

Definición 2.11 *Un proceso estocástico $\{Y_t\}_{t\in\mathbb{Z}}$ será* ***invertible*** *(AR(∞)) se admite unha representación da forma*

$$Y_t = c + a_t + \sum_{i=1}^{+\infty} \pi_i Y_{t-i}, \text{ con} \sum_{i=1}^{+\infty} |\pi_i| < \infty,$$

onde $\{a_t\}_{t\in\mathbb{Z}}$ é ruído branco e os coeficientes π_i representan os pesos asignados ás observacións pasadas para explicar o valor da actual.

Unha vez introducidas estas definicións, presentarase un resultado fundamental que garante que calquera proceso estacionario ou ben é lineal ou ben se pode transformar para que o sexa.

Teorema 2.1 *Se un proceso estocástico $\{Y_t\}_{t\in\mathbb{Z}}$ é estacionario e non ten compoñentes deterministas, entón admite unha representación do tipo*

$$Y_t = \sum_{i=0}^{+\infty} \psi_i a_{t-i}, \text{ con } \psi_0 = 1 \text{ e } \sum_{i=0}^{+\infty} \psi_i^2 < \infty,$$

onde $\{a_t\}_{t\in\mathbb{Z}}$ é ruído branco e os coeficientes ψ_i determinan a influencia de cada innovación pasada e presente sobre o valor actual do proceso.

Daquela, nesta sección presentáronse os fundamentos básicos no contexto da análise das series temporais. A continuación, introducirase unha aproximación clásica ao estudo deste tipo de procesos.

2.3.2. Modelos de Box-Jenkins

Como se indicou na sección previa, atopar o proceso estocástico xerador dunha serie de tempo non é unha tarea fácil. Nesta sección tratarase de construír un modelo estocástico sinxelo que poida xerar a serie de tempo dispoñible. Para iso, empregarase a metodoloxía Box-Jenkins, consistente en tres etapas: identificación e selección do modelo xerador da serie, estimación dos parámetros do modelo seleccionado e validación ou diagnose das hipóteses do mesmo.

Fundamentalmente nesta sección seguirase a estrutura da referencia Aneiros (2024).

Procesos ARIMA(p,d,q)

Para comezar, defínense posibles modelos para series estacionarias.

Definición 2.12 *Un* ***proceso autorregresivo*** *de orde p, AR(p), é un proceso estocástico* $\{Y_t\}_{t\in\mathbb{Z}}$ *que admite unha represetación do tipo*

$$Y_t = c + \phi_1 Y_{t-1} + ... + \phi_p Y_{t-p} + a_t, \tag{2.3}$$

sendo $c, \phi_1, ..., \phi_p$ *constantes con* $\phi_p \neq 0$ *e* $\{a_t\}_{t\in\mathbb{Z}}$ *un proceso de ruído branco.*

É preciso sinalar que a representación dada por (ec. 2.3) dá lugar a un proceso estacionario se, e só se, o polinomio característico $\phi(z) = 1 - \phi_1 z - ... - \phi_p z^p$ non ten raíces de módulo 1. En particular, será un proceso causal se tampouco ten raíces de módulo menor ca 1. Ademais, sempre é invertible.

Definición 2.13 *Un* ***proceso de médias móbiles*** *de orde q, MA(q), é un proceso estocástico* $\{Y_t\}_{t\in\mathbb{Z}}$ *que admite unha representación do tipo*

$$Y_t = c + a_t + \theta_1 a_{t-1} + ... + \theta_q a_{t-q}, \tag{2.4}$$

con $c, \theta_1, ..., \theta_q$ *constantes,* $\theta_q \neq 0$ *e* $\{a_t\}_{t\in\mathbb{Z}}$ *un proceso de ruído branco.*

Para este tipo de procesos cúmprese que sempre son estacionarios e causais, mentres que serán invertibles se, e só se o polinomio característico $\theta(z) = 1 + \theta_1 z + ... + \theta_q z^q$ non ten raíces de módulo menor ou igual ca 1.

Agora ben, se nun mesmo proceso existe estrutura autorregresiva (AR) e de médias móbiles (MA), xorden os modelos ARMA que se definen a continuación.

Definición 2.14 *Un* ***proceso ARMA*** *de ordes p e q, ARMA(p, q) é un proceso estocástico* $\{Y_t\}_{t\in\mathbb{Z}}$ *que admite a representación*

$$\begin{aligned} Y_t = c &+ \phi_1 Y_{t-1} + ... + \phi_p Y_{t-p} \\ &+ a_t + \theta_1 a_{t-1} + ... + \theta_q a_{t-q}, \end{aligned} \tag{2.5}$$

con $c, \phi_1, ..., \phi_p, \theta_1, ..., \theta_q$ $(\phi_p \neq 0, \theta_q \neq 0)$ *constantes e* $\{a_t\}_{t\in\mathbb{Z}}$ *un proceso de ruído branco.*

Observación 2.2 *Considerando o operador retardo B tal que*

$$BY_t = Y_{t-1},$$

pódese reescribir de forma máis compacta o modelo (2.5) da seguinte forma

$$\phi(B)Y_t = c + \theta(B)a_t,$$

sendo

$$\phi(B) = 1 - \phi_1 B - ... - \phi_p B^p \text{ e } \theta(B) = 1 + \theta_1 B + ... + \theta_q B^q.$$

Tendo en conta as propiedades dos procesos autorregresivos e de medias móbiles, obsérvase que a representación anterior dá lugar a un proceso estacionario se, e só se, o polinomio $\phi(z)$ non ten raíces de módulo 1. Será, en efecto, causal se dito polinomio tampouco ten raíces con módulo menor ca 1. Ademais, será invertible se o polinomio $\theta(z)$ tampouco ten raíces de módulo menor ou igual a 1.

Os modelos ARMA presentados até o momento constitúen unha familia de procesos estacionarios. Porén, na práctica non é habitual atoparse con series de tempo xeradas por procesos constantes en media e varianza, en especial ao traballar con series de demanda como as que se estudarán na práctica (Capítulo 4). A solución ante esta situación reside en transformar a serie para ter a estacionariedade e posteriormente axustar un modelo ARMA.

Unha das principais causas da falta de estacionariedade dunha serie é a presenza de tendencia no nivel da serie, isto é, existe unha relación entre o nivel medio da mesma e o tempo. Para enfrontar este problema, proponse diferenciar a serie de forma regular.

Definición 2.15 *Dada unha serie temporal* $\{Y_t\}_{t\in\mathbb{Z}}$*, defínese a* ***primeira diferenza regular*** *(*$d = 1$*) como o proceso* $\{W_t\}_{t\in\mathbb{Z}}$ *obtido mediante:*

$$W_t = Y_t - Y_{t-1},\ \forall t \in \mathbb{Z}.$$

En particular, se unha vez diferenciada a serie regularmente segue observándose tendencia, repetirase o proceso sobre a serie diferenciada (en xeral, é suficiente con $d \leq 3$).

Unha vez presentada esta idea, xorden un novo tipo de modelos coñecidos como ARIMA, que consisten en aplicar d diferenzas regulares á serie de tempo orixinal e posteriormente axustar un modelo ARMA sobre a serie estacionaria.

Definición 2.16 *Un* ***proceso ARIMA*** *de ordes p, d e q, ARIMA(p,d,q) é un proceso estocástico* $\{Y_t\}_{t\in\mathbb{Z}}$ *que admite unha representación do tipo*

$$\phi(B)(1-B)^d Y_t = c + \theta(B)a_t,$$

onde c é unha constante, $\phi(B)$ *é o polinomio correspondente á parte autorregresiva,* $\theta(B)$ *o polinomio correspondente á parte de medias móbiles e* $\{a_t\}_{t\in\mathbb{Z}}$ *un proceso de ruído branco. Ademais,* $\phi(z)$ *non ten raíces de módulo 1.*

Observación 2.3 *Destáquese que un proceso* $\{Y_t\}_{t\in\mathbb{Z}}$ *é ARIMA(*p, d, q*) se, e só se,* $\{(1-B)^d Y_t\}_{t\in\mathbb{Z}}$ *é un proceso ARMA(*p, q*).*

A identificación das ordes óptimas (p, d, q) para as series de Trucksters realizarase seguindo criterios de parsimonia e minimización de indicadores como o AIC, proceso que se detallará máis adiante na presente sección.

Procesos ARIMA estacionais

Os procesos ARIMA(p, d, q) estudados no apartado anterior son capaces de capturar a falta de estacionariedade causada pola presenza de tendencia. Non obstante, non recollen a ausencia de estacionariedade debida á presenza de compoñente estacional, isto é, a presenza de patróns repetitivos. Ditos patróns obsérvanse graficamente tanto no comportamento da serie coma nos correspondentes correlogramas (*fas* e *fap*) cunha frecuencia ou **período** s. Isto motiva o estudo de novos procesos: os modelos ARMA estacionais.

Definición 2.17 *Defínese un* ***proceso ARMA estacional*** *de ordes P e Q e con período estacional s como un proceso estocástico* $\{Y_t\}_{t\in\mathbb{Z}}$ *que admite unha representación da forma*

$$\begin{aligned} Y_t = {} & c + \Phi_1 Y_{t-s} + \ldots + \Phi_P Y_{t-Ps} \\ & + a_t + \Theta_1 a_{t-s} + \ldots + \Theta_Q a_{t-Qs}, \end{aligned} \tag{2.6}$$

onde $c, \Phi_1, \ldots, \Phi_P, \Theta_1, \ldots, \Theta_Q$ *son constantes tales que* $\Phi_P \neq 0, \Theta_Q \neq 0$ *e* $\{a_t\}_{t\in\mathbb{Z}}$ *é un proceso de ruído branco.*

Observación 2.4 *Cómpre sinalar varias consideracións sobre os procesos ARMA estacionais:*

- *De forma análoga aos procesos ARMA, a representación anterior (ec. 2.6) pode reescribirse da seguinte forma*

  $$\Phi(B^s)Y_t = c + \Theta(B^s)a_t,$$

  sendo agora

  $$\Phi(B^s) = 1 - \Phi_1 B^s - \ldots - \Phi_P B^{Ps} \text{ e } \Theta(B^s) = 1 + \Theta_1 B^s + \ldots + \Theta_Q B^{Qs},$$

  e B^s *representa o operador retardo estacional,* $B^s Y_t = Y_{t-s}$.

- *Os procesos $ARMA(P,Q)_s$ son, en particular, procesos $ARMA(sP,sQ)$ con moitos coeficientes nulos, polo cal as condicións de causalidade, invertibilidade e estacionariedade derivan directamente das propiedades xa expostas para os procesos ARMA.*

Os procesos ARMA estacionais así definidos modelan a dependencia estacional, é dicir, a relación entre Y_t e as observacións sucedidas en intres de tempo separados por múltiplos do período estacional s.

Agora ben, é posible que nun mesmo proceso existan dependencia regular e estacional. Así, combinando os modelos ARMA(p,q) e os modelos ARMA$(P,Q)_s$ xorden os modelos ARMA estacionais multiplicativos.

Definición 2.18 *Un **proceso ARMA estacional multiplicativo** de ordes p, q, P e Q e período estacional s, $ARMA(p,q)\times(P,Q)_s$, é un proceso estocástico $\{Y_t\}_{t\in\mathbb{Z}}$ que admite a seguinte representación*

$$\phi(B)\Phi(B^s)Y_t = c + \theta(B)\Theta(B^s)a_t, \tag{2.7}$$

sendo c unha constante, $\phi(B)$ e $\Phi(B^s)$ os correspondentes polinomios autorregresivos, $\theta(B)$ e $\Theta(B^s)$ os polinomios de medias móbiles e $\{a_t\}_{t\in\mathbb{Z}}$ un proceso de ruído branco.

Observación 2.5 *Unha forma análoga de construír os modelos anteriores sería a partires dun ARMA non estacional onde $\{a_t\}_{t\in\mathbb{Z}}$ en lugar de ser ruído branco é un ARMA estacional.*

Como se indicou previamente, é posible que a serie non sexa estacionaria como consecuencia da presenza de compoñente estacional. Cando unha serie presenta dependencia estacional e o nivel medio da serie depende de dito período estacional, será preciso diferenciar a serie, pero neste caso de forma estacional.

Definición 2.19 *Dada unha serie temporal $\{Y_t\}_{t\in\mathbb{Z}}$ con dependencia estacional de período s, defínese a **primeira diferenza estacional** ($D=1$) como o proceso $\{V_t\}_{t\in\mathbb{Z}}$ obtido mediante:*

$$V_t = Y_t - Y_{t-s},\ \forall t\in\mathbb{Z}.$$

Permitindo así introducir os procesos ARIMA estacionais.

Definición 2.20 *Un **proceso ARIMA estacional** de ordes P, D, Q e con período estacional s, $ARIMA(P,D,Q)_s$ é un proceso estocástico $\{Y_t\}_{t\in\mathbb{Z}}$ que admite a seguinte representación*

$$\Phi(B^s)(1-B^s)^D Y_t = c + \Theta(B^s)a_t, \tag{2.8}$$

sendo c unha constante, $\Phi(B^s)$ o polinomio autorregresivo, $\Theta(B^s)$ o polinomio de medias móbiles e $\{a_t\}_{t\in\mathbb{Z}}$ un proceso de ruído branco.

Recompilando todas as ideas descritas ata o momento, constrúense os modelos ARIMA estacionais multiplicativos, a partir dos cales se poden axustar unha gran parte das series de tempo como as do presente traballo. Ademais, ditos procesos inclúen todos os anteriores presentados como casos particulares.

Definición 2.21 *Un **proceso ARIMA estacional multiplicativo** de ordes p, d, q, P, D, Q e período estacional s, $ARIMA(p,d,q)\times(P,D,Q)_s$ é un proceso estocástico $\{Y_t\}_{t\in\mathbb{Z}}$ que se pode escribir como segue*

$$\phi(B)\Phi(B^s)(1-B)^d(1-B^s)^D Y_t = c + \theta(B)\Theta(B^s)a_t, \tag{2.9}$$

onde c é unha constante, $\phi(B)$ e $\Phi(B^s)$ son os correspondentes polinomios autorregresivos, $\theta(B)$ e $\Theta(B^s)$ os de medias móbiles e $\{a_t\}_{t\in\mathbb{Z}}$ un proceso de ruído branco.

Así, a forma de proceder ante unha serie non estacionaria pola presenza de tendencia na práctica (Capítulo 4) será a seguinte:

- Comézase eliminando (en caso de existir) a tendencia aplicando d diferenzas regulares.
- Posteriormente, elimínase (se é preciso) a compoñente estacional aplicando D diferenzas estacionais. En xeral, é suficiente con $D = 1$.
- Unha vez que a serie diferenciada é estacionaria, modelizarase a través dun modelo ARMA como os presentados (regular, estacional ou multiplicativo).

Queda daquela explicada a forma de proceder ante series non estacionarias debido á tendencia. Agora ben, recordando a Definición 2.5 de serie estacionaria, outra posible causa da falta de estacionariedade sería a heterocedasticidade, isto é, varianza non constante. Este fenómeno é habitual en series temporais de demanda loxística como as de Trucksters. Para solucionar a falta de homocedasticidade, o máis habitual na práctica é aplicar unha transformación logarítmica aos datos $x = \log(y)$, aínda que existen outro tipo de transformacións, coñecidas como transformacións Box-Cox (véxase Shumway e Stoffer, 2017):

$$x_t = \begin{cases} \frac{y_t^\lambda - 1}{\lambda}, & \text{para } \lambda \neq 0, \\ \log(y_t), & \text{para } \lambda = 0 \end{cases} \tag{2.10}$$

Observación 2.6 *Nótese que as transformacións anteriores aplícanse unicamente a series $\{y_t\}_{t\in\mathbb{Z}}$ con todos os seus valores positivos. En caso de que a serie conteña algún valor non positivo, é preciso aplicar a transformación á serie $\{x_t\}_t = \{y_t\}_t + c$, $\forall t \in \mathbb{Z}$ sendo c unha constante que garanta que $x_t > 0$ $\forall t \in \mathbb{Z}$.*

Para a selección do λ óptimo existen varios métodos como o de máxima verosimilitude, sobre os que non se pretende afondar neste traballo.

Metodoloxía de construción do modelo

Recordemos que, tal e como se determinou na Sección 1.3, o obxectivo fundamental do traballo reside en facer predicións sobre o comportamento futuro das series temporais de demanda. Así, unha vez establecido o marco teórico dos procesos estocásticos e tras introducir os modelos clásicos de series temporais, o seguinte paso sería decidir cal dos modelos xa presentados se axusta mellor ás series temporais que se estudarán práctica.

Por conseguinte, o presente apartado detallará as distintas etapas da metodoloxía de Box-Jenkins, unha das metodoloxías que se aplicarán no Capítulo 4 para estudar as series de demanda.

Identificación e selección dun modelo xerador da serie

Na práctica, o primeiro paso consiste en seleccionar un modelo que poida ser xerador da serie de tempo. Para iso, é recomendable comezar levando a cabo unha análise exploratoria sobre a serie temporal. Para iso, cómpre apoiarse no gráfico secuencial da serie, así como nos gráficos de autocorrelacións simples (*fas*) e parciais (*fap*) (véxase Figura 2.3). Esta análise exploratoria permite extraer unha primeira idea de como é o comportamento da serie de demanda.

Inicialmente, abordarase o estudo da variabilidade da serie. Así, se no gráfico secuencial se observan problemas de heterocedasticidade, é necesario facer unha transformación de Box-Cox (ec. 2.10) sobre a serie orixinal. Na práctica, aplicarase fundamentalmente a transformación logarítmica (Capítulo 4).

Posteriormente, o seguinte aspecto a tratar será a presenza de tendencia. En particular, os problemas de tendencia poden detectarse ou ben no gráfico secuencial ou ben na gráfica da *fas* mostral, que tomaría valores positivos que decaen lentamente a cero a medida que aumenta o retardo. Ante esta situación, aplicaranse d diferenzas regulares, como xa se explicou na sección previa.

Tras eliminar a tendencia, o seguinte paso residirá no estudo da presenza de compoñente estacional. Este fenómeno observarase unha vez máis no gráfico secuencial da serie (comportamento cíclico con

período s); así como na gráfica da *fas* mostral, que amosará unha forte correlación positiva no retardo estacional, converxendo lentamente a cero cando crece o retardo e, ademais, con periodicidade do mesmo período ca serie. Daquela, no caso de que a serie presente tendencia estacional, aplicaranse D diferencias estacionais de período s, como se detallou previamente.

Unha vez se teña a serie diferenciada, o seguinte paso consistirá en determinar as ordes dos polinomios autorregresivos e de medias móbiles: parte regular (p, q) e estacional (P, Q). Como primeira idea, pódese observar novamente o gráfico secuencial así como os gráficos das funcións de autocorrelacións simple e parcial mostrais (véxase Aneiros, 2024 e o Capítulo 6 de Box et al., 2015). Porén, esta metodoloxía por si soa carece de validez, pois tan só é un criterio gráfico. En particular, este enfoque é subxectivo e pode provocar a identificación de varios posibles procesos como xeradores da serie e, en calquera caso, os modelos identificados serán relativamente sinxelos.

Por conseguinte, a forma de proceder será suxerir un modelo mediante a interpretación gráfica e, a continuación, seleccionar as ordes mediante a minimización dalgún criterio de información. Ditos índices permiten seleccionar o modelo óptimo buscando o equilibrio entre a bondade de axuste (maximizar a verosimilitude) e a complexidade do modelo (penalizar o número de parámetros). Os criterios máis empregados na práctica son os seguintes:

- **Criterio de información de Akaike** (AIC):

$$AIC(M) = -2\log(L(\hat{\varphi}_{r+1})) + 2r.$$

- **Criterio de información de Akaike corrixido** (AICc):

$$AICc(M) = -2\log(L(\hat{\varphi}_{r+1})) + 2(rT + r + 2)/(T - r - 2).$$

- **Criterio de información Bayesiano** (BIC):

$$BIC(M) = -2\log(L(\hat{\varphi}_{r+1})) + r\log(T).$$

Sendo $L(\hat{\varphi}_{r+1})$ a función de verosimilitude do modelo avaliada nos estimadores de máxima verosimilitude; r a cantidade de parámetros (sen contar a varianza das innovacións) do modelo; T o tamaño mostral; e M o modelo correspondente con vector de parámetros φ_{r+1}. Na práctica, co fin de evitar o sobreaxuste, ante modelos con diferenzas no criterio de información inferiores a 2 unidades, seleccionarase sempre o modelo máis sinxelo (menor número de parámetros).

Estimación do modelo seleccionado

Unha vez seleccionado un modelo como posible xerador da serie, o seguinte paso será estimar os parámetros do devandito modelo.

Por simplicidade, explicarase o caso dun modelo ARMA[1] e presentaranse dous métodos: estimación por mínimos cadrados e por máxima verosimilitude.

Dado entón un modelo ARMA(p, q), os parámetros que se deben estimar son os seguintes:

$$c, \phi_1, \dots, \phi_p, \theta_1, \dots, \theta_q \text{ e } \sigma_a^2,$$

onde σ_a^2 representa a varianza das innovacións. Considérese unha estimación de ditos parámetros, o vector das estimacións vaise denotar como:

$$\tilde{\beta} = (\tilde{c}, \tilde{\phi}_1, \dots, \tilde{\phi}_p, \tilde{\theta}_1, \dots, \tilde{\theta}_q, \tilde{\sigma}_a^2).$$

[1] O caso dos modelos ARMA estacionais será análogo incluíndo máis parámetros.

Ademais, os residuos asociados a dita estimación defínense como

$$\hat{a}_t = y_t - (\tilde{c} + \tilde{\phi}_1 y_{t-1} + ... + \tilde{\phi}_p y_{t-p} + \tilde{\theta}_1 \hat{a}_{t-1} + ... + \tilde{\theta}_q \hat{a}_{t-q}),$$

e a súa suma de cadrados

$$S(\tilde{\beta}) = \sum_{t=1}^{T} \hat{a}_t^2.$$

Estimación por mínimos cadrados e mínimos cadrados condicionados. A estimación dos parámetros polo método dos mínimos cadrados obtense mediante os valores de $\hat{\beta}$ que minimicen a función S,

$$\hat{\beta} = \arg\min_{\tilde{\beta}} S(\tilde{\beta}).$$

Ao buscar este mínimo poden aparecer dúas dificultades:

- Se $p > 0$, os valores dos residuos $\hat{a}_1, ..., \hat{a}_p$ dependen dos valores non observados de $Y_0, ..., Y_{1-p}$.
- Se ademais $q > 0$, $\hat{a}_{p+1}$ depende dos valores $\hat{a}_p, \hat{a}_{p-1}, ..., \hat{a}_{p+1-q}$, que á súa vez dependen de valores non observados da serie.

Co fin de solventar ambos problemas, xorde o método dos mínimos cadrados condicionados:

$$\min_{\tilde{\beta}} \sum_{t=p+1}^{T} \hat{a}_t^2,$$
$$\text{s. a. } \hat{a}_p = \hat{a}_{p-1} = ... = \hat{a}_{p+1-q} = 0.$$

Estimación por máxima verosimilitude. Para estimar por máxima verosimilitude os parámetros buscarase neste caso o vector $\hat{\beta}$ que maximiza a función de verosimilitude:

$$\hat{\beta} = \arg\max_{\tilde{\beta}} f_{\tilde{\beta}}(y_1, ..., y_T),$$

sendo $f_{\tilde{\beta}}$ a función de densidade conxunta asociada a un vector aleatorio $(\tilde{Y}_1, ..., \tilde{Y}_T)'$ procedente dun proceso ARMA(p, q) con coeficientes $\tilde{\beta}$.

Observación 2.7 *É preciso sinalar que, baixo certas condicións de regularidade, cúmprese que:*

- *Os estimadores de máxima verosimilitude dos parámetros do modelo excepto o da varianza das innovacións ($\hat{\sigma}_a^2$) son asintoticamente inesgados, eficientes e con distribución normal.*
- *O estimador de máxima verosimilitude da varianza das innovacións é consistente.*

Partindo da normalidade asintótica, validarase o modelo contrastando a significación individual de cada parámetro mediante o test da t de Student. Na práctica, seguirase un procedemento iterativo: se algún parámetro non resulta estatisticamente significativo, eliminarase do modelo (fixándoo a 0) e procederase a reaxustar o modelo reducido.

Diagnose do modelo

Unha vez estimados os parámetros, procederase á fase de diagnose, cuxo obxectivo é validar que os residuos do modelo axustado satisfán as hipóteses fundamentais de ruído branco (independencia, homocedasticidade e normalidade). Este proceso concíbese como un ciclo iterativo: no caso de que algunha destas hipóteses se incumpra, o modelo considerarase inadecuado, o que obrigará a retornar á etapa de identificación para propoñer unha nova estrutura e repetir o procedemento ata obter un modelo estatisticamente válido.

A suposición fundamental é que as innovacións $\{a_t\}_{t\in\mathbb{Z}}$ sexan ruído branco, é dicir, deben ser incorreladas, con media cero e varianza constante.

Como hipótese adicional tamén se comproba que as innovacións sexan gaussianas. Aínda que esta condición non é estritamente limitante para a validez do modelo, resulta adecuada. En caso de non cumprirse esta última hipótese, en caso de calcular intervalos de predición non se "garantirá" o nivel de confianza dos intervalos construídos en base á distribución do erro de predición.

Nótese que, como as innovacións $(a_1, \dots, a_T)$ non son observables, os contrastes realizaranse sobre os residuos do modelo $(\hat{a}_1, \dots, \hat{a}_T)$.

Antes de facer os distintos contrastes, pode ser de axuda representar o gráfico secuencial dos residuos. Isto permite notar problemas de tendencia ou variabilidade non constante principalmente, invalidando a hipótese de ruído branco.

Contraste de incorrelación. En primeiro lugar, cómpre verificar se os residuos están incorrelados. Para iso, adoptarase un enfoque mixto: gráfico e analítico.

Por unha parte, analizarase o correlograma dos residuos, que consiste na representación gráfica das funcións de autocorrelación (*fas* e *fap*) definidas anteriormente. Baixo a hipótese de ruído branco e fixado un nivel de significación do 5 %, espérase que aproximadamente o 95 % dos coeficientes de autocorrelación mostral se atopen dentro das bandas de confianza de Anderson, definidas por $\pm 1.96/\sqrt{T}$. Visualmente, a existencia de barras que sobresaen de xeito significativo destes límites indicaría a presenza de información sistemática non recollida polo modelo, o que obrigaría a revisar a súa estrutura.

Por outra banda, para unha validación formal, empregarase o contraste de Ljung-Box (Capítulo 8 de Box et al., 2015), que contrasta a hipótese nula de incorrelación conxunta ata un retardo m, isto é:

$$\begin{aligned} H_0 &: \rho_1 = \rho_2 = \cdots = \rho_m = 0 \\ H_1 &: \exists j \in \{1, \dots, m\} \text{ tal que } \rho_j \neq 0 \end{aligned}$$

O estatístico de proba defínese como:

$$Q_m = T(T+2)\sum_{k=1}^{m} \frac{\hat{\rho}_k^2}{T-k} \tag{2.11}$$

onde T é o tamaño da mostra e $\hat{\rho}_k$ a autocorrelación mostral de orde k dos residuos.

Baixo a hipótese nula, o estatístico Q_m (ec. 2.11) segue asintoticamente unha distribución χ^2 con $m-(p+q)$ graos de liberdade, onde p e q son as ordes da parte regular e estacional do modelo. Para o nivel de significación fixado (5 %), rexeitarase a hipótese de incorrelación se o valor de Q_m supera o cuantil 0.95 da distribución $\chi^2_{m-(p+q)}$[2].

Contraste de media cero. Unha vez comprobada a incorrelación dos residuos, contrastarase se a media dos mesmos é nula. Para levar a cabo isto, empregarase o test da t de Student para a media dunha mostra. Neste caso o estatístico vén dado por:

$$t = \frac{\bar{\hat{a}}}{\hat{s}_{\hat{a}}/\sqrt{T}}, \tag{2.12}$$

onde $\bar{\hat{a}}$ e $\hat{s}_{\hat{a}}$ denotan a media e a cuasivarianza mostral dos residuos e T o tamaño mostral.

[2]No caso dos ARMA estacionais multiplicativos a única diferenza reside nos graos de liberdade da distribución χ^2 sendo nese caso $m-(p+q+P+Q)$

Baixo a hipótese nula ($H_0 : \mu_a = 0$) dito estatístico segue unha distribución t de Student que se aproxima asintoticamente por unha $N(0,1)$. Daquela, fixado un nivel de significación do 5 %, rexeitarase que os residuos teñen media nula se

$$|\bar{\hat{a}}| \geq 1.96 \frac{\hat{s}_{\hat{a}}}{\sqrt{T}}.$$

Contraste de homocedasticidade. Para verificar a estabilidade da varianza nos residuos, na práctica (Capítulo 4) empregarase o estatístico de contraste H_k. Este test avalía se a varianza dos residuos é constante ao longo da serie comparando a variabilidade no tramo final da mostra coa do tramo inicial.

Formalmente, a hipótese nula (H_0) asume que a varianza é constante (homocedasticidade), mentres que a hipótese alternativa (H_1) suxire que a varianza aumenta ou diminúe co tempo (heterocedasticidade). O estatístico calcúlase como a razón entre a suma de cadrados dos residuos no último terzo da serie e a do primeiro terzo:

$$H_k = \frac{\sum_{t=T-k+1}^{T} \hat{a}_t^2}{\sum_{t=d+1}^{d+k} \hat{a}_t^2}, \tag{2.13}$$

sendo k o número de observacións considerado en cada tramo (na práctica tomarase un terzo da mostra).

Baixo a hipótese nula, o estatístico H_k (ec. 2.13) segue unha distribución $F_{k,k}$. Así, rexeitarase a homocedasticidade se o p-valor asociado ao estatístico é inferior ao nivel de significación considerado ($\alpha = 0.05$ neste traballo).

Contraste de normalidade. A última etapa da validación do modelo céntrase na distribución das innovacións. Como xa se sinalou previamente, a normalidade non é indispensable para a validez do modelo, mais si é adecuada.

Para avaliar esta hipótese, empregaranse criterios gráficos así como contrastes formais. Sobre a análise gráfica, pode resultar de utilidade un gráfico **Q-Q plot** dos residuos. Dito gráfico representa os cuantís mostrais fronte aos dunha distribución normal estándar. No caso en que se cumpra que as innovacións son i.i.d. con distribución normal, a disposición da mostra debería ser aproximadamente lineal (en particular aliñánse sobre a diagonal principal). Pola contra, desviacións nos extremos representarían a presenza de colas pesadas ou asimetría, o que indicaría a non normalidade das innovacións.

Para unha validación analítica rigorosa existen varias opcións. Unha delas é o test de Jarque-Bera que se basea en verificar se os coeficientes de asimetría e curtose dos residuos coinciden cos dunha distribución normal.[3]

Así, o estatístico da proba sería:

$$JB = T\left(\frac{S^2}{6} + \frac{(K-3)^2}{24}\right), \tag{2.14}$$

onde T é o tamaño da mostra, S o coeficiente de asimetría mostral e K o coeficiente de curtose mostral.

Baixo a hipótese nula (H_0 : a distribución das innovacións é gaussiana) o estatístico JB (ec. 2.14) distribúese de acordo a unha χ^2 con dous graos de liberdade. Daquela, se JB supera o cuantil de orde 0.95 dunha χ^2_2 rexeitarase a normalidade dos residuos.

Cómpre sinalar que o test de Jarque-Bera non é a única opción para contrastar a normalidade das innovacións. Outra opción sería o contraste de Shapiro-Wilk (véxase Aneiros, 2024).

[3] Recordemos que a distribución $N(0,1)$ ten simetría 0 e coeficiente de curtose 3.

Predición

Unha vez validado o modelo e tras comprobar que os residuos satisfán as hipóteses de ruído branco, procédese á etapa final e obxectivo principal deste estudo: a predición de valores futuros. En particular, na práctica buscarase predicir os valores da serie para un horizonte h baseándose na información dispoñible ata o intre T.

Considérese o preditor que minimice o erro cadrático medio de predición (MSE) que será:

$$\hat{Y}_T(h) = \mathbb{E}[Y_{T+h}|Y_1, \dots, Y_T].$$

Para obter estas predicións na práctica, seguirase un procedemento recursivo que se fundamenta na expresión do modelo axustado (ec. 2.9):

1. **Expansión do modelo:** reescríbese o modelo ARIMA de xeito que no lado esquerdo queden os termos Y_t e no lado dereito quedarán o resto de termos.
2. **Desprazamento temporal:** substitúese o índice temporal t por $T + h$.
3. **Substitución de valores:** reemplázanse os termos teóricos polas súas respectivas estimacións ou valores coñecidos de acordo ás seguintes indicacións: os valores observados (Y_t, $t \leq T$) polos seus valores reais; as futuras observacións (Y_t, $t > T$) polas correspondentes predicións (calculadas recursivamente); as innovacións pasadas (a_t, $t \leq T$) polos residuos estimados; e as innovacións futuras pola súa esperanza, que é cero.

A predición puntual obtida, $\hat{Y}_T(h)$, denomínase predición con orixe en T e horizonte h.

Observación 2.8 *Cómpre ter en conta as seguintes consideracións:*

- *Se $d = 0$, as predicións a longo prazo converxerán á media do proceso.*
- *Se $d = 1$, as predicións a longo prazo converxerán, ou ben a unha constante se $c = 0$, ou ben a unha recta se $c \neq 0$.*
- *Se $d > 1$, as predicións a longo prazo converxerán, ou ben a unha constante se $c = 0$, ou ben presentarán tendencia cadrática se $c \neq 0$. Nos futuros axustes deste proxecto (Capítulo 4) , decidiuse non incluír a constante neste caso, posto que unha tendencia cuadrática é perigosa á hora de predicir.*

No caso de modelos estacionais (ARIMA estacionais multiplicativos), este comportamento da tendencia sucede de forma análoga, pero co patrón estacional superposto á tendencia base definida por d e c.

Ademais de obter as predicións puntuais, tamén se poderían construír intervalos de predición. Para iso existen dous posibles escenarios: se as innovacións son gaussianas, os intervalos constrúense a partires desta propiedade; se non se cumpre a normalidade, a alternativa reside en aplicar técnicas de remostreo (Bootstrap sobre os residuos). Para profundizar sobre a construción dos mesmos, poden consultarse as referencias Cao e Fernández (2022) e o Capítulo 5 de Box et al. (2015).

Extensións dos modelos Box-Jenkins: Análise de intervencións e atípicos

Na modelización de series temporais económicas e loxísticas, é frecuente a aparición de observacións que diverxen significativamente do comportamento estocástico esperado do proceso base. Estas anomalías, se non son tratadas adecuadamente, poden distorsionar a identificación do modelo ARIMA adecuado, comprometendo a precisión das predicións.

Na presente sección, abordarase a caracterización teórica destes fenómenos, distinguindo dous tipos de factores segundo a dispoñibilidade a priori sobre o momento de ocurrencia.

Por unha banda, unha serie de tempo pode estar afectada por factores exóxenos de cronoloxía coñecida, tales como eventos especiais (Black Friday) ou cambios na política de mercado da empresa. A análise de **intervención** permite incorporar estes efectos deterministas ao modelo.

Por outra banda, estudarase a detección de valores atípicos que suceden en intres descoñecidos. Dende unha perspectiva teórica, estes impactos clasifícanse segundo a súa natureza en **atípicos aditivos (AO)**, que afectan só á observación puntual, e **atípicos innovativos (IO)**, que afectan á innovación e o seu efecto propágase no tempo segundo a estrutura do modelo. Cómpre sinalar que, no axuste dos modelos do Capítulo 4, os atípicos modelaranse mediante variables *dummy*.

Durante este apartado seguirase fundamentalmente a referencia Box et al. (2015).

Análise de intervención

A análise de intervención estuda modelos de series temporais nos que se inclúen variables ficticias para representar sucesos exóxenos que producen efectos deterministas. As intervencións clasifícanse en dous tipos: os **efectos permanentes**, que provocan un cambio no nivel da serie a partires dun intre coñecido; e os **efectos transitorios**, que provocan cambios en algúns valores da serie, mais a longo prazo o nivel da mesma non se ve afectado.

Durante a presente análise, suporase que a intervención tivo lugar nun momento h coñecido. Ademais, denotarase por $\{Y_t\}_{t\in\mathbb{Z}}$ o proceso que xera a serie intervida, e $\{X_t\}_{t\in\mathbb{Z}}$ o proceso xerador da serie se non sufrise intervención. En particular, dito proceso será un ARIMA dos explicados na Sección 2.3.2.

Antes de explicar a modelización dos efectos permanentes e transitorios, cómpre definir dúas funcións que serán de gran utilidade.

Definición 2.22

- *Defínese a* ***función chanzo*** $\left(S_t^{(h)}\right)$ *como*

$$S_t^{(h)} = \begin{cases} 0, & si\ t < h, \\ 1, & si\ t \geq h. \end{cases} \tag{2.15}$$

- *Defínese a* ***función impulso*** $\left(P_t^{(h)}\right)$ *como*

$$P_t^{(h)} = \begin{cases} 0, & si\ t \neq h, \\ 1, & si\ t = h. \end{cases} \tag{2.16}$$

Observación 2.9 *As funcións definidas anteriormente cumpren que*

- $S_t^{(h)} - S_{t-1}^{(h)} = P_t^{(h)}$.
- $S_t^{(h+j)} = S_{t-j}^{(h)}$ *se* $j \geq 0$.

Así, o modelo para unha serie con intervención será da forma:

$$Y_t = \omega(B) f_t^{(h)} + X_t, \tag{2.17}$$

onde $\omega(B)$ denomínase función de transferencia e $f_t^{(h)}$ será a función chanzo (ec. 2.15), se o efecto é permanente, ou a función impulso (ec. 2.16), se o efecto é transitorio.

A clave da análise de intervención reside en como se modela a función de transferencia, $\omega(B)$, que describe o impacto da intervención.

Modelización dos efectos permanentes. Os efectos permanentes caracterízanse por alterar o nivel medio ou tendencia da serie a partires do intre da intervención $t = h$. A continuación preséntanse tres tipos de estruturas básicas para modelar estes efectos, baseadas na función chanzo $S_t^{(h)}$.

En primeiro lugar, considérese o caso no cal a intervención provoca un desprazamento inmediato e constante no nivel da serie. O efecto consiste nun salto de ω_0 unidades a partires do intre $t = h$. Así, o modelo sería:

$$Y_t = \omega_0 S_t^{(h)} + X_t.$$

Outro caso dase cando o salto non é brusco, senón que o nivel da serie cambia gradualmente dende o momento da intervención ($t = h$) ata outro intre $t = h + l$. Daquela, o modelo (ec. 2.17) escribirase como:

$$Y_t = (\omega_0 + \omega_1 B + ... + \omega_l B^l) S_t^{(h)} + X_t.$$

Por último, modélase o caso onde o efecto da intervención consiste en variar gradualmente o nivel da serie dende o intre $t = h$. Este comportamento dinámico captúrase mediante un parámetro δ ($0 < \delta < 1$). O correspondente modelo será:

$$Y_t = \frac{\omega_0}{1 - \delta B} S_t^{(h)} + X_t.$$

En particular, a función de transferencia pódese reescribir como $\omega(B) = \omega_0(1 + \delta B + ... + \delta^j B^j + ...)$.

Modelización dos efectos transitorios. A diferenza dos anteriores, os efectos transitorios tan só cambian algúns valores da serie, sen presenciarse efecto no nivel asintoticamente. De forma análoga aos efectos permanentes, preséntanse tres estruturas para modelizar os efectos, baseadas neste caso na función impulso $P_t^{(h)}$.

Un exemplo paradigmático deste tipo de fenómenos no sector loxístico é o Black Friday. Este evento xera un incremento masivo na demanda de transporte durante unha semana específica, pero unha vez transcorrido o período de ofertas, o volume de pedidos tende a retornar aos seus niveis habituais. Dependendo de como se disipe este impacto (de forma instantánea ou gradual), empregarase unha das seguintes estruturas:

En primeiro lugar, considérase o caso no cal a intervención provoca un cambio de ω_0 unidades unicamente no intre no cal sucede ($t = h$). Entón, o modelo resulta:

$$Y_t = \omega_0 P_t^{(h)} + X_t.$$

Outro posible efecto transitorio dunha intervención consiste en variar tódolos valores da serie dende o intre $t = h$ ata o momento $t = h + l$. Neste caso, o modelo será:

$$Y_t = (\omega_0 + \omega_1 B + ... + \omega_l B^l) P_t^{(h)} + X_t.$$

Por último, establécense os modelos no caso de que o efecto transitorio provoque variacións na serie dende o intre $t = h$ ata o final. Neste caso as variacións estarán xeradas a partires dos parámetros ω_0 e δ ($0 < \delta < 1$). Así, o modelo (ec. 2.17) reescríbese como segue:

$$Y_t = \frac{\omega_0}{1 - \delta B} P_t^{(h)} + X_t.$$

Partindo dos modelos de intervención presentados ata o momento poderían construírse outros máis sofisticados (véxase Aneiros, 2024).

Neste momento nace unha pregunta moi natural: *como se constrúe un modelo de intervención na práctica?*

O primeiro a decidir será o tipo de efecto e a función de transferencia seleccionada. Isto identificarase facendo a exploración da serie completa $(y_1, \dots, y_T)$. Tras isto, deberase seleccionar un modelo ARIMA como posible xerador do proceso sen intervención $\{X_t\}_{t\in\mathbb{Z}}$. Para tal fin existen dúas opcións:

- Propoñer un modelo en base ao estudo das observacións previas á intervención.
- Suxerir como modelo o ARIMA que minimice algún dos criterios de información xa explicados (AIC, BIC ou AICc) na estimación do modelo para o proceso completo $\{Y_t\}_{t\in\mathbb{Z}}$.

O seguinte paso sería estimar os parámetros do modelo ARIMA seleccionado e da función de transferencia $\omega(B)$. Isto realizarase mediante máxima verosimilitude, mais non se entrará en detalle neste procedemento.

Análise de atípicos

En ocasións, as observacións dunha serie de tempo poden estar afectadas por erros ou perturbacións que crean efectos espurios nas series e resultan en comportamentos "inusuais". Estas observacións que se saen do comportamento habitual da serie denomínanse *outliers* ou valores atípicos. A presenza deste tipo de observacións nunha serie de tempo compromete a selección e estimación dun modelo ARIMA, podendo provocar nesgos nas estimacións dos parámetros ou distorsionar a identificación da estrutura de correlación (*fas/fap*). A casuística de que o momento h no cal sucede un evento que dá lugar a un atípico fose coñecido, estaría cuberto pola análise de intervención descrita previamente. Porén, na práctica, a presenza de *outliers* é descoñecida ao comezo da análise da serie, co cal será preciso buscar algunha alternativa para tratar estes casos. Neste apartado explicaranse outras metodoloxías para modelar o efecto destes atípicos na serie de tempo.

É preciso facer varias consideracións previas ao desenvolvemento do presente apartado:

- Suporase que só hai un atípico no intre $t = h$.
- A serie con efecto do atípico provén do proceso $\{Y_t\}_{t\in\mathbb{Z}}$.
- O proceso xerador da serie sen efecto do atípico denotarase por $\{X_t\}_{t\in\mathbb{Z}}$ e provén dun proceso ARIMA dos definidos na Sección 2.3.2.

Agora ben, como xa se indicou previamente, na teoría distinguiranse dous tipos de valores atípicos: os atípicos aditivos (AO) e os atípicos innovativos (IO).

Por unha parte, os atípicos aditivos (AO) serán aqueles nos que o valor da serie no intre $t = h$ foi xerado de forma distinta ao resto. En particular, terase que

$$Y_t = \omega_A P_t^{(h)} + X_t,$$

sendo $P_t^{(h)}$ a función impulso definida por (ec. 2.16).

Por outra banda, nos atípicos innovativos (IO) a innovación no intre $t = h$ foi xerada de forma distinta ao resto. Neste caso, as innovacións do proceso $\{Y_t\}_{t\in\mathbb{Z}}$ serán:

$$e_t = \omega_I P_t^{(h)} + a_t,$$

sendo a_t as innovacións do proceso $\{X_t\}_{t\in\mathbb{Z}}$ e $P_t^{(h)}$ de novo a función impulso.

Dende un punto de vista práctico, a detección de atípicos realízase a través da análise dos residuos do modelo estimado. Por conseguinte, resulta fundamental caracterizar o efecto que exercen os atípicos aditivos e innovativos sobre a estrutura dos residuos.

Asumindo que o proceso ARIMA é invertible, póndense expresar as innovacións do proceso sen "contaminar" en función da serie mediante a relación $a_t = \pi(B)X_t$, onde $\pi(B) = 1 - \sum_{j=1}^{\infty} \pi_j B^j$. De forma análoga, é posible facer o propio sobre as innovacións do proceso co efecto do atípico: $e_t = \pi(B)Y_t$. Baixo estas condicións, estúdase o impacto dos atípicos sobre os residuos.

No caso dos atípicos aditivos, o efecto no residuo non se limita ao intre h, senón que se propaga aos momentos posteriores dependendo da estrutura autorregresiva do modelo. En particular:

$$e_t = \omega_A \pi(B) P_t^{(h)} + a_t.$$

Isto implica que o residuo no intre $t = h$ sofre un incremento ω_A, mais nos momentos seguintes ($t > h$), o residuo vese afectado por $-\omega_A \pi_{t-h}$.

En canto ao efecto dos atípicos innovativos, xa se explicou que o efecto deste tipo de *outliers* afecta directamente ás innovacións. De feito tense que

$$e_t = \omega_I P_t^{(h)} + a_t.$$

Neste caso o único residuo que se ve afectado é o do momento do atípico ($t = h$).

Partindo destas relacións, o problema de detección de atípicos redúcese a estimar o parámetro ω nun modelo de regresión simple sobre os residuos $e_t = \omega \hat{z}_t + a_t$ e contrastar a súa significación (mediante o contraste da t de Student). A variable explicativa $\hat{z}_t$ dependerá do tipo de atípico (AO ou IO) e dos parámetros do modelo ARIMA estimado.

Na práctica, descoñécese a priori o número, localización (h) e natureza dos atípicos. Por ese mesmo motivo, será preciso implementar un procedemento de detección iterativo aplicado sobre a totalidade das observacións. O algoritmo que se empregará na práctica (Capítulo 4) baséase no método de Bonferroni e consta das seguintes etapas:

1. **Estimación inicial:** Axústase un modelo ARIMA á serie orixinal e obtéñense os residuos estimados.
2. **Cálculo de estatísticos:** Para cada intre t, calcúlanse os estatísticos t de Student baixo as hipóteses nulas de que dita observación é un atípico aditivo ou innovativo.
3. **Selección de candidatos:** Identifícanse como candidatos todos aqueles puntos cuxos p-valores asociados sexan inferiores a $0.05/T$ (na práctica fixarase un nivel de significación do 5 %).
4. **Identificación e estimación:** De entre tódolos candidatos seleccionarase aquel de menor p-valor. Incorpórase o efecto de dito atípico ao modelo e recalcúlanse os residuos.
5. **Converxencia.** O proceso repítese iterativamente sobre os novos residuos ata que non se detecte ningún atípico adicional significativo.

2.3.3. Modelos *Prophet*

Unha vez analizada a metodoloxía clásica de Box-Jenkins, nesta sección preséntase unha alternativa moderna e flexible, especialmente deseñada para modelar series con fortes compoñentes estacionais e influencias de eventos de calendario, características típicas das series de demanda: o modelo *Prophet*.

Este algoritmo, desenvolto por Taylor e Letham (2018) no equipo de *Core Data Science* de **Facebook**, afástase do corrente dos modelos ARIMA. Mentres que os métodos clásicos proxectan o erro a través dunha estrutura de dependencia temporal, *Prophet* baséase no enfoque dos **Modelos Aditivos Xeneralizados (GAM)**. Conceptualmente, aborda o problema de predición como un exercicio de axuste de curvas (*curve fitting*).

A formulación matemática do *Prophet* fundaméntase na descomposición da serie temporal $y(t)$ en tres compoñentes principais: tendencia, estacionalidade e festivos, aos que se lles suma un termo de erro.

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t. \tag{2.18}$$

No modelo (2.18) tense que:

- $g(t)$ representa a tendencia ou crecemento non periódico (cambios no nivel da serie).
- A compoñente $s(t)$ modela a estacionalidade periódica (patróns semanais, anuais,...).
- $h(t)$ recolle o efecto de festivos e eventos especiais (Black Friday).
- Por último, ε_t é o termo do erro, para o que se asume unha distribución normal.

Esta estrutura aditiva aporta unha serie de vantaxes: flexibilidade á hora de axustar estacionalidades máis complexas; robustez fronte a datos faltantes e non precisa que as observacións sexan equiespacidadas (a diferenza dos modelos ARIMA); o axuste é computacionalmente rápido; e, ademais, os parámetros do modelo son facilmente interpretables.

Nas seguintes subseccións detallarase a especificación das distintas compoñentes e como esta estrutura permite incorporar coñecemento do sector (festivos, Black Friday,...) de forma máis intuitiva ca nos modelos clásicos.

Modelo da tendencia

A compoñente da tendencia, $g(t)$, captura os cambios non periódicos no nivel da serie temporal. O modelo *Prophet* permite implementar dous tipos de crecemento para esta compoñente: un **modelo de crecemento loxístico saturado** e un **modelo de crecemento lineal segmentado**.

O crecemento loxístico é axeitado para problemas onde se coñece un límite físico de capacidade. Porén, neste proxecto non se empregará esta opción, pois non se imporá un límite superior á capacidade de crecemento da empresa. Se o lector quere afondar neste tipo de modelos pode consultar a referencia básica desta sección Taylor e Letham (2018). Por conseguinte, considerarase un modelo de crecemento lineal segmentado.

Neste tipo de modelos, os cambios na tendencia incorpóranse definindo unha serie de puntos de cambio. Ditos puntos denotaranse por t_j con $j = 1, \dots, S$. Defínese ademais para cada punto de cambio (t_j):

$$a_j(t) = \begin{cases} 1, & \text{si } t \geq t_j, \\ 0, & \text{si } t < t_j. \end{cases}$$

Así, a tendencia formúlase como:

$$g(t) = \left(k + \sum_{j=1}^{S} a_j(t)\delta_j \right) t + \left(m + \sum_{j=1}^{S} a_j(t)\gamma_j \right), \tag{2.19}$$

onde:

- k é a taxa de crecemento base.
- δ_j representa o cambio na taxa de crecemento que se produce no intre t_j.
- m é o parámetro de desprazamento.
- $\gamma_j = -t_j\delta_j$ para asegurar a continuidade de $g(t)$.

Na definición da función de tendencia $g(t)$ (ec. 2.19) obsérvase un problema: *como se seleccionan os puntos de cambio t_j?* Os puntos de cambio poden ser especificados manualmente se as datas de cambios de política na empresa ou novos contratos son coñecidos de antemán. Porén, o máis habitual na práctica é seleccionalos automaticamente a partires dun conxunto de candidatos.

O modelo especifica unha gran cantidade de puntos de cambio, empregando a distribución *a priori* $\delta_j \sim \text{Laplace}(0, \tau)$[4]. O parámetro τ permite controlar a flexibilidade do modelo á hora de detectar cambios na tendencia. Cómpre sinalar que, as consideracións sobre δ_j non teñen impacto sobre a taxa de crecemento base k, co cal se τ é próximo a 0 o axuste redúcese a un modelo lineal estándar (non segmentado).

Agora ben, á hora de facer as predicións, a tendencia terá unha taxa constante. Mostrouse que o modelo para a tendencia asume que existen S puntos de cambio dunha mostra de T puntos, producíndose en cada un deles un cambio da forma $\delta_j \sim \text{Laplace}\ (0, \tau)$. Así, simularanse cambios de taxa futuros que imitan os axustados reemplazando τ por unha varianza estimada a partir dos datos. En particular, podería empregarse **inferencia Bayesiana** partindo da distribución *a priori* de τ, ou ben estimar mediante **máxima verosimilitude** o parámetro de escala $\lambda = \frac{1}{S}\sum_{j=1}^{S} |\delta_j|$.

Os futuros puntos de cambios mostréanse aleatoriamente de forma que a frecuencia coincida coa do histórico, é dicir, $\forall j > T$:

$$\begin{cases} \delta_j = 0 \text{ con probabilidade } \frac{T-S}{T}, \\ \delta_j \sim \ \text{Laplace}(0, \lambda) \text{ con probabilidade } \frac{S}{T}. \end{cases}$$

En consecuencia, a incerteza na predición da tendencia mídese asumindo que o futuro vai presentar a mesma frecuencia e magnitude media de cambios no nivel da serie ca o histórico.

Ademais, unha vez estimado o parámetro λ, pódese usar este modelo para simular tendencias futuras e contruír intervalos de predición. Non obstante, a hipótese de que a tendencia ten un comportamento similar á observada é moi forte, co cal non se espera que os intervalos así construídos teñan cobertura exacta. Porén, poden resultar de utilidade á hora de notar un problema de sobreaxuste. En particular, se τ é grande o modelo presenta unha maior flexibilidade á hora de axustarse á serie histórica e o erro de entrenamento diminuirá. A consecuencia disto será unha maior amplitude dos intervalos de predición.

Estacionalidade

A segunda compoñente estrutural do modelo é a estacionalidade, $s(t)$, que captura as variacións periódicas da serie temporal (patróns semanais, mensuais ou anuais). A diferenza dos modelos ARIMA que axustan a estacionalidade mediante diferenciación ou retardos estacionais ríxidos, *Prophet* modela estes efectos como unha función continua baseada en series de Fourier (véxase López Pouso, 2019).

Esta aproximación permite modelar múltiples períodos estacionais simultaneamente e manexar periodicidades non enteiras (por exemplo, un ano como 365.25 días). Formalmente, sexa P o período estacional, a compoñente $s(t)$ pode aproximarse pola seguinte suma de senos e cosenos:

$$s(t) = \sum_{n=1}^{N} \left(a_n \cos\left(\frac{2\pi n t}{P}\right) + b_n \operatorname{sen}\left(\frac{2\pi n t}{P}\right)\right). \tag{2.20}$$

Para poder axustar a estacionalidade será necesario estimar $2N$ coeficientes: a_n e b_n, $n = 1, \ldots, N$ que seguen unha distribución *a priori* $N(0, \sigma^2)$.

O parámetro N xoga un papel fundamental no axuste. Un valor baixo de dito parámetro xera unha estacionalidade suave, mentres que un N elevado captura estacionalidades con cambios bruscos e alta

[4] A distribución de Laplace resulta da diferenza de dúas variables aleatorias exponenciais i.i.d.

frecuencia, podendo caer no sobreaxuste. *Prophet* emprega os valores $N = 10$ para a estacionalidade anual e $N = 3$ para a estacionalidade semanal. Na práctica, a selección de dito parámetro levarase a cabo mediante unha busca en malla (*Grid Search*) minimizando algún criterio do erro (Sección 4.3).

Matricialmente, esta compoñente intégrase no modelo xeral xerando una matriz de regresores $X(t)$ formada polos termos sinusoidais de xeito que $s(t) = X(t)\beta$ sendo β o vector dos $2N$ coeficientes $(a_1, b_1, \dots, a_N, b_N)$.

Festivos e eventos

Os festivos ou campañas especiais teñen un impacto significativo sobre as series temporais no ámbito económico e loxístico. A miúdo os efectos destas datas non seguen patróns estacionais, co cal non son capturados pola compoñente $s(t)$ definida previamente.

Prophet é capaz de modelar tales efectos mediante a terceira compoñente do modelo (ec. 2.18): $h(t)$. En particular, dita compoñente permite incorporar ao modelo unha listaxe de datas asumindo que os efectos son independentes entre si.

É posible incluír unha fiestra de efecto para cada un dos festivos, o que permite reflectir o impacto do evento tanto nos días previos coma nas datas posteriores á mesma. Matematicamente, para cada data especial i, sexa D_i o conxunto dos días previos e posteriores ao evento. Engádese unha función indicadora para representar se a observación t sucede dentro da fiestra do evento i, e a cada data especial asígnase un parámetro κ_i que captura o efecto do festivo no modelo. Daquela xérase unha matriz de regresores

$$Z(t) = [1(t \in D_1), \dots, 1(t \in D_L)],$$

onde L é o número total de eventos considerados. Así, a compoñente do modelo $h(t)$ escribirase como segue

$$h(t) = Z(t)\kappa, \tag{2.21}$$

sendo κ o vector dos parámetros de impacto (κ_i) para cada festivo i. Para evitar o sobreaxuste, impónse sobre o vector de parámetros κ unha distribución *a priori* $N(0, v^2)$, onde o parámetro v permite regular a flexibilidade.

2.4. Medidas de adecuación das predicións

A formalización matemática dos modelos de series temporais cobra sentido práctico ao ser contrastada coa súa eficacia preditiva. Co obxectivo de garantir unha toma de decisións óptima na planificación operativa de Trucksters, resulta imprescindible definir un conxunto de métricas de erro que faciliten a comparación entre as diferentes arquitecturas e aseguren a selección do modelo con maior exactitude na estimación da demanda futura.

En particular, consideraranse catro métricas distintas: **Root Mean Squared Error (RMSE)**, **Mean Absolute Error (MAE)**, **Mean Absolute Percentage Error (MAPE)** e **Mean Absolute Scaled Error (MASE)** que se definen a continuación.

No que segue, denotarase por $e_T(h) = y_{T+h} - \hat{y}_T(h)$ o erro de predición observado con orixe en T e horizonte h e consideraranse as predicións obtidas para horizontes $h = 1, \dots, H$.

Definición 2.23 *Defínese o* ***RMSE*** *como a seguinte métrica do erro:*

$$RMSE = \sqrt{\frac{1}{H}\sum_{h=1}^{H} e_T^2(h)}. \tag{2.22}$$

O RMSE é unha métrica éstandar que penaliza fortemente os grandes erros. O preditor que minimiza teoricamente esta métrica é a media. Porén, presenta unha limitación principal: depende das unidades da serie observada. Ademais, é unha métrica sensible ante a presenza de atípicos.

Definición 2.24 *Defínese o* ***MAE*** *como o promedio dos erros absolutos, é dicir:*

$$MAE = \frac{1}{H}\sum_{h=1}^{H}|e_T(h)|. \tag{2.23}$$

O MAE resulta unha métrica máis robusta ante *outliers* ca o RMSE e a súa interpretación é directa. Ademais, o preditor que a minimiza é a mediana. Non obstante, presenta o mesmo inconvinte ca métrica anterior, pois depende das unidades da serie.

Definición 2.25 *Defínese o* ***MAPE*** *como a seguinte medida porcentual:*

$$MAPE = \frac{1}{H}\sum_{h=1}^{H}|\rho_T(h)|, \tag{2.24}$$

sendo $\rho_T(h) = \frac{e_T(h)}{y_{T+h}} \cdot 100$.

O MAPE trátase dun erro porcentual, co cal a súa interpretación é moi sinxela. Ademais, soluciona o problema das métricas anteriores, pois non depende das unidades. Pola contra, non é unha medida útil cando a serie presenta algún valor próximo a 0, pois resultarían en erros desproporcionados.

Definición 2.26 *Defínese o* ***MASE*** *como:*

$$MASE = \frac{1}{H}\sum_{h=1}^{H}|p_T(h)|, \tag{2.25}$$

onde

$$p_T(h) = \frac{e_T(h)}{\frac{1}{T-1}\sum_{t=2}^{T}|y_t - y_{t-1}|}.$$

En particular, o MASE é un erro escalado que se constrúe a partires do MAE calculado sobre o preditor *naïve*[5] na mostra histórica.

Esta última métrica ten unha interpretación moi interesante, pois un valor inferior a 1 indicaría que o modelo axustado predí, en media, mellor que o preditor naïve a horizonte 1, é dicir, a predición é mellor ca simplemente considerar o valor observado anterior como predición. Por ser adimensional e estable en cero, considérase unha das métricas máis robustas para comparar series heteroxéneas.

Observación 2.10 *No caso de que a serie temporal teña compoñente estacional de período s, o preditor naïve consiste en predecir o valor nun intre t (y_t) como o respectivo valor observado no período anterior (y_{t-s}). Daquela, na fórmula do MASE (ec. 2.25) o termo do erro escalado reescribiríase como:*

$$p_T(h) = \frac{e_T(h)}{\frac{1}{T-s}\sum_{t=s+1}^{T}|y_t - y_{t-s}|}.$$

En resumo, a selección destas catro métricas configuran un marco de avaliación rigoroso. O uso complementario do RMSE e o MAE permite cuantificar o impacto operativo dos erros de magnitude, mentres que o MASE e o MAPE facilitan a comparativa do rendemento dos distintos modelos.

[5] O preditor naïve consiste en predecir o valor da serie nun intre t (y_t) mediante o valor observado no momento anterior $t-1$ (y_{t-1})

2.5. Ferramentas Computacionais

A implementación práctica de todas as técnicas descritas neste capítulo levouse a cabo empregando a linguaxe de programación Python (consúltese Python Software Foundation, 2026). A elección desta linguaxe fundamentouse na súa versatilidade, a potencia das súas bibliotecas especializadas en series temporais, ademais de ser o entorno de traballo empregado pola empresa Trucksters.

O desenvolvemento organizouse de forma modular (dividindo o traballo en diferentes *scripts*) para garantir a reprodutivilidade e a escalabilidade do código.

Entorno de desenvolvemento e xestión de datos

Utilizouse o entorno interactivo **Jupyter Notebook**, que permitiu combinar a execución de código coa documentación técnica e a visualización inmediata de resultados. Para a manipulación estruturada de datos e os cálculos numéricos básicos, as bibliotecas fundamentais foron `Pandas` e `NumPy`.

Bibliotecas específicas empregadas no traballo

Dependendo da técnica estatística aplicada, utilizáronse as seguintes librerías de referencia:

Análise Multivariante e *Outliers*. Para a análise de correspondencias utilizouse a librería `prince`, mentres que para o *clustering* xerárquico e a detección de anomalías mediante *Isolation Forest* empregáronse `SciPy` e `scikit-learn` respectivamente.

Modelización de Series Temporais. A metodoloxía de Box-Jenkins implementouse mediante `statsmodels` e a biblioteca de automatización `pmdarima`. Para a metodoloxía baseada en compoñentes, empregouse `Prophet`, desenvolta polo equipo *Core Data Science* de **Facebook**.

Representación gráfica. Os diferentes gráficos de resultados e diagnóstico xeráronse mediante `Matplotlib` e `Seaborn`.

Funcións propias. Un aspecto clave do deseño computacional deste traballo foi a creación dun módulo de funcións personalizadas denominado `Funcions.py` e que se pode consultar en Airas (2026). Este ficheiro actúa como o centro da análise de series temporais, recollendo os procedementos precisos na aplicación das distintas metodoloxías de series temporais:

- **Preprocesado:** Automatización da carga de datos e agregación semanal.
- **Diagnóstico:** Funcións para a xeración de correlogramas (*fas/fap*) e contrastes de validación.
- **Visualización de compoñentes:** Desenvolvemento de funcións gráficas específicas para descompoñer o impacto de festivos, estacionalidade e regresores extra nos modelos *Prophet*.

Esta estrutura modular permite que o código sexa facilmente adaptable a estudos futuros, sen necesidade de comezar o traballo dende cero, ademais de axudar a un mellor seguimento dos *scripts* empregados no presente proxecto.

Dispoñendo agora de todas estas ferramentas, poderase afrontar a parte práctica do traballo. O seguinte capítulo comezará coa presentación e análise exploratoria da base de datos de Trucksters. Posteriormente, no Capítulo 4 procederase á implementación computacional e ao axuste dos modelos propostos no presente capítulo. O proxecto rematará coa especificación das conclusións en base aos resultados acadados.

Parte II

Aplicación práctica

Capítulo 3

Análise exploratoria dos datos

Unha vez establecida a fundamentación metodolóxica no capítulo precedente, este terceiro bloque ten como obxectivo implementar ditas técnicas mediante unha Análise Exploratoria de Datos (EDA) exhaustiva. Esta fase resulta crucial para transformar os datos brutos en coñecemento estatístico, permitindo validar as hipóteses de partida e adecuar a estratexia de modelización.

Ao longo deste capítulo, procederase á caracterización das variables que integran a base de datos, poñendo especial atención na análise clúster (descrita na Sección 3.2) para identificar patróns de comportamento entre os diversos perfís de clientes. Así mesmo, examinaranse as relacións e dependencias entre as variables de interese, o que permitirá detectar posibles estruturas de correlación e valores atípicos. Todo este proceso servirá de ponte necesaria para garantir a robustez e interpretabilidade dos modelos de series temporais que se desenvolverán no Capítulo 4.

Co obxectivo de garantir a reprodutivilidade dos resultados, o código en Python empregado para este tratamento de datos e a xeración de visualizacións pode consultarse no seguinte repositorio de GitHub: Airas (2026).

3.1. Preprocesamento e preparación de datos

O obxectivo deste proxecto consiste en estudar distintos modelos para predicir o volume de envíos solicitados nunha data determinada por un cliente e, daquela, ver se é posible anticipar a chamada *peak season*. Para iso, traballouse cunha serie de variables obtidas da base de datos relacional PostgreSQL da empresa, aloxada en Google Cloud, mediante consultas SQL adaptadas á súa estrutura.

En particular, na base de datos recóllese información acerca de cada pedido que se leva a cabo. A continuación, inclúese un resumo das variables presentes na base de datos, aínda que para unha explicación máis detallada recoméndase consultar o Apéndice A.

- **Identificadores e datos básicos:** Inclúense variables coma o ID do pedido, que permite rastrexar cada envío de forma única; ou información sobre cliente, orixe e destino, datos clave para estudar patróns xeográficos e de demanda. A información sobre o cliente será fundamental para a agrupación da carteira de clientes mediante a análise clúster da Sección 3.2.
- **Variables de volume e distancia:** Disponse de variables indicadoras da eficiencia e aproveitamento da flota, como son os quilómetros con carga e os quilómetros percorridos en baleiro. Tamén se inclúe información sobre o tipo de remolque, que pode afectar á capacidade de transporte e á selección de rutas.

- **Variables económicas:** Información acerca dos prezos por cliente e provedor, variables que permiten avaliar marxes e custos asociados.
- **Variables temporais:** Conteñen información sobre a data do envío, axudando a analizar tendencias ao longo do tempo e identificar épocas de alta demanda.
- **Festivos e datas especiais:** Indicadores de se o pedido foi realizado durante un festivo ou algunha data especial coma o Black Friday. A información detallada sobre os festivos e eventos especiais considerados no traballo pode consultarse no Apéndice B.
- **Variables derivadas (*feature engineering*):** Co obxectivo de extraer un maior potencial analítico da base de datos, xeráronse variables que permiten estandarizar a información e facilitar a modelización posterior. Entre estas destacan o prezo por quilómetro (€/km), que actúa como un indicador de rendemento homoxéneo; a pertenza a un determinado clúster de cliente (obtida mediante a análise que se detalla na Sección 3.2); e as agregacións temporais en escalas semanais e mensuais, indispensables para a análise de series temporais.

Tras unha entrevista co responsable de operacións na empresa, identificouse que a demanda de envíos depende fortemente do sector do cliente e de picos estacionais, o que orientou a selección de variables e xustificou a necesidade da creación de clústeres de clientes.

Cómpre sinalar que, como a información sobre o sector ao que pertence cada cliente non estaba codificada na base de datos relacional, engadiouse de forma manual a partir do coñecemento por parte do responsable da empresa. Así, dada a elevada carteira de clientes da que dispón Trucksters, o traballo limitouse ás 100 rutas con maior volume histórico. Por conseguinte, esta decisión introduce un nesgo de selección deliberado que debe ser tido en conta á hora xeneralizar as conclusións do proxecto.

3.1.1. Filtrado dos datos orixinais

Unha vez introducidas as novas variables e antes de iniciar a análise exploratoria, resulta fundamental realizar un primeiro filtrado dos datos. O obxectivo é eliminar observacións que non reflictan un comportamento real ou que poidan introducir nesgos nos modelos, incluíndo aquelas que poden deberse a erros na recollida ou rexistro dos datos.

En primeiro lugar, identificáronse diferentes tipos de erros ou valores atípicos nos datos recollidos:

- Algúns envíos presentan prezos negativos para os clientes, representando compensacións que se lles realizan debido a problemas no reparto.
- Outros envíos teñen a distancia percorrida nula, correspondendo a casos "artificiais" creados pola empresa para representar penalizacións por cancelacións de última hora ou axustes derivados da flutuación nos prezos do combustible.
- Finalmente, observouse que os prezos por quilómetro (tanto para clientes como para provedores) superiores a 3€ son consecuencia de erros na recollida dos datos. A lóxica destes erros foi revisada co responsable de negocio, quen confirmou que se trataba de observacións que non reflectían a operativa real da empresa.

Tras esta validación, optouse por excluír todos estes rexistros da base de datos, asegurando que a información utilizada na análise exploratoria e nos modelos preditivos reflicta correctamente o comportamento real dos envíos.

Por último, débese ter en conta que a finalidade deste traballo é predicir o volume de pedidos nunha data determinada. Polo tanto, é fundamental considerar que os datos presentan unha dependencia temporal. Así, ao analizar a base de datos obsérvase que só unha pequena parte dos rexistros é anterior ao ano 2022 (concretamente o 3.2 %). Isto débese a dous factores: o forte crecemento da empresa a partir do propio ano 2022 e á falta de rexistros nos anos previos. En consecuencia, pódese asumir que

os datos anteriores ao 2022 non reflicten adecuadamente a realidade actual da empresa e que a súa inclusión podería ter un impacto negativo nos modelos que se desenvolverán. Por este motivo, decidiuse traballar coa información dispoñible a partir do 1 de xaneiro de 2022.

En relación aos datos máis recentes, a consulta da base de datos realizouse a comezos da segunda semana de outubro de 2025. Dado que a base de datos se actualiza diariamente, disponse de rexistros ata a primeira semana do devandito mes. Así, ao non contar coa información completa de outubro, optouse por empregar os datos ata a última semana de setembro de 2025, inclusive.

Tras a aplicación destes criterios de exclusión e o filtrado temporal, o conxunto de datos final para o estudo pasou de 21271 observacións iniciais a un total de 19869 rexistros. Esta mostra constitúe a base sobre a que se aplicará o estudo realizado neste traballo.

3.2. Análise clúster

Tras falar co responsable na empresa, formulouse a hipótese de que a demanda depende do tipo de cliente e o sector ao que pertence. Isto fai pensar que un modelo xenérico non sexa capaz de modelar o comportamento da demanda, o cal indica que sería práctico engadir máis detalle ao modelo. Se se pensa no nivel extremo, é dicir un modelo por cliente, atoparíanse certos problemas para clientes dos que non existe suficiente mostra, polo cal o modelo non sería estatisticamente robusto (presentaría moita variabilidade), ademais de supoñer un maior tempo computacional. Outra solución, á vista da información dispoñible, sería agrupar aos clientes directamente por sector, mais isto suporía unha perda da información do tipo de contrato. Por conseguinte, este traballo optou por un punto intermedio, agrupando por tipo de cliente tendo en conta o sector e o contrato ou acordo que teñen con Trucksters, o que permite detectar posibles relacións entre distintos tipos de clientes. Para levar a cabo esta agrupación, seguirase o procedemento explicado na Sección 2.1: **análise de correspondencias** seguido dun **algoritmo aglomerativo xerárquico**.

Por conseguinte, as variables seleccionadas para o *clustering* serán o sector do cliente e o tipo de contrato (ver Apéndice A), sendo estas variables categóricas. Por unha banda, cómpre sinalar que os clientes clasifícanse en 7 sectores distintos: alimentación, electrónica, froita, industria, paquetería, transporte e *retail*. Por outra banda, existen 2 tipos de contratos entre estes clientes e Trucksters: fixo (*fix*) ou puntual (*spot*). Os clientes con contrato fixo presentan unha demanda máis estable por existir un acordo con Trucksters, mentres que para os clientes con contrato puntual a demanda presenta maior variabilidade.

Tendo en conta o tipo de variables consideradas para o *clustering*, é fundamental percatarse que os algoritmos de formación de grupos mediante aglomeración dos individuos baséanse no cálculo de distancias. Como estas variables son categóricas, para poder aplicar o algoritmo correspondente, previamente é necesario facer unha análise de correspondencias (CA) (ver Sección 2.1.1). Dita técnica de análise multivariante permite estudar a relación entre dúas variables categóricas e proxectalas nun espazo de menor dimensionalidade. Así, as coordenadas resultantes do CA serán os valores cos que se calcularán as distancias na análise clúster. O primeiro paso para levar a cabo a análise de correspondencias consiste en calcular a táboa de continxencia das variables de interese. Ditas táboas reflexan a distribución conxunta do par formado polas variables categóricas. O Cadro 3.1 amosa a distribución conxunta para as variables de estudo.

Á hora de estudar a relación entre dúas variables, a situación máis extrema prodúcese cando as variables son independentes. En consecuencia, antes de proceder ca CA, é esencial verificar a existencia dunha relación estatisticamente significativa entre as variables categóricas. Empregarase para iso o test de independencia **chi-cadrado** (χ^2) detallado na Sección 2.1.1. No código complementario ao traballo (consúltese Airas, 2026) pode verse que, fixando un nivel de significación do 5 %, rexéitase a hipótese de independencia. Esta relación contrastada xustifica a aplicación da análise de correspondencias.

Sector	**Fixo (*fix*)**	**Puntual (*spot*)**	**Total**
Alimentación	1063	111	1174
Electrónica	781	490	1271
Froita	572	2	574
Industria	1173	188	1361
Paquetería	9535	211	9746
Retail	1878	1.661	3539
Transporte	1233	971	2204
Total	**16235**	**3634**	**19869**

Cadro 3.1: Táboa de continxencia: Frecuencias de pedidos por sector e tipo de contrato. A distribución amosa unha clara predominancia de contratos fixos en sectores como a paquetería, mentres que no *retail* a modalidade *spot* ten un peso relativo moito maior.

Cómpre recordar que a análise de correspondencias consiste en aplicar unha análise de compoñentes principais (PCA) aos perfís de fila (ou columnas) estandarizados, calculados a partires da táboa de continxencia (Cadro 3.1). Tendo en conta que ditos perfís cumpren a condición de que suman 1 por columnas (ou por filas), o número máximo de dimensións que aportan información será o mín$(I - 1, J - 1)$, sendo I e J o número de categorías das filas e columnas, respectivamente. Daquela, como as variables de interese constan de 7 sectores e 2 tipos de contrato, a representación das compoñentes resultantes será unidimensional.

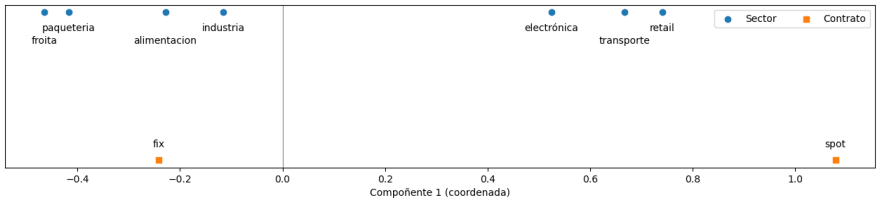


Figura 3.1: Proxección conxunta dos perfís de fila e columna resultante da análise de correspondencias. A proximidade entre puntos reflicte a asociación entre as categorías de ambas variables.

Na Figura 3.1 obsérvase a representación gráfica do resultado da CA levada a cabo. Como xa se indicou, a primeira compoñente concentra a totalidade da inercia entre o sector e o tipo de contrato. Así, vese como os sectores de electrónica, transporte e *retail* están fortemente asociados a non ter un acordo asinado. Isto suxire que os clientes de ditos sectores traballan coa empresa de forma menos regular e a prezos de mercado. Por outra banda, o resto de sectores están asociados a un contrato fixo. Serán entón clientes cos cales Trucksters traballa baixo algún tipo de acordo (número de rutas, volume semanal...), co cal espérase que a demanda sexa máis estable ou previsible.

Agora ben, as coordenadas obtidas mediante a CA empregaranse como *inputs* na análise clúster.

A construción dos grupos levouse a cabo mediante un algoritmo aglomerativo considerando o método do máximo (véxase Sección 2.1.2), que define a distancia entre clústeres como a maior distancia entre os individuos dos grupos. O paso crítico no algoritmo reside na selección do número de grupos (K) que se forman. Observando o dendrograma do *clustering* (Figura 3.2), o corte considerouse no maior salto de distancia entre unións (liña punteada), dando lugar a tres clústeres ben diferenciados.

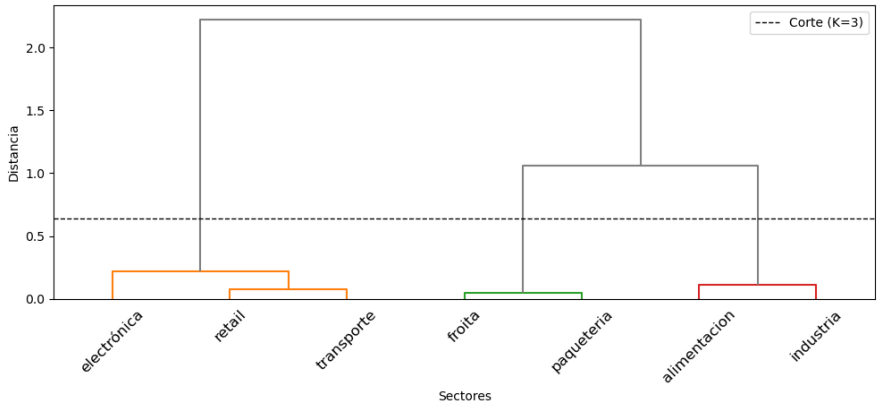


Figura 3.2: Dendrograma resultante do procedemento de clúster xerárquico. A árbore ilustra as etapas de fusión das observacións ata a formación de $K = 3$ grupos homoxéneos. A liña discontinua marca a altura de corte considerada.

Os grupos formáronse, en particular, agrupando os distintos sectores, tal e como se describe a continuación:

- Clúster 1: formado polos clientes dos sectores de electrónica, *retail* e transporte. Este grupo consta de 7014 observacións, representando un 35.3 % da totalidade dos datos.
- Clúster 2: formado polos clientes de sectores de froita e paquetería, resulta o grupo máis representativo cun total de 10320 observación (51.9 % da mostra).
- Clúster 3: conformado polos clientes adicados á alimentación e á industria. Este grupo tan só presenta 2525 pedidos, representando un 12.8 % do total.

Nas seccións seguintes, centrarase o estudo no comportamento das distintas variables dispoñibles en cada un dos clústeres por separado, o que aportará unha maior información dos datos cos que se traballa neste proxecto.

3.3. Análise descritiva

Unha vez definida a estrutura da carteira de clientes mediante a análise de clústeres, o seguinte paso metodolóxico consiste na caracterización detallada de cada grupo. Esta fase de análise exploratoria non só pretende describir o estado actual dos datos, senón identificar as dinámicas subxacentes que determinarán o comportamento das series temporais de demanda e prezos.

Un aspecto fundamental ao realizar unha análise exploratoria dos datos reside no estudo das correlacións entre as variables. Tendo en conta a natureza continua das variables de interese nesta sección, traballouse co coeficiente de correlación de Pearson (ρ). Cómpre recordar que dito coeficiente é unha medida da relación lineal entre pares de variables, calculado como o cociente entre a covarianza e o produto das desviacións típicas de ambas magnitudes.

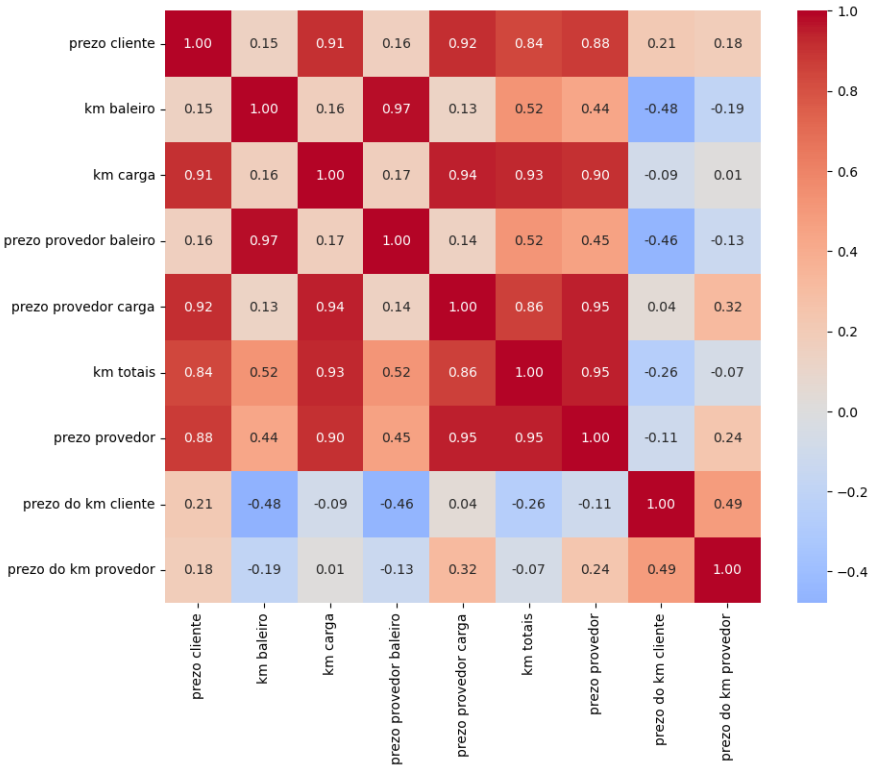


Figura 3.3: Matriz de correlación de Pearson das variables numéricas. Obsérvase unha colinealidade severa entre as variables orixinais de distancia e prezo ($\rho > 0,80$), que se ve mitigada ao empregar as variables ratio (prezos por quilómetro).

A Figura 3.3 representa graficamente a matriz de correlacións. Nela obsérvase un problema de colinealidade severo entre o grupo de variables orixinais. Por exemplo, entre as variables que representan distancias e prezos totais, os coeficientes de correlación son superiores a 0.80. Isto indica que ditas variables recollen información redundante.

Por outra banda, é importante sinalar que as variables derivadas (calculadas a partir dos datos orixinais seguindo os criterios da Sección 3.1) mitigan esta colinealidade. En efecto, obsérvase que as variables que representan os prezos por quilómetro presentan coeficientes de correlación moito máis moderados ($-0,48 < \rho < 0,49$). Por conseguinte, esta análise xustifica centrar o estudo descritivo detallado nas variables derivadas, evitando así a duplicidade de información e garantindo unha análise máis eficiente e robusta.

Prezo por quilómetro.

En primeiro lugar, estudarase a distribución do prezo por quilómetro, diferenciando entre clientes e provedores.

Dende unha perspectiva de negocio, Trucksters xestiona transporte tanto con flota propia como con colaboradores externos. Cada colaborador ten un prezo ou custo por quilómetro percorrido. A nivel contractual, este prezo adoita ser fixo. Porén, en períodos de alta demanda, cando a empresa precisa aumentar a capacidade, pode incorporar colaboradores adicionais a un custo superior para cubrir os picos, xerando así variacións no prezo do provedor.

De forma análoga, Trucksters distingue entre dous tipos de clientes: fixos e puntuais. Por unha

banda, os clientes con contrato fixo (*fix*) constitúen rutas que se pactan a longo prazo, normalmente anual, cun prezo previamente acordado. Por outra banda, os clientes con contrato puntual (*spot*) presentan prezos que se fixan segundo a dinámica do mercado en cada servizo.

En consecuencia, tanto as flutuacións no prezo do provedor coma as do prezo dos clientes poden reflectir cambios na demanda. En particular, un aumento de prezos pode indicar o inicio da *peak season*, pois requerirase de maior capacidade e os custos axústanse á situación de mercado. Por conseguinte, o prezo por quilómetro pode aportar información sobre a serie da demanda, xustificando así o seu estudo.

Dende unha perspectiva analítica, na Figura 3.4 obsérvase, graficamente, o comportamento do prezo por quilómetro de todos os clientes. Á esquerda represéntase un histograma que permite ver que a variable ten unha forma unimodal, presentando unha forte concentración dos datos arredor da tendencia central. En efecto, o rango intercuartílico (que contén o 50 % das observacións) é pequeno (IQR = 0.22 €/km). Respecto á simetría, a gráfica amosa unha certa asimetría positiva, apoiada no feito de que a media (1.20 €/km) é lixeiramente superior á mediana (1.19 €/km). Isto indica que a cola dereita contén os valores extremos máis influentes. No diagrama de caixas da dereita (*boxplot*) reafírmase esta asimetría positiva, vendo como se teñen máis datos extremos na parte dereita (os que caen fóra do "bigote").

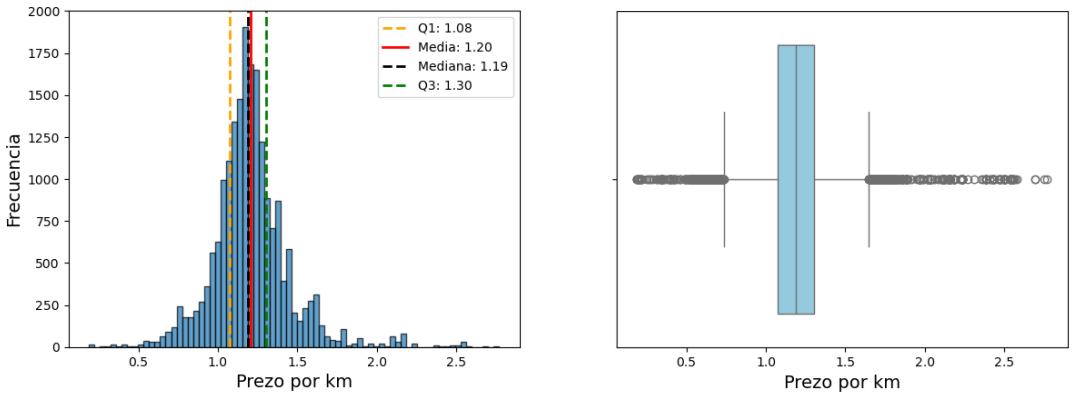


Figura 3.4: Análise exploratoria do prezo por quilómetro para os clientes. O histograma amosa a frecuencia e concentración dos prezos, mentres que o *boxplot* permite visualizar a dispersión e os prezos significativamente afastados da media.

A análise clúster levada a cabo na Sección 3.2 permite estudar o comportamento desta variable en cada un dos tres clústeres. A Figura 3.5 representa, na gráfica superior, os histogramas do prezo por quilómetro dentro de cada clúster, mentres que a gráfica inferior representa unha comparación entre os tres clústeres mediante un *boxplot*. Así, obsérvanse certas diferenzas tanto na tendencia central coma na forma da distribución.

Para os clientes de electrónica, *retail* e transporte (clúster 1) vense os prezos máis baixos cunha media de 1.15 €/km que coincide neste caso coa mediana. É salientable a presente simetría, que contrasta co comportamento global e co que sucede nos outros dous clústeres. Porén, este grupo é o que presenta un rango intercuartílico maior (IQR = 0.27 €/km), o que suxire unha maior heteroxeneidade na distribución dos prezos. Pódese pensar que, o motivo da dispersión e da tendencia máis barata dos prezos presente resida nos clientes do sector de transporte, que se trataban de cargas revendidas por outra empresa de transporte, co cal os prezos non resultan tan competitivos.

Para os clientes de froita e paquetería (clúster 2), a media (1.23 €/km) é lixeiramente superior á mediana (1.20 €/km) presentando daquela unha certa asimetría positiva. Estatisticamente, é o grupo

máis homoxéneo, co rango intercuartílico máis curto dos tres clústeres (IQR = 0.20 €/km).

Por último, os clientes dos sectores de alimentación e industria (clúster 3) presentan os prezos medios máis caros, e a asimetría positiva máis clara, sendo neste caso a media (1.26 €/km) e a mediana (1.21 €/km). Así, este comportamento pode responder seguramente a clientes cos que se acordou algunha ruta con prezos moi competitivos.

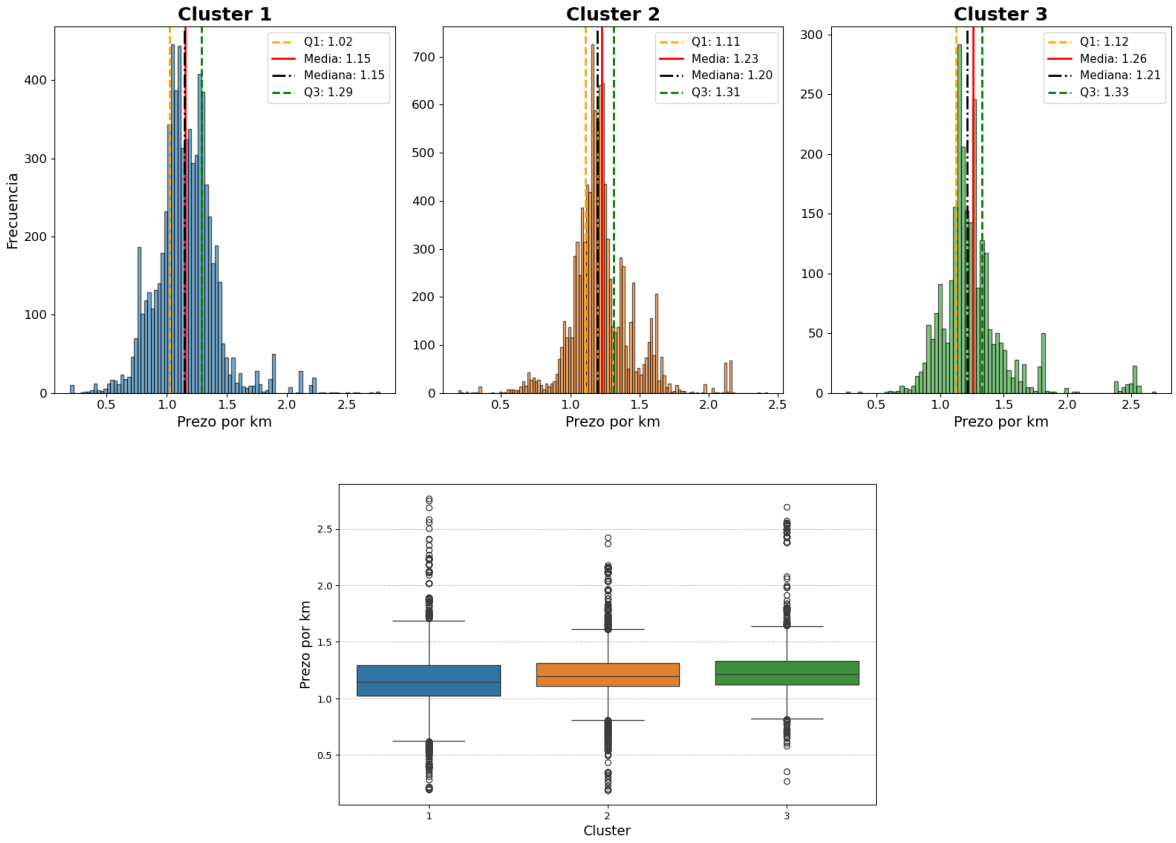
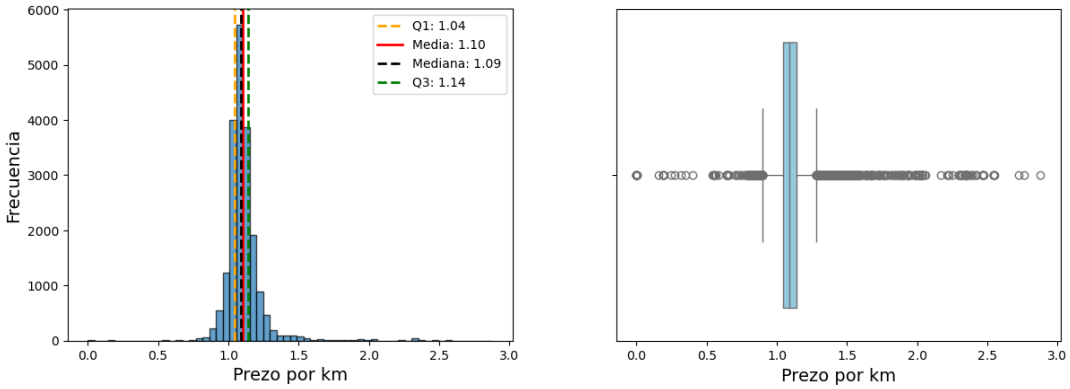


Figura 3.5: Análise exploratoria do prezo por quilómetro para os clientes segmentado por clúster. A figura superior mostra os histogramas e a inferior os boxplots. Clúster 1 (azul), clúster 2 (laranxa) e clúster 3 (verde).

Unha vez estudada a distribución dos prezos para os clientes, é preciso facer o propio cos prezos dos provedores, estudando en primeiro lugar o comportamento global e vendo posteriormente se hai diferenzas no comportamento entre clústeres.

Na Figura 3.6 obsérvase mediante o histograma (Figura (3.6a)) e o diagrama de caixas (Figura (3.6b)) como a distribución dos prezos por quilómetro para os provedores presenta unha clara forma unimodal, cunha concentración en torno á media todavía superior ao caso dos clientes. En termos da tendencia, os prezos resultan inferiores, sendo agora a media 1.10 €/km fronte os 1.20 €/km para os clientes. En cuanto á forma, a variable presenta certa asimetría positiva (mediana 1.09 €/km) e, como xa se sinalou, os prezos están todavía máis concentrados neste caso, sendo o rango intercuartílico (IQR) 0.10 €/km. Así, obsérvase unha maior estabilidade dos prezos no caso dos provedores.

Facendo agora o estudo desagregado por clúster, na Figura 3.7 represéntanse os histogramas para cada clúster na parte superior, e os *boxplot* na gráfica inferior. Mediante estas representacións gráficas, destácase a similitude do comportamento dos prezos dos provedores para os dous primeiros clústeres, en



(a) Histograma (b) Boxplot

Figura 3.6: Análise exploratoria do prezo por quilómetro para os provedores. O histograma amosa a frecuencia e concentración dos prezos, mentres que o *boxplot* permite visualizar a dispersión e os prezos significativamente afastados da media.

contra do que sucedía no caso dos prezos dos clientes, onde se presenciaban maiores diferenzas entre os clientes do clúster 1 e o resto. Porén, neste caso o clúster 3 (alimentación e industria) presenta o prezo en media máis alto (1.15 €/km), sendo aproximadamente 0.06 €/km superior aos outros clústeres. Ademais da tendencia central lixeiramente superior, cómpre salientar para este tipo de clientes unha maior dispersión dos prezos, presentando o maior rango intercuartílico (IQR = 0.13 €/km).

En síntese, a segmentación estratéxica mediante a análise de clústeres permite validar a existencia de patróns de comportamento diferenciados na política de prezos. No que respecta aos clientes, os resultados suxiren que a tendencia e a volatilidade do prezo están fortemente condicionadas pola modalidade contractual predominante en cada segmento (fixa ou puntual). Pola contra, a estrutura de custos asociada aos provedores amosa unha maior dependencia do sector de actividade do cliente, independentemente do tipo de acordo comercial subxacente. Esta caracterización resulta fundamental para o obxecto do presente traballo: a heteroxeneidade detectada nos ratios de prezo, confirma que un modelo de predición global sería insuficiente.

Evolución temporal do prezo por quilómetro

Tras caracterizar a distribución estática das ratios de prezo, resulta imperativo abordar o seu comportamento desde unha perspectiva dinámica. Dada a dependencia temporal intrínseca dos rexistros e o obxectivo de predicir o volume de envíos, a análise da evolución cronolóxica das variables clave é un paso crítico. Este procedemento non só busca identificar as compoñentes estruturais clásicas (tendencia, estacionalidade e ciclos), senón tamén avaliar a heteroxeneidade destas dinámicas entre os clústeres identificados. Comprender se as flutuacións de prezos seguen un patrón simultáneo ou se, pola contra, cada segmento de clientes responde a ciclos de mercado diferenciados, será determinante para a especificación e comprensión dos modelos preditivos que se desenvolverán no Capítulo 4.

Para este fin, calcúlanse os prezos medios mensual e diariamente, sendo fundamental sinalar que, con motivo de mellorar a interpretación, para a visualización diaria empregaranse os datos diarios suavizados (véxase Airas, 2026). De igual forma que no apartado anterior, estúdase en primeiro lugar o caso dos clientes.

A Figura 3.8 representa a evolución temporal no caso dos clientes tanto mensualmente (Figura (3.8a)), como diariamente (Figura (3.8b)). De forma xeral, obsérvase unha clara evolución crecente dos prezos ao longo destes últimos anos.

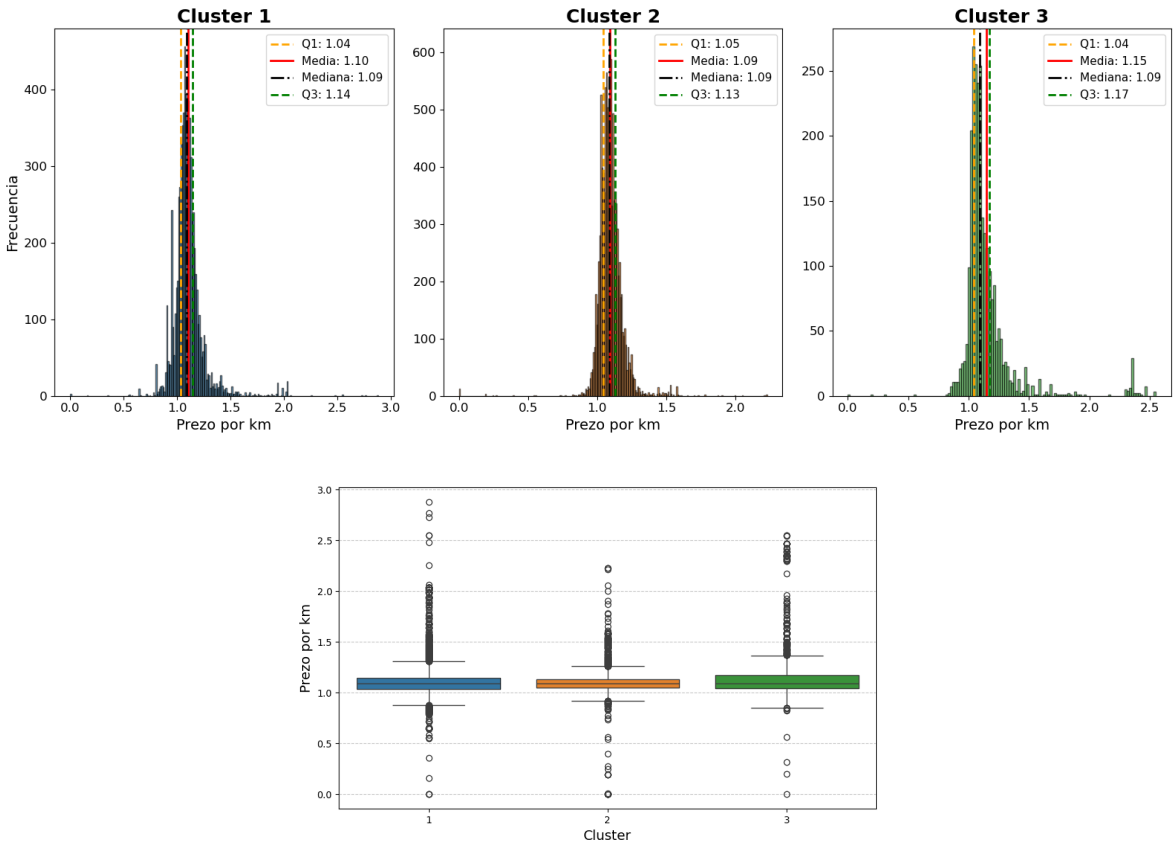
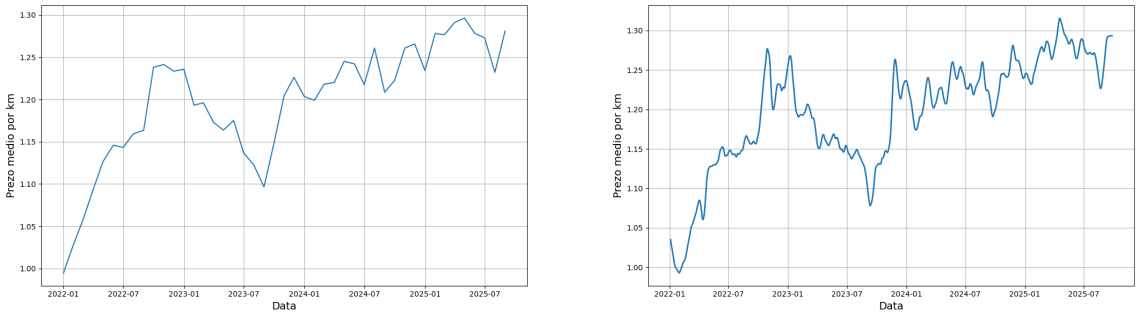


Figura 3.7: Análise exploratoria do prezo por quilómetro para os provedores segmentado por clúster. A figura superior mostra os histogramas e a inferior os boxplots. Clúster 1 (azul), clúster 2 (laranxa) e clúster 3 (verde).



(a) Mensualmente

(b) Diariamente

Figura 3.8: Evolución temporal do prezo medio por quilómetro para os clientes (2022-2025). Á esquerda represéntase a serie agregada por meses, mentres que á dereita móstrase o prezo diario suavizado. Obsérvase tendencia crecente dos prezos.

O ano 2022 comeza cunha forte tendencia crecente, partindo de prezos bastante baixos (1 €/km), chegando durante o mes de novembro a prezos medios superiores a 1.25 €/km. Este aumento tan significativo dos prezos é consecuencia do efecto rebote que sufriu o sector tras a caída provocada pola

pandemia dous anos atrás.

Durante a primeira metade do ano 2023, obsérvase a única fase de descenso dos prezos nos datos dispoñibles, caendo ata valores medios inferiores aos 1.10 €/km. Tras isto, os prezos recuperan a tendencia á alza, sendo salientable novamente o efecto que causa nos mesmos o Black Friday, provocando que os prezos medios máximos do ano sucedan novamente neste último trimestre.

A partires do 2024, a tendencia dos prezos segue a ser creciente pero máis estable ca nos anos previos. Este aumento constante dos últimos dous cursos fundaméntase no crecemento da empresa, que presenta unha posición de maior poder no mercado, conseguindo traballar a prezos máis beneficiosos para a mesma.

Para un estudo máis detallado, será de interese repetir a análise para cada clúster por separado. Así, a Figura 3.9 amosa as tres series temporais resultantes, revelando patróns de tendencia e estacionalidade distintos. A maior discrepancia atópase entre os clientes de alimentación e industria (clúster 3) con respecto ao resto, posto que durante o último trimestre do 2022 presentan un pico que chega a superar os 1.6 €/km. Consultando esta información co responsable de negocio, identificouse un grupo de viaxes a finais de 2022 correspondentes a envíos procedentes de España con dirección Reino Unido. En consecuencia, o elevado prezo está motivado polas aduanas (traballo extra), así como o eurotúnel (custo extra). Cómpre sinalar que este tipo de exportacións deixaron de levarse a cabo tras estes meses debido á pouca rendabilidade das mesmas, motivo polo cal non se aprecian picos tan elevados noutros anos.

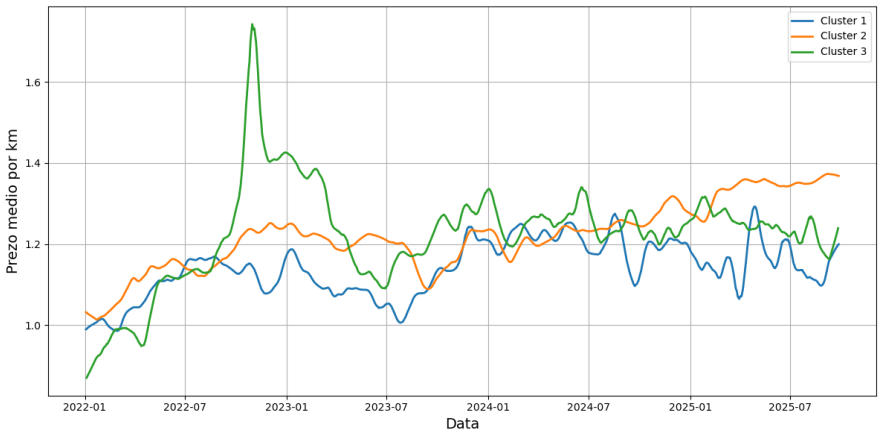
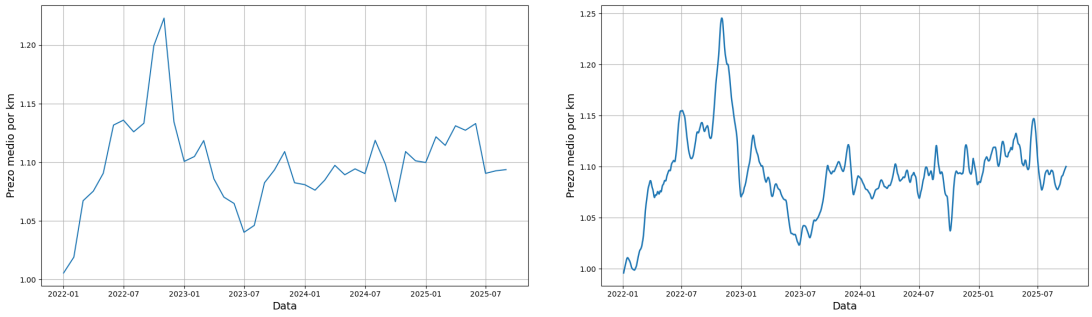


Figura 3.9: Evolución temporal do prezo medio por quilómetro para os clientes segmentado por clústers (2022-2025). Os clústeres diferéncianse por cores: clúster 1 (azul), clúster 2 (laranxa) e clúster 3 (verde).

Este fenómeno concorda co observado ao estudar a distribución dos prezos na Figura 3.5, onde se apreciaba que o clúster 3 presentaba unha asimetría positiva considerable. Posteriormente, os prezos dos clientes de alimentación e industria caen e aseméllanse ao comportamento do clúster 1, conformado polos clientes de transporte, *retail* e electrónica. Por outra banda, do clúster formado polos clientes dos sectores de froita e paquetería obsérvase unha tendencia crecente e estable durante os últimos dous anos, provocando que nas últimas datas das que se teñen datos presenten prezos medios claramente superiores ao resto (1.4 €/km).

Para os provedores, a Figura 3.10 ilustra a evolución temporal do prezo medio por quilómetro mensualmente (Figura (3.10a)) e diariamente (Figura (3.10b)). De forma similar aos clientes, durante o ano 2022 obsérvase unha clara tendencia crecente que alcanza o pico máximo no mes de novembro, dándose prezos medios cercanos a 1.25 €/km. Como xa se indicou, esta tendencia inicial vén explicada seguramente pola recuperación do sector trala pandemia. Posteriormente os prezos caen de forma

considerable ata comezos do ciclo seguinte.



(a) Mensualmente (b) Diariamente

Figura 3.10: Evolución temporal do prezo medio por quilómetro para os provedores (2022-2025). Á esquerda represéntase a serie agregada por meses, mentres que á dereita móstrase o prezo diario suavizado. Obsérvase tendencia decrecente dos prezos durante os últimos anos.

O ano 2023 establece un patrón de maior estabilidade, aínda que se observa unha tendencia lixeiramente decrecente durante a primeira metade do ano. Novamente, neste curso obsérvase o efecto que ten o Black friday nos datos, pois os prezos crecen ata rematar o mes de novembro e caen durante decembro. Non obstante, a oscilación dos prezos neste ano é menor á observada no curso anterior, sen chegar a superar en ningún intre os 1.15 €/km.

Os valores durante o 2024 mantéñense nun rango moito máis constante (próximo aos 1.10 €/km), sendo salientable a caída que se produce no mes de outubro, aspecto que parece influír no impacto que ten neste ano o Black Friday, sendo os picos de agosto nalgúns momentos superiores aos de novembro. Como xa se explicou, este comportamento tan diferente aos anos previos vén explicado polo crecemento de Trucksters, presentando unha posición de mercado máis forte e podendo negociar prezos máis competitivos.

Da mesma maneira ca no caso dos clientes, cómpre facer a análise segmentada pos clúster e ver así as diferenzas nos prezos medios dos provedores en función dos tipos de clientes.

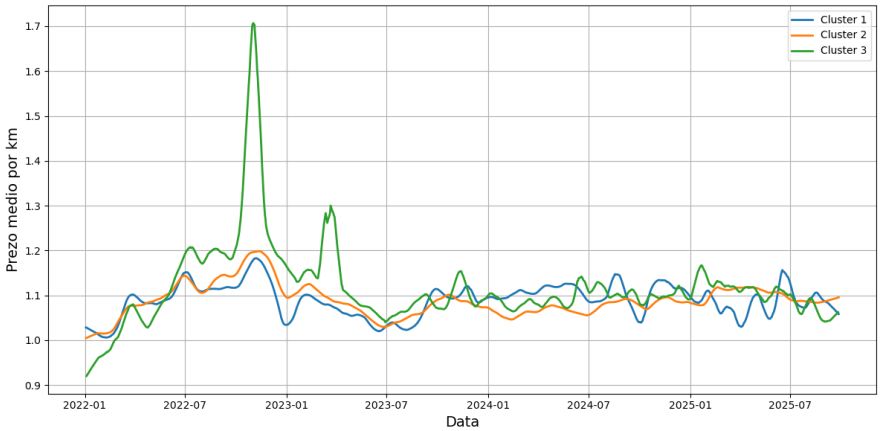
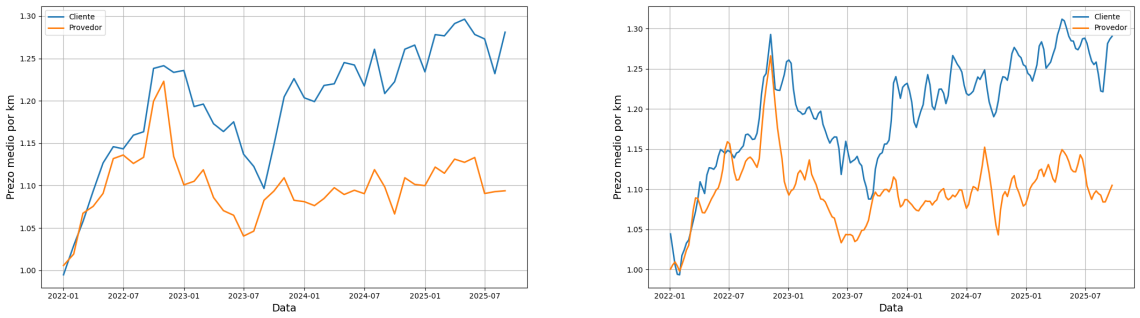


Figura 3.11: Evolución temporal do prezo medio por quilómetro para os provedores segmentado por clústers (2022-2025). Os clústers diferéncianse por cores: clúster 1 (azul), clúster 2 (laranxa) e clúster 3 (verde).

Mediante a Figura 3.11 represéntanse as tres series temporais. Consecuentemente aos prezos dos

clientes, obsérvase un pico extremo a finais do 2022 para o clúster 3, chegando a valores próximos a 1.70 €/km. Tras isto, obsérvase unha caída moi rápida nos prezos de dito clúster, mais a diferenza dos prezos dos clientes, existe un pequeno repunte durante o primeiro trimestre de 2023. Ademais disto, neste caso apenas se notan máis diferenzas nos prezos dos provedores entre clúster amosando para todos os clústeres un comportamento estable en torno aos 1.10 €/km dende mediados do 2023.

Daquela, obsérvase como os prezos medios para os provedores presentan unha menor volatilidade ca os prezos para os clientes. Esta idea pode contrastarse representando nunha mesma gráfica a tendencia temporal dos prezos para os clientes e os provedores.



(a) Mensualmente (b) Diariamente

Figura 3.12: Comparativa da evolución do prezo medio por quilómetro entre clientes e provedores (2022-2025). Á esquerda represéntase a serie mensual e á dereita a diaria con suavizado temporal. Obsérvase un incremento na marxe bruta a partir de 2024, cunha estabilización dos custos de transporte fronte ao aumento das tarifas de venda.

En efecto, a Figura 3.12 presenta a comparación directa da evolución histórica do prezo medio por quilómetro para os clientes e os provedores. A análise mostra unha diverxencia clara nas tendencias a partires da segunda metade do 2023. Mentres os prezos dos provedores mostran unha tendencia á estabilización arredor de 1.10 €/km, os clientes manteñen a tendencia crecente, sendo a diferenza nos últimos meses próxima a 0.20 €/km.

Facendo agora o propio para cada clúster (Figura 3.13), afínase un pouco a interpretación. Daquela, cómpre sinalar que as diferenzas dos prezos medios observadas débense especialmente aos clientes de froita e paquetería. Para os sectores que forman este clúster, os prezos dos clientes manteñen, dende a segunda metade do 2023, a tendencia crecente ata superar os 1.35 €/km nas últimas datas, sendo os dos provedores próximos a 1.10 €/km. Porén, é notable sinalar como tanto para o clúster 1, que presenta maiores flutuacións nos prezos, coma no clúster 3, máis estable tralo pico xa sinalado, o comportamento dos prezos para clientes e provedores apenas difire.

En definitiva, a análise cronolóxica revela un desacoplamento significativo entre as series de prezos de clientes e provedores a partir da segunda metade de 2023. Esta diverxencia, especialmente acentuada no clúster 2 (froita e paquetería), constitúe unha evidencia empírica da consolidación de Trucksters no mercado loxístico. A capacidade de manter unha tendencia ascendente nos prezos de venda mentres se estabilizan os custos de subcontratación amosa unha optimización da rendibilidade estratéxica. Desde o prisma da modelización, este fenómeno é de vital importancia: confirma que a dinámica de prezos non responde a un proceso estacionario único, senón que está influenciada por cambios estruturais na posición competitiva da empresa.

Evolución temporal do volume de pedidos

Como culminación desta fase exploratoria, abórdase o estudo da variable obxectivo central deste traballo: o volume de envíos. Tras analizar os prezos dos clientes e provedores, resulta fundamental

Figura 3.13: Comparativa da evolución do prezo medio por quilómetro entre clientes e provedores segmentado por clúster (2022-2025). Obsérvanse maiores diferenzas no clúster 2 (gráfico intermedio), fronte a maior similitude no clúster 1 (superior) e clúster 3 (inferior).

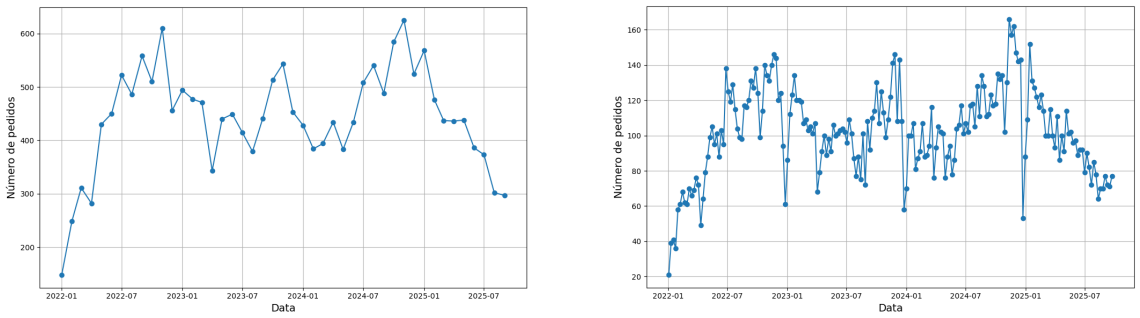
caracterizar o comportamento dinámico da demanda, xa que estas series constituirán o obxecto de predición nos modelos que se desenvolverán no Capítulo 4.

Para iso, mediante a Figura 3.14, represéntase a evolución mensual (Figura (3.14a)) e semanal (Figura (3.14b)). Por outra banda, a Figura 3.15 permite afondar no comportamento histórico ao representar a evolución semanal desagregada por ano.

Con respecto á tendencia, a serie comeza no 2022 cunha clara tendencia crecente, consecuente coa fase de crecemento da empresa e recuperación do sector trala pandemia. Non entanto, os anos 2023 e 2024 amosan unha estabilización do nivel de volume, observando un comportamento máis cíclio. O comportamento da serie durante o ano 2025 difire dos cursos anteriores, presenciándose unha tendencia lixeiramente decrecente. A pesares de que isto poda parecer estraño, a explicación reside nas políticas de mercado da empresa. En particular, como consecuencia do crecemento de Trucksters nos anos previos, a estratexia seguida durante o presente ano céntrase en minorar o volume de envío, mais incrementando a rentabilidade dos mesmos. Este feito concorda co comportamento dos prezos medios estudados anteriormente (véxanse Figuras 3.12 e 3.13).

En cuanto á estacionalidade, pese a non observar comportamentos con exactamente os mesmos patróns anualmente, si se nota o impacto do Black Friday, provocando que as últimas semanas de novembro constitúan o período de maior volume de envíos cada ano. Porén, en contra do comportamento esperado por intuición durante o Nadal, no mes de decembro obsérvase unha caída salientable. En particular, nos anos 2023 e 2024 ditas semanas constitúen o volume mínimo de pedidos anual.

Pola súa parte, a Figura 3.15 confirma a alta variabilidade semanal e permite ver con maior claridade o inicio da *peak season* nos anos 2023 e 2024, principalmente. Porén, é natural cuestionarse se o impacto do Black Friday ten a mesma importancia en todos os sectores.



(a) Volume mensual

(b) Volume semanal

Figura 3.14: Evolución do volume de envíos: comparativa entre agregación mensual (esquerda) e semanal (dereita). A representación semanal pon de manifesto a volatilidade e os patróns cíclicos da demanda que a agregación mensual tende a ocultar.

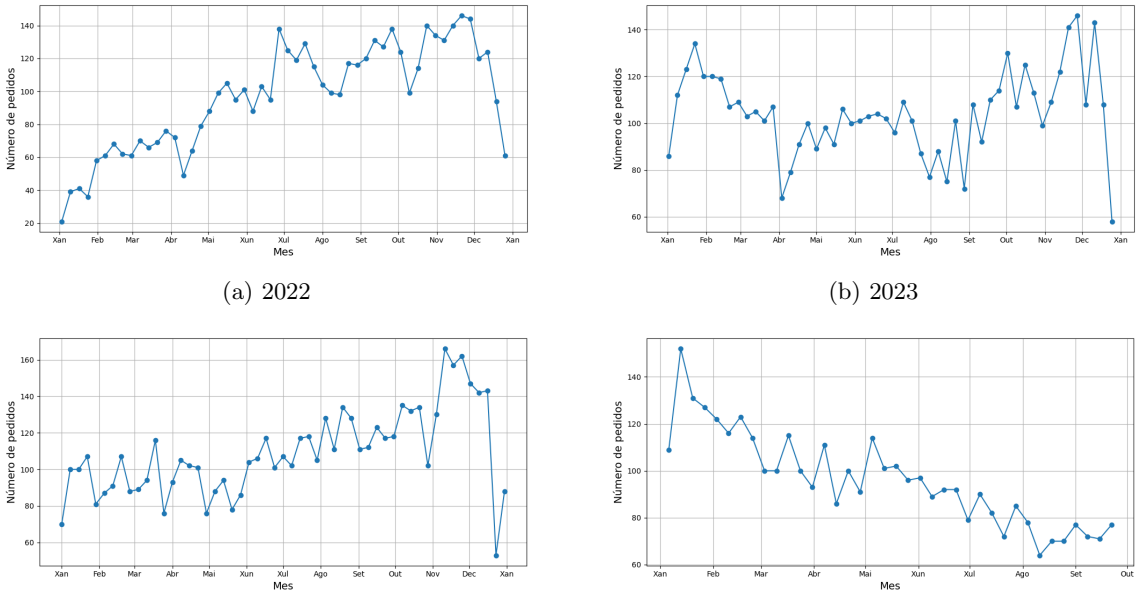
Por conseguinte, unha vez feita esta primeira análise xeral e seguindo coa metodoloxía do proxecto até o momento, repítese este estudo para cada grupo de clientes desagregado. Para iso, mediante a Figura 3.16 represéntase o volume de envíos por semana para cada clúster no espazo temporal considerado no proxecto.

Esta representación gráfica permite ter un coñecemento máis refinado do comportamento do volume de envíos. En particular, con respecto ao impacto do Black Friday, o clúster máis influenciado por dito evento é o conformado polos sectores de froita e paquetería (curva laranxa). En efecto, nos anos 2022 e 2024 dito clúster mostra un crecemento significativo a comezos do mes de novembro, alcanzando máximos absolutos de volume (100 envíos semanais en 2022 e máis de 80 en 2024). Ademais destes picos, o devandito clúster mantén un volume base máis alto e estable ca os outros grupos durante a maior parte do ano.

Por outra banda, sobre o clúster 1 (sectores de transporte, *retail* e electrónica) cómpre sinalar a súa gran variabilidade, presentando flutuacións semanais máis significativas (serie azul). O efecto do Black Friday non é tan acusado neste grupo, aínda que no ano 2023 si se observa un pico a finais de novembro, que repunta tamén en decembro. Recórdese que a dispersión no comportamento deste clúster pode estar asociado á ausencia de contrato por parte da empresa cos clientes dos sectores correspondentes. Ademais disto, é notable a caída considerable do volume durante o ano 2025, sendo en particular o clúster cunha tendencia máis descendente no presente curso.

Por último, os clientes dos sectores de alimentación e industria (curva verde) conforman o clúster con volume base menor, operando principalmente entre 10 e 30 pedidos semanais. Neste grupo, o efecto da *peak season* é residual e a súa tendencia é bastante constante, sendo sectores máis estables. Porén, durante o 2025 obsérvanse picos pronunciados (superando os 40 envíos semanais) en varias semanas do primeiro trimestre. Este aspecto terá gran influenza á hora de avaliar os modelos que se axustarán no Capítulo 4.

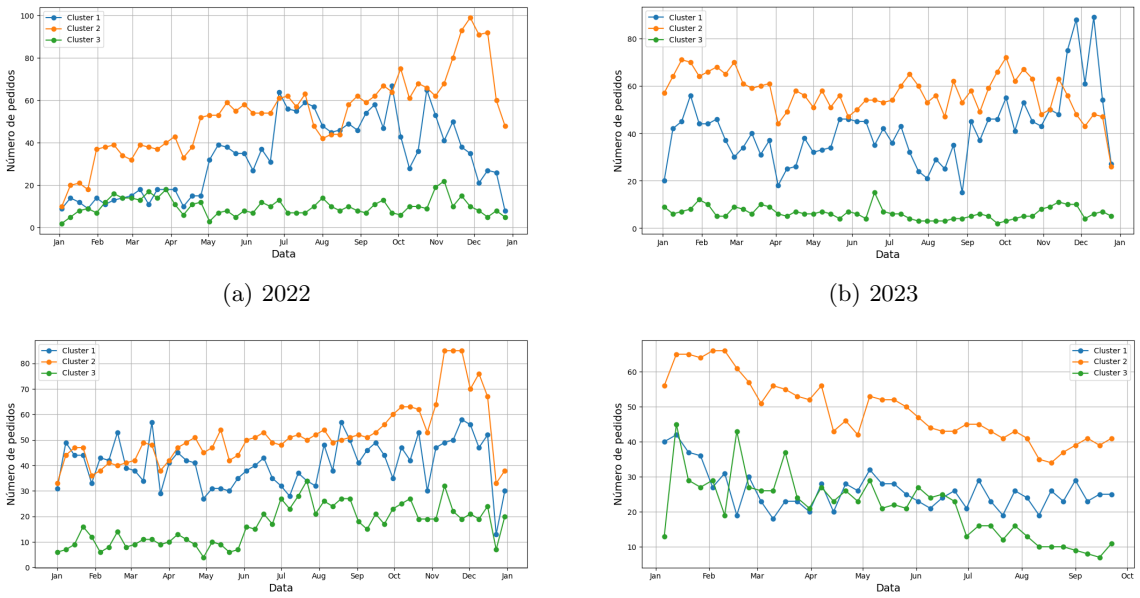
En resumo, o estudo cronolóxico do volume de envíos confirma a elevada heteroxeneidade da demanda, validando a segmentación estratéxica realizada na Sección 3.2. A disparidade observada demostra que a estacionalidade e as tendencias non son factores transversais a toda a carteira, senón que responden á idiosincrasia operativa de cada sector. Así mesmo, o cambio de tendencia detectado no ano 2025 reflicte o éxito do axuste estratéxico cara á rendibilidade comercial, o que supón un reto adicional para a modelización preditiva.



(a) 2022 (b) 2023

(c) 2024 (d) 2025

Figura 3.15: Comparativa anual do volume de pedidos semanal (2022-2025). Obsérvanse patróns estacionais recorrentes e tendencia decrecente no 2025.



(a) 2022 (b) 2023

(c) 2024 (d) 2025

Figura 3.16: Comparativa anual do volume de pedidos semanal segmentada por clúster (2022-2025). O clúster 2 (laranxa) amosa o maior impacto da *peak season* e o nivel base máis elevado.

3.4. Detección e impacto dos atípicos

Unha fase crítica na análise exploratoria consiste na identificación e tratamento de valores atípicos ou *outliers*. Nun contexto de modelización estatística, estas observacións representan desviacións significativas respecto ao comportamento xeral da mostra que, de non seren xestionadas, poderían comprometer a robustez e a capacidade de xeneralización dos modelos preditivos. O obxectivo deste apartado é identificar ditas anomalías, analizar a súa natureza e avaliar o seu impacto nas variables clave anteriormente descritas.

Cómpre distinguir este proceso do cribado inicial realizado na Sección 3.1.1. Namentres aquel filtrado se centrou na eliminación de erros de rexistro e operacións non reais (como os prezos superiores a 3 €/km), a presente sección aborda a detección de anomalías estatísticas derivadas da propia dinámica do mercado.

Tendo en conta a análise destas variables levada a cabo durante o presente capítulo, presentando as distribucións colas pesadas e certa asimetría positiva, non é recomendable empregar os criterios clásicos de detección de atípicos (IQR ou *Z-score*). Daquela, optouse por unha metodoloxía non paramétrica como é o algoritmo *Isolation Forest* (véxase Sección 2.2), que resulta máis eficaz illando anomalías en distribucións complexas sen asumir normalidade.

A eficacia da detección mediante este algoritmo reside na correcta especificación do hiperparámetro de **contaminación** (*contamination*), que define a proporción esperada de observacións anómalas no conxunto de datos. A selección deste limiar require un compromiso cauteloso: unha eliminación excesiva de rexistros podería inducir a unha perda artificial de variabilidade, derivando en modelos con risco de sobreaxuste (*overfitting*) que non representarían fielmente a volatilidade real do mercado.

Partindo do suposto de que os atípicos representan unha pequena proporción dos datos, o valor da contaminación fixouse en 0.0235 (2.35 %). Esta cifra non é arbitraria, senón que responde a un proceso de axuste iterativo baseado na distribución empírica dos prezos. Tras realizar probas de sensibilidade, observouse que este valor identifica con precisión aquelas observacións que exceden os rangos operativos validados pola empresa, sen comprometer a estrutura das series temporais. Deste xeito, garántese que o filtrado actúe exclusivamente sobre as colas máis extremas e potencialmente erróneas da distribución.

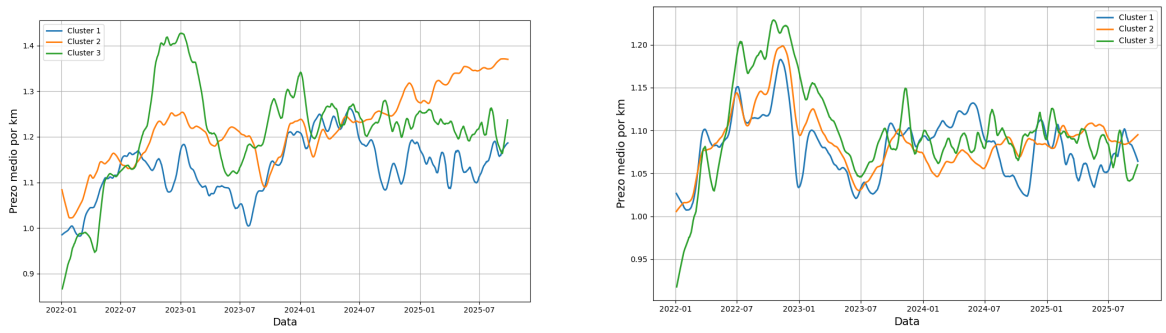
Tras aplicar o algoritmo, créase un conxunto de datos reducido que se empregará para levar a cabo unha análise de sensibilidade dos modelos que se axustarán no Capítulo 4.

Agora ben, de forma exploratoria, cómpre estudar o impacto de prescindir de ditas observacións. Como se detectan os *outliers* mediante as variables dos prezos por quilómetro, as diferenzas máis salientables veranse ao representar a evolución temporal das mesmas. Recórdese, en particular, que ao estudar estas variables sobre os datos orixinais, o clúster 3 presentaba picos elevados dos prezos medios por quilómetro. Daquela, é de esperar que os clientes de alimentación e industria presenten as diferenzas máis marcadas.

Na Figura 3.17, represéntase a evolución temporal dos prezos medios por día para cada clúster tanto para os clientes (Figura (3.17a)), coma provedores (Figura (3.17b)). Comparando estas gráficas coas orixinais (Figuras 3.9 e 3.11), obsérvase unha atenuación dos picos que se detectaban no clúster 3, especialmente no caso dos provedores. Así, os prezos no conxunto de datos depurados presentan unha dispersión moito menor, tendo un comportamento máis estable.

Por outra banda, existe o risco de que os prezos elevados, marcados como atípicos polo algoritmo, sexan en realidade consecuencia natural do incremento da demanda polo Black Friday. Así, estaríase cometendo un erro ao diminuír o volume de pedidos durante a *peak season*, tendo isto impacto negativo sobre o obxectivo do proxecto.

Para descartar este risco, a Figura 3.18 representa a evolución temporal do volume de pedidos mensual por clúster, comparando a serie orixinal (liña continua) coa serie resultante trala eliminación

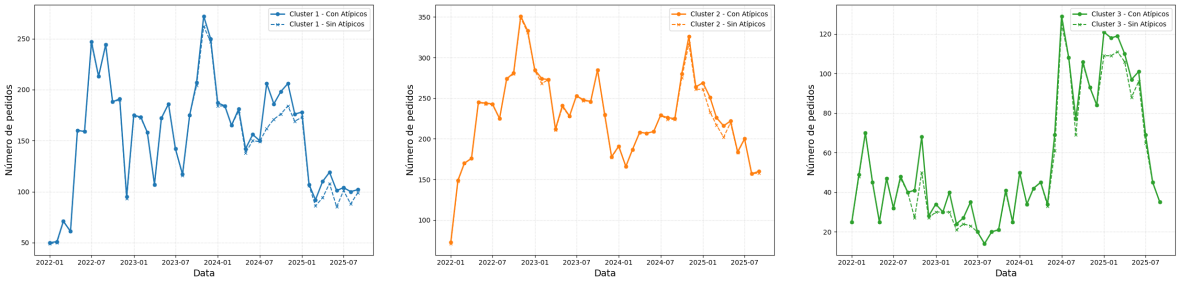


(a) Clientes

(b) Provedores

Figura 3.17: Impacto dos atípicos no prezo medio diario suavizado dos clientes e provedores segmentado por clúster (2022-2025). O clúster 3 (verde) presenta as maiores diferenzas con respecto á serie orixinal.

dos *outliers* (liña discontinua). Esta comparación resulta fundamental para asegurar que ao eliminar os atípicos non se prescinde fundamentalmente dos picos provocados polo Black Friday.



(a) Clúster 1: Pedidos

(b) Clúster 2: Pedidos

(c) Clúster 3: Pedidos

Figura 3.18: Impacto da eliminación de atípicos no volume de pedidos mensuais por clúster (2022-2025). O maior impacto obsérvase na serie do clúster 3 (verde), correspondente aos sectores de alimentación e industria.

Para o clúster 1 (Figura 3.18a), obsérvanse as maiores diferenzas a partires de 2024, onde principalmente cae o volume de envíos no último trimestre, así como durante os meses do 2025. Non obstante, a tendencia decrecente así como os ciclos estacionais mantéñense representados.

Pola súa parte, para os clientes do clúster 2 (Figura 3.18b) apenas se detectan cambios na demanda de pedidos. Isto concorda coa análise elaborada ao longo do presente capítulo, onde xa se detallou que os clientes de froita e paquetería presentaban unha maior estabilidade de mercado, con prezos que flutuaban en menor medida.

Por último, para o clúster 3 (Figura 3.18c) apréciase unha diminución máis considerable dos envíos, sendo notable en períodos de alta volatilidade nos prezos (finais do 2022). Como xa se explicou na análise previa, os clientes de dito clúster presentaban os picos máis marcados nos prezos por quilómetro.

Como conclusión deste capítulo, pódese afirmar que a análise exploratoria desenvolta permitiu acadar un coñecemento profundo da natureza e complexidade do problema obxecto de estudo. O carácter multidimensional dos datos, a identificación de patróns de comportamento diferenciados por clústeres e o estudo temporal das variables de interese constitúen a base necesaria para afrontar con garantías metodolóxicas os obxectivos propostos pola empresa. No vindeiro capítulo, procederase á especificación, axuste e avaliación de diversos modelos de series temporais, cuxo rendemento comparativo

permitirá determinar a estratexia óptima para a predición do volume de envíos.

Capítulo 4

Implementación de modelos de series temporais

Tras a caracterización exhaustiva dos datos e a consolidación do marco metodolóxico nos capítulos precedentes, este capítulo aborda a fase de implementación e validación dos modelos estocásticos detallados na Parte I. O obxectivo fundamental radica en identificar a metodoloxía preditiva que mellor capture a estrutura temporal da demanda, proporcionando a Trucksters unha ferramenta robusta para a anticipación das súas necesidades loxísticas.

En coherencia coa estratexia de segmentación definida no capítulo anterior, a modelización estrutúrase en dúas dimensións complementarias. En primeiro lugar, desenvólvese un estudo global sobre o volume total de pedidos para establecer unha liña base de rendemento. Nunha segunda instancia, o estudo desagrégase por clústeres co propósito de capturar as dinámicas específicas de cada segmento de clientes. Esta dualidade permite avaliar se o incremento na complexidade da análise non só aporta unha maior interpretabilidade desde a óptica de negocio, senón que tamén se traduce nunha mellora significativa da precisión estatística.

No que respecta ás técnicas de estimación, a Sección 4.2 dedícase ao axuste da metodoloxía de Box-Jenkins. Complementariamente, na Sección 4.3 explórase o potencial do algoritmo *Prophet*, baseado nun modelo aditivo xeneralizado que facilita a xestión de estacionalidades complexas. Finalmente, na Sección 4.4 realízase un estudo de robustez mediante a comparación do rendemento dos modelos sobre o conxunto de datos orixinal fronte á serie depurada de atípicos. Este procedemento resulta esencial para cuantificar o impacto das anomalías na estabilidade das previsións e determinar a configuración óptima para a empresa.

Se o lector quere afondar no código empregado durante o presente capítulo, pode consultar Airas (2026).

4.1. Deseño experimental

Antes de proceder ao axuste dos modelos, resulta imperativo definir o marco experimental que garante a comparabilidade e a validez estatística dos resultados. Esta sección detalla o protocolo seguido para avaliar cal é a mellor arquitectura de previsión para a demanda de Trucksters.

Obxectivos e xustificación dos modelos

Tal e como se estableceu no primeiro capítulo do traballo (Sección 1.3), o obxectivo central do experimento é determinar o modelo de previsión con maior capacidade preditiva. Para tal fin, a selección de métodos fundamentouse na complementariedade das súas aproximacións:

- **ARIMA:** Seleccionouse como representante da metodoloxía clásica de Box-Jenkins. É o estándar para modelar series temporais baseándose na súa propia estrutura de dependencia lineal (autocorrelación), resultando ideal para series cun comportamento estocástico estable.
- ***Prophet*:** Escolleuse pola súa flexibilidade para xestionar series temporais de negocio que presentan estacionalidades múltiples e efectos de calendario complexos. Ao basearse nun modelo aditivo de descomposición, permite integrar de forma directa coñecemento experto sobre festividades e cambios de tendencia.

Definición dos conxuntos de adestramento e test

Dada a natureza temporal dos datos, optouse por dividir a serie histórica nunha mostra de adestramento e outra de test:

- **Conxunto de Adestramento (01/01/2022 - 31/12/2024).** Empregarase para a identificación e axuste dos modelos (157 semanas).
- **Conxunto de Test (01/01/2025 - 30/09/2025).** As observacións do ano 2025 resérvanse exclusivamente para a avaliación dos modelos (38 semanas).

En particular, para a avaliación no conxunto de test, empregarase unha metodoloxía de ***Rolling Window Forecasting***. Esta técnica fudaméntase en readestrar e reavaliar iterativamente o modelo semana a semana, sendo unha aproximación máis adecuada ca predicir directamente para todo o horizonte de 2025. Isto permite á empresa simular un escenario operativo real onde as previsións se actualizan a medida que pechan novos datos.

Métricas de avaliación e hipóteses

A selección do "mellor" modelo basearase nas métricas do erro definidas na Sección 2.4.

- **MAE e RMSE:** Permiten cuantificar a magnitude do erro en unidades reais de pedidos. En particular, o RMSE penaliza con maior severidade as desviacións de gran magnitude.
- **MAPE:** Permite interpretar o erro en termos porcentuais.
- **MASE:** Para comparar o rendemento do modelo fronte a un preditor "inxenuo" (naïve), sendo unha métrica fundamental en series con estacionalidade.

As hipóteses fundamentais que se pretenden validar son:

1. Os modelos segmentados por clúster superan en precisión ao modelo global ao capturar dinámicas específicas dos sectores.
2. A inclusión de variables exóxenas (festivos e atípicos) incrementa a robustez do modelo ante picos de demanda como Black Friday.

Estratexia de modelización nos modelos ARIMA

Para cada serie temporal, seguirase o esquema iterativo explicado na Sección 2.3.2: selección e axuste do modelo, diagnose ou análise dos residuos e predición.

Porén, cómpre recordar que, un elemento que xogaba un papel fundamental neste proxecto eran os festivos ou eventos especiais. En particular, durante a análise exploratoria realizada no capítulo previo,

detallouse o efecto do Black Friday sobre as distintas series temporais. Daquela, co fin de capturar este efecto mediante os modelos ARIMA, previamente a seleccionar o posible modelo xerador da serie, levarase a cabo unha análise de intervención, construíndo variables *dummy* para modelar o efecto do Black Friday.

Estratexia de modelización nos modelos *Prophet*

Aínda que os modelos *Prophet* e ARIMA parten de aproximacións matemáticas diferentes, a análise realizada pola metodoloxía de Box-Jenkins proporciona información estrutural clave sobre as series temporais de Trucksters. Co obxectivo de garantir a coherencia entre as distintas etapas do proxecto e maximizar a precisión preditiva, aproveitarase o coñecemento aportado polos modelos ARIMA á hora de aplicar a metodoloxía *Prophet*.

Tratamento de atípicos

Por unha banda, *Prophet* é capaz de manexalos atípicos presentes na mostra de adestramento, mais só no caso de ser axustados como cambios na tendencia. De non ser así, a mellor opción para tratar a presenza de *outliers* sería eliminándoos manualmente, posto que o algoritmo non presenta problemas coa presenza de datos faltantes. Porén, para non prescindir desta información, optouse por incorporar os atípicos detectados na fase ARIMA como regresores externos.

Deste xeito, introdúcense variables binarias (*dummy*) para as datas sinaladas (O_t). Isto permite ao modelo estimar un coeficiente específico para o impacto de cada anomalía, illando o seu efecto e evitando que distorsione a estimación das compoñentes de tendencia e estacionalidade.

Tratamento de festividades e efectos de calendario

A natureza da actividade loxística de Trucksters implica unha forte dependencia do calendario laboral, onde certas festividades teñen impacto sobre a demanda de pedidos. *Prophet* incorpora unha compoñente ($h(t)$) para modelar estes eventos, mediante a que se capturará o efecto dos festivos considerados (Apéndice B). Porén, debido á condición semanal dos datos, a data da festividade debe moverse ao día da semana empregado como representante da mesma, neste caso luns. Ademais, cómpre sinalar que, por ter datos semanais, o efecto de moitos festivos será capturado pola compoñente estacional do modelo. Daquela, tan só será preciso engadir aqueles festivos que suceden en diferentes semanas ao longo da serie.

Aínda que *Prophet* dispón desta compoñente específica para captar o impacto de festividades ou eventos especiais, a análise exploratoria (Sección 3.3) revelou unha heteroxeneidade significativa no impacto da *peak season* entre os distintos clústeres. Mentres que certos sectores presentan un comportamento anticipado ou de corrección posterior, outros amósanse insensibles ao evento.

En consecuencia, en lugar de empregar unha configuración xenérica de festividade, decidiuse modelar o Black Friday mediante as variables exóxenas resultantes da análise de intervención realizada nos modelos ARIMA. Esta aproximación outorga unha maior flexibilidade, permitindo activar ou desactivar o regresor segundo a sigificación estatística observada en cada clúster, garantindo así un axuste adaptado á realidade operativa de cada segmento de clientes.

Selección de hiperparámetros e validación cruzada

A diferenza dos modelos ARIMA, onde a elección da orde (p, d, q) se basea principalmente na análise de correlogramas e criterios de información (AIC), en *Prophet* a complexidade do modelo regúlase a través de hiperparámetros que controlan a flexibilidade da tendencia e da estacionalidade.

Para garantir a capacidade de xeneralización do modelo e evitar o sobreaxuste (*overfitting*), implementouse un procedemento de busca en grella (*Grid Search*) combinado cunha estratexia de validación

cruzada sobre unha orixe deslizante (*rolling origin cross-validation*).

En concreto, defínese un espazo de busca centrado en tres principais parámetros:

- Un parámetro para controlar a sensibilidade ante o cambio de tendencia nos datos (`changepoint_prior_scale`). Valores elevados permiten unha tendencia máis flexible (risco de sobreaxuste) mentres que valores baixos forzan unha tendencia máis ríxida.
- Outro parámetro encargado de regular a magnitude das flutuacións estacionais permitidas (`seasonality_prior_scale`). De forma similar, valores elevados permiten axustar grandes oscilacións nos ciclos periódicos.
- Un último parámetro encargado de controlar o impacto das festividades incluídas (`holidays_prior_scale`) e cunha interpretación similar aos anteriores.

Adicionalmente, definiuse a estrutura da estacionalidade (`seasonality_mode`) de forma determinista para cada clúster, aproveitando a diagnose de varianza que se realizará na análise ARIMA. Así, para aqueles clústeres onde se detecte heterocedasticidade e se precise transformación logarítmica, fixouse unha estacionalidade de tipo multiplicativo. Pola contra, nos casos de varianza estable, mantívose a estacionalidade aditiva. Esta transferencia de información estrutural entre modelos permite reducir o espazo de busca sen comprometer a precisión teórica.

Unha vez definida a grella, o rendemento de cada combinación de parámetros avaliouse mediante validación cruzada temporal. Este método consiste en realizar unha serie de cortes (*cutoffs*) na historia da serie, adestrando o modelo cos datos anteriores ao corte realizando predicións sobre un horizonte temporal fixo.

Debido ás características dos datos dispoñibles neste traballo, o esquema de validación configurouse do seguinte xeito: estableceuse un período inicial de adestramento suficiente para captar a estacionalidade anual, seguido de cortes trimestrais e proxectando un horizonte de predición dun mes. Finalmente, seleccionouse aquela combinación de hiperparámetros que minimizou o RMSE promediado sobre todos os cortes realizados.

Por último, é preciso sinalar que se fixou un nivel de significación do 5 % para todos os contrastes que se levarán a cabo.

4.2. Modelos ARIMA

Nesta sección abordarase a implementación práctica da metodoloxía clásica de Box-Jenkins, construíndo modelos ARIMA capaces de capturar a estrutura de dependencia lineal e posibles patróns estacionais da demanda do transporte.

Seguindo a estrutura definida no deseño experimental, comezarase polo caso da serie global e posteriormente a axustarase un modelo para cada un dos clústeres construídos.

4.2.1. Modelo Global

Análise de intervención

O primeiro paso consiste en caracterizar o impacto do Black Friday na demanda agregada. A Figura 4.1 mostra a evolución temporal da serie global, onde se marcaron en vermello as semanas centrais do evento para os anos 2022, 2023 e 2024.

A inspección visual revela que o efecto do evento é de tipo transitorio. Porén, non se observa efecto só na propia semana do Black Friday, tamén se aprecia impacto tanto nas semanas previas coma un efecto rebote nas posteriores. Dada a heteroxeneidade deste comportamento, non sería correcto modelalo cunha única variable *dummy* de ventá. No seu lugar, defínense tres regresores de intervención

independentes que se incluirán no modelo ARIMA como variables exóxenas: unha variable binaria activa nas dúas semanas previas, unha variable binaria activa na semana do Black Friday, e unha última variable *dummy* activa nas dúas semanas posteriores ao evento.

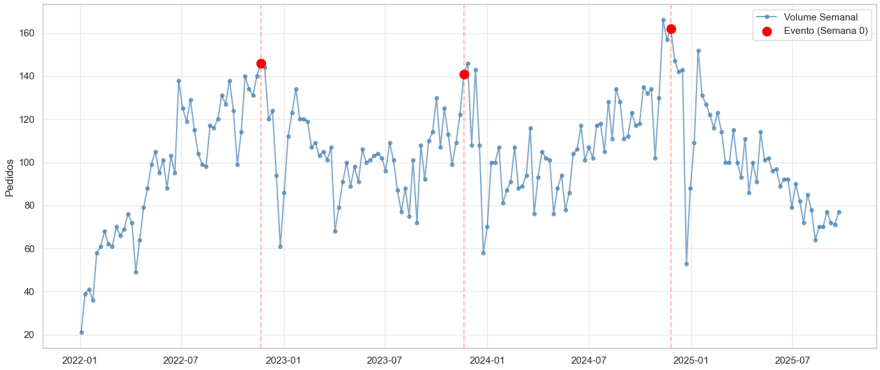
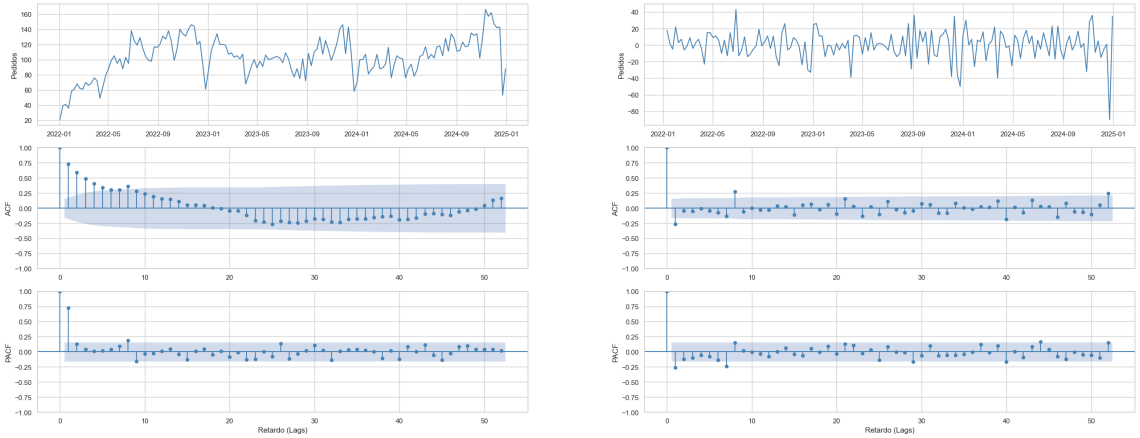


Figura 4.1: Análise de intervención na serie de demanda global: Impacto do Black Friday. As semanas centrais do evento márcanse en vermello.

Selección e axuste do modelo

Unha vez definidas as variables exóxenas para capturar o efecto do Black Friday, procédese á identificación da estrutura estocástica do modelo.



(a) Serie orixinal

(b) Serie diferenciada

Figura 4.2: Gráfico secuencial, *fas* e *fap* da serie de demanda global. Represéntase á esquerda a serie orixinal e á dereita a serie tras unha diferenza regular ($d = 1$). Non se observan problemas de heterocedasticidade.

A Figura 4.2 mostra a evolución temporal e os correlogramas (*fas* e *fap*) para a serie orixinal (Figura (4.2a)). Tanto no gráfico secuencial coma na lenta caída no *fas* apréciase un claro comportamento non estacionario (tendencia estocástica). En consecuencia, aplícase unha diferenza regular ($d = 1$). Nas gráficas da Figura (4.2b) obsérvase que a diferenciación logrou estabilizar a media da serie. A análise visual non suxire problemas de heterocedasticidade nin presenza de compoñente estacional. Centrando o estudo na estrutura de autocorrelación da serie diferenciada:

- A *fas* presenta un pico significativo no retardo 1, suxerindo un termo de media móbil ($q = 1$).

- A *fap* mostra da mesma forma un valor significativo no retardo 1, o que podería indicar alternativamente a presenza dun termo autorregresivo ($p = 1$)[1].

Isto permite propoñer como candidatos iniciais un ARIMA$(0, 1, 1)$ ou ben un ARIMA$(1, 1, 0)$.

Para a selección definitiva, compáranse os modelos suxeridos graficamente co modelo óptimo en termos do AIC, resultando neste caso un ARIMA$(1, 1, 1)$. Recórdese que, co fin de evitar o sobreaxuste, seleccionábase o modelo máis sinxelo (menos parámetros) que diste menos de dúas unidades ao AIC óptimo.

Modelo	**AIC**
$ARIMA(0, 1, 1)$	**1325.59**
$ARIMA(1, 1, 0)$	1334.93
$ARIMA(1, 1, 1)$	1325.08

Cadro 4.1: Comparativa dos valores AIC para os modelos candidatos na serie de demanda global. En negro destácase o modelo seleccionado en termos do AIC óptimo e tendo en conta o criterio de parsimonia.

Á vista do Cadro 4.1, obsérvase que a diferenza entre o modelo automático ARIMA$(1, 1, 1)$ e o modelo ARIMA$(0, 1, 1)$ é inferior a 2 unidades. Daquela seleccionarase como modelo candidato un ARIMA$(0, 1, 1)$.

Unha vez escollido o modelo, o seguinte paso será detectar a presenza de atípicos: innovativos (IO) e aditivos (AO). Cómpre recordar que na práctica ambos tipos de *outliers* modelaranse mediante variables *dummy*, que se incluirán no modelo como variables exóxenas (da mesma forma que o efecto do Black Friday). Así, implementando o algoritmo explicado na parte teórica (Sección 2.3.2), detectáronse dous atípicos: as semanas **25/12/2023 e 23/12/2024**.

Tras isto, axústase o modelo resultante: ARIMA$(0, 1, 1)$ con variables exóxenas (efecto do Black Friday e atípicos). Nesta etapa o procedemento será estimar por máxima verosimilitude os parámetros do modelo. Así, en caso de que algún dos parámetros non sexa significativo (nivel de significación do 5 %), suprimirase dito parámetro e reaxústase o modelo. Este proceso repítese ata que todos os elementos do modelo sexan estatisticamente significativos.

O Cadro 4.2 permite ver as estimacións, xunto co erro asociado á estimación e o p-valor correspondente ao contraste de significación dos mesmos. En particular, todos os parámetros considerados no modelo orixinal son estatisticamente significativos.

A análise dos coeficientes recollidos no Cadro 4.2 permite extraer conclusións fundamentais sobre a dinámica da demanda global de Trucksters:

- **Impacto do Black Friday**: Os tres regresores asociados ao evento presentan coeficientes positivos e altamente significativos. O impacto neto é máximo durante a semana central (BF_{peak}) cun incremento estimado de aproximadamente 29 pedidos sobre a tendencia base. É salientable que o efecto se anticipa nas dúas semanas previas ($BF_{pre} \approx 17$) e mantense, aínda que con menor intensidade, nas dúas posteriores ($BF_{pre} \approx 16$), o que valida a estratexia de modelar o evento como unha ventá de intervención ampliada e non como un punto illado.

[1] As bandas de confianza asumen un nivel de significación $\alpha = 0.05$. Baixo a hipótese nula de incorrelación, espérase estatisticamente que o 5 % das autocorrelacións mostrais queden fóra das bandas. Polo tanto, picos illados en retardos altos deben interpretarse con cautela, pois adoitan ser atribuíbles ao azar e non a un patrón estrutural real.

Parámetro	Estimación	Erro Estándar	Estatístico z	P-valor
$\mathrm{BF}_{\mathrm{pre}}$	17.010	5.987	2.841	0.004
$\mathrm{BF}_{\mathrm{peak}}$	28.962	11.723	2.470	0.013
$\mathrm{BF}_{\mathrm{post}}$	15.873	5.228	3.036	0.002
$\mathrm{O}_{23/12/24}$	-64.659	8.799	-7.349	< 0.001
$\mathrm{O}_{25/12/23}$	-42.758	9.636	-4.437	< 0.001
θ_1 (MA_1)	-0.533	0.084	-6.325	< 0.001
σ^2 (Varianza)	221.822	24.100	9.204	< 0.001

Cadro 4.2: Resumo da estimación dos parámetros para o modelo ARIMA$(0, 1, 1)$ con variables de intervención. Inclúense os coeficientes asociados ao impacto do Black Friday e eventos atípicos, todos eles significativos cun nivel de significación do 5 %.

- **Impacto dos atípicos**: Os *outliers* detectados coinciden coas semanas de Nadal de 2023 e 2024. Os coeficientes negativos e de gran magnitude (-42.76 e -64.66 respectivamente) cuantifican o drástico descenso da actividade loxística durante o peche industrial e comercial de fin de ano.
- **Compoñente estocástica (θ_1)**: O parámetro de media móbil (MA(1)) presenta un valor de -0.533. A significación deste termo indica que a serie ten memoria de curto prazo respecto aos erros de predición pasados, en particular o correspondente ao período inmediatamente anterior.

Diagnose do modelo

Unha vez estimados os parámetros, é imperativo validar as hipóteses estruturais subxacentes. O obxectivo é confirmar que os residuos ($\hat{a}_t$) se comportan como un proceso de ruído branco gaussiano (incorrelado, de media cero e distribución normal).

A Figura 4.3 presenta o panel de diagnose gráfica. Na gráfica secuencial (superior esquerda) non se aprecian problemas de heterocedasticidade nin incumprimento da hipótese de media 0. Mediante o correlograma (*fas*) descártase a presenza de autocorrelación, posto que ningún retardo excede as bandas de confianza significativas.

Respecto á normalidade, o histograma e o gráfico dos cuantís mostran un axuste razoable á distribución teórica, aínda que se aprecia unha lixeira asimetría nas colas. Para confirmar visualmente estas propiedades, empréganse os contrastes formais.

O Cadro 4.3 resume os resultados dos tests estatísticos. O contraste de Ljung-Box ($p = 0.530$) confirma a ausencia de correlación serial. No tocante á normalidade, obsérvase unha discrepancia marxinal entre o test de Shapiro-Wilk ($p = 0.091$, que acepta a normalidade) e o de Jarque-Bera ($p = 0.050$, no límite do rexeitamento). Dado o tamaño da mostra ($N \approx 150$), priorízase o resultado de Shapiro-Wilk por ser máis robusto en mostras finitas, concluíndo que a hipótese de normalidade é aceptable para os fins do modelo.

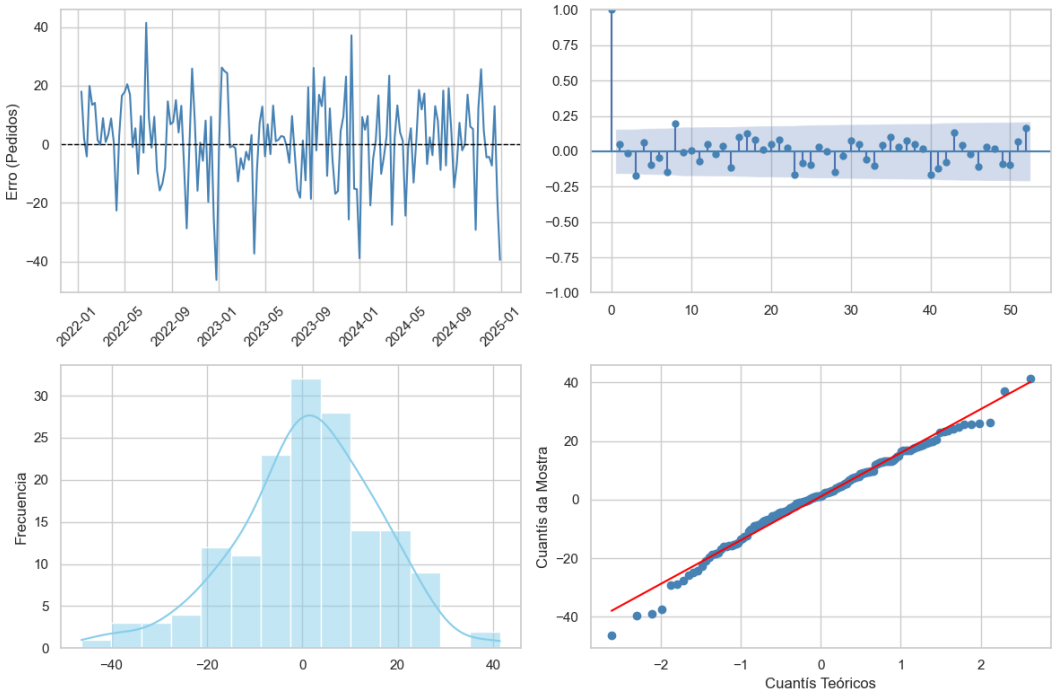


Figura 4.3: Diagnose de residuos para o modelo de demanda global: evolución temporal para o contraste de homocedasticidade (superior esquerda), correlograma *fas* para verificar a incorrelación (superior dereita), e histograma xunto co gráfico Q-Q para a validación da normalidade (inferiores).

Contraste	**P-valor**
Media 0 (t-test)	0.370
Incorrelación (Ljung-Box)	0.530
Homocedasticidade (test H)	0.440
Normalidade (Shapiro-Wilk)	0.091
Normalidade (Jarque-Bera)	0.050

Cadro 4.3: Resultados dos contrastes de hipóteses sobre os residuos do modelo. O modelo ARIMA$(0, 1, 1)$ resulta válido e con residuos gaussianos.

Predición

Finalmente, avalíase a capacidade preditiva do modelo nun escenario real. Seguindo a estratexia de *Rolling Window* definida no deseño experimental, emprégase o modelo ARIMA$(0, 1, 1)$ estimado con datos ata 2024 para proxectar a demanda semana a semana durante os datos do ano 2025 (horizonte $h = 1$). En cada paso, actualízase o conxunto de información co dato real observado para predicir o seguinte.

A Figura 4.4 compara a serie real (azul) coa predición do modelo (laranxa). O axuste visual é notablemente preciso: o modelo ARIMA, axudado polas variables de intervención, é capaz de capturar fielmente a tendencia decrecente estrutural que sofre a demanda ao longo do 2025. Isto indica que o modelo non só memorizou o pasado, senón que xeneraliza correctamente a dinámica do proceso.

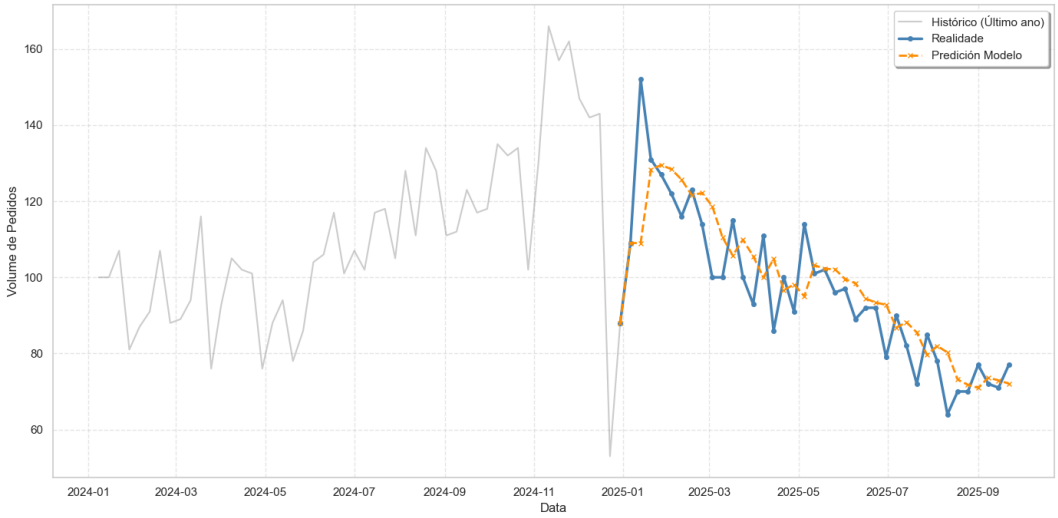


Figura 4.4: Gráfica comparativa da serie real (azul) na mostra de test (2025) fronte ás predicións do ARIMA$(0, 1, 1)$ realizadas mediante a técnica *rolling window* a horizonte $h = 1$ (laranxa).

Para cuantificar con precisión o rendemento do modelo na mostra de test (2025), calculáronse as métricas do erro definidas no marco teórico (Sección 2.4). O Cadro 4.4 resume os valores obtidos.

Métrica	**Valor**
MAE	7.90
$RMSE$	11.15
$MAPE$	8.18 %
$MASE$	0.60

Cadro 4.4: Avaliación da capacidade predictiva do modelo ARIMA$(0, 1, 1)$ para a serie global: métricas de erro calculadas sobre o conxunto de test (38 semanas). RMSE e MAE mídense en envíos semanais. MAPE é medida porcentual. MASE < 1 indica mellor rendemento ca o preditor naïve.

A interpretación destas medidas permite valorar a utilidade do modelo. Por unha parte, mediante o MAE e o RMSE pódese medir a precisión en volume. Por construción o RMSE penaliza fortemente os erros grandes, mais neste caso non dista moito do valor do MAE (aproximadamente 11 e 8). Tendo en conta o volume semanal de pedidos, desviacións de 8 envíos semanais supoñen unha gran aproximación para a empresa.

Por outra banda, o MAPE é unha medida porcentual. No sector do transporte, un erro do 8.18 % considérase satisfactorio.

Por último, o valor do MASE (0.60) constitúe o indicador máis relevante de calidade do axuste. Ao ser inferior a 1, indica que o modelo proposto supera ao preditor "inxenuo" (que consiste en considerar o último valor observado como predición). Especificamente, o modelo reduce nun 40 % o erro cometido polo preditor naïve de horizonte 1.

En conclusión, o modelo de Box-Jenkins axustado, incorporando a análise de intervención, demostra un rendemento preditivo robusto e constitúe unha ferramenta fiable para anticipar as necesidades loxísticas a nivel global.

Aínda que o modelo global proporciona unha estimación robusta da carga de traballo total, como xa se observou no Capítulo 3, a agregación dos datos pode enmascarar dinámicas diverxentes entre os distintos perfís de clientes. Por conseguinte, procédese ao axuste da metodoloxía de Box-Jenkins para cada un dos tres clústeres.

4.2.2. Modelo para o Clúster 1

O primeiro grupo estaba formado polos clientes dos sectores de transporte, *retail* e electrónica.

Análise de intervención

Iníciase o estudo caracterizando o impacto do Black Friday nos clientes dos sectores considerados. A Figura 4.5 representa a evolución temporal da demanda dos clientes do primeiro clúster, representándose en vermello as semanas do Black Friday.

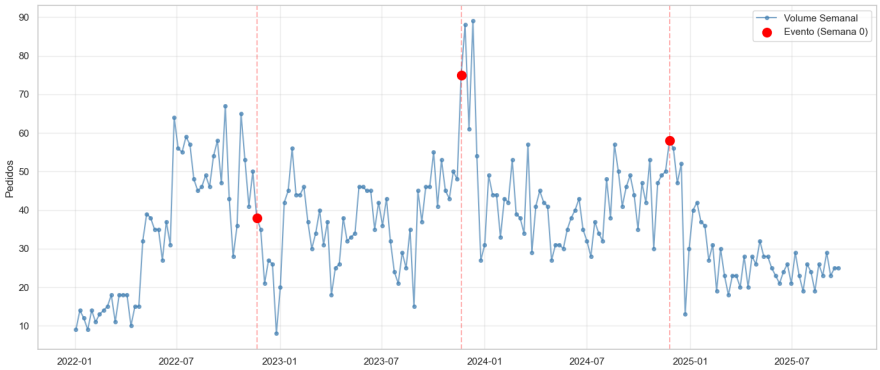


Figura 4.5: Análise de intervención na serie de demanda dos sectores de transporte, *retail* e electrónica: Impacto do Black Friday. As semanas centrais do evento márcanse en vermello.

Novamente o efecto do evento é de tipo transitorio (non afecta ao nivel medio da serie). Porén, a diferenza do caso xeral, neste clúster obsérvase unha maior volatilidade e heteroxeneidade interanual no impacto do evento.

Mentres que o 2024 se asemella ao patrón estándar (pico na semana central), o ano 2023 presenta un comportamento anómalo cun repunte da actividade nas semanas posteriores e no 2022 o impacto é apenas considerable.

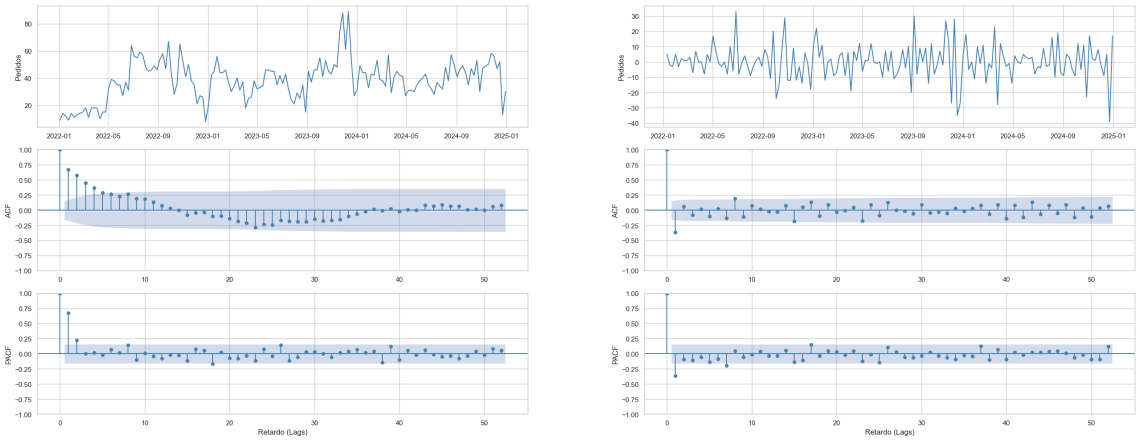
Ante a presente dispersión, e co obxectivo de capturar tanto a anticipación como o efecto rebote, defínese unha estrutura de intervención máis delimitada ca o caso global, composta por tres regresores puntuais: unha variable para a semana previa, outra para a semana central, e unha última variable para a observación posterior.

Esta configuración busca o equilibrio entre capturar o efecto da *peak season* e evitar o sobreaxuste debido ao ruído inherente á serie desagregada.

Selección e axuste do modelo

Tras construír os regresores externos que formarán parte do modelo, procédese á identificación da estrutura estocástica do mesmo.

A Figura 4.6 mostra a evolución temporal e os correlogramas para a serie orixinal (Figura (4.6a)). No gráfico secuencial e na caída lenta presente na *fas* obsérvase falta de estacionariedade pola presenza de tendencia. Para corrixilo, diferénciase regularmente a serie ($d = 1$) e observando as gráficas correspondentes á serie diferenciada (Figura (4.6b)), non se aprecian problemas na estacionariedade da serie (compoñente estacional ou heterocedasticidade). Centrando o estudo na estrutura de correlación da serie diferenciada, tanto o *fas* coma o *fap* mostran picos para o primeiro retardo, o que leva a propoñer como modelos candidatos para a serie orixinal un ARIMA$(0, 1, 1)$ ou ben un ARIMA$(1, 1, 0)$.



(a) Serie orixinal

(b) Serie diferenciada

Figura 4.6: Gráfico secuencial, *fas* e *fap* da serie de demanda do clúster 1. Represéntase á esquerda a serie orixinal e á dereita a serie tras unha diferenza regular ($d = 1$). Non se observan problemas de heterocedasticidade.

Tras esta primeira suxerencia gráfica, selecciónase o modelo óptimo en termos do criterio de información de Akaike (AIC), resultando un ARIMA$(1, 1, 2)$. Comparando os distintos modelos propostos en termos do AIC (Cadro 4.5) o ARIMA$(0, 1, 1)$ resulta o modelo seleccionado por ser máis sinxelo e distar en menos de dúas unidades do criterio de información óptimo.

Modelo	**AIC**
$ARIMA(0, 1, 1)$	**1184.64**
$ARIMA(1, 1, 0)$	1190.18
$ARIMA(1, 1, 2)$	1183.63

Cadro 4.5: Comparativa dos valores AIC para os modelos candidatos na serie de demanda do clúster 1. En negro destácase o modelo seleccionado en termos do AIC óptimo e tendo en conta o criterio de parsimonia.

A seguinte etapa consiste na detección dos atípicos (tanto aditivos coma innovativos). Ao aplicar o

algoritmo correspondente ao modelo seleccionado non se atopa ningún *outlier*, polo que non será preciso engadir máis regresores ao modelo resultante.

En consecuencia, axústase o ARIMA$(0,1,1)$ con tres variables exóxenas. Estímanse os parámetros do modelo por máxima verosimilitude e, de forma recursiva, elimínanse aqueles coeficientes que non resulten significativos. Así, o modelo final é:

Parámetro	**Estimación**	**Erro Estándar**	**Estatístico z**	**P-valor**
BF_{peak}	10.910	3.904	2.794	0.005
BF_{post}	14.612	4.420	3.306	0.001
θ_1 (MA_1)	-0.480	0.068	-7.010	< 0.001
σ^2 (Varianza)	110.521	9.272	11.920	< 0.001

Cadro 4.6: Resumo da estimación dos parámetros para o modelo ARIMA$(0,1,1)$ con variables de intervención. Inclúense os coeficientes asociados ao impacto do Black Friday todos eles significativos cun nivel de significación do 5 %. Non se detectaron atípicos.

A análise dos resultados recollidos no Cadro 4.6 permite debullar a dinámica específica da demanda para os sectores de transporte, *retail* e electrónica:

- **Efecto diferido do Black Friday**: Un dos achados máis notables deste modelo é a significación e magnitude dos regresores de intervención. Mentres que no modelo global o impacto era máximo na semana central, no clúster 1 obsérvase que o efecto rebote ou posterior ($\text{BF}_{\text{post}} \approx 14.6$) supera o impacto da propia semana do evento ($\text{BF}_{\text{peak}} \approx 10.9$). Este fenómeno valida a observación visual previa sobre o comportamento anómalo de 2023 e suxire que, para estes sectores, a presión loxística non se esgota na semana do evento, senón que se estende ás semanas seguintes, posiblemente debido á xestión de entregas pendentes ou a unha demanda máis dilatada no tempo.
- **Ausencia de anticipación e atípicos:** A variable correspondente ao efecto da semana previa ao Black Friday (BF_{pre}) non resultou ter un efecto significativo. Ademais, a nula detección de atípicos, suxire que as flutuacións da serie quedan explicadas pola propia estrutura estocástica e os regresores do Black Friday.

Diagnose do modelo

Unha vez estimados os parámetros, contrástase a validez do modelo axustado.

En primeiro lugar, lévase a cabo un diagnóstico gráfico mediante a Figura 4.7. Na representación secuencial dos residuos (superior esquerda) non se presencia incrumpimento nas hipótese de homocedasticidade nin de media cero. Grazas ao correlograma (*fas*) descártase a presenza de autocorrelación nos residuos, ao non superarse para ningún retardo as bandas de confianza.

Sobre a hipótese de normalidade, o histograma e o gráfico dos cuantís amosan certa asimetría nas colas, o que leva a dubidar sobre o cumprimento de dita hipótese.

Para confirmar a interpretación visual, cómpre aplicar os contrastes correspondentes. No Cadro 4.7 resúmense os resultados dos tests estatísticos. Os contrastes de Ljung-Box (p-valor $= 0.64$) e de

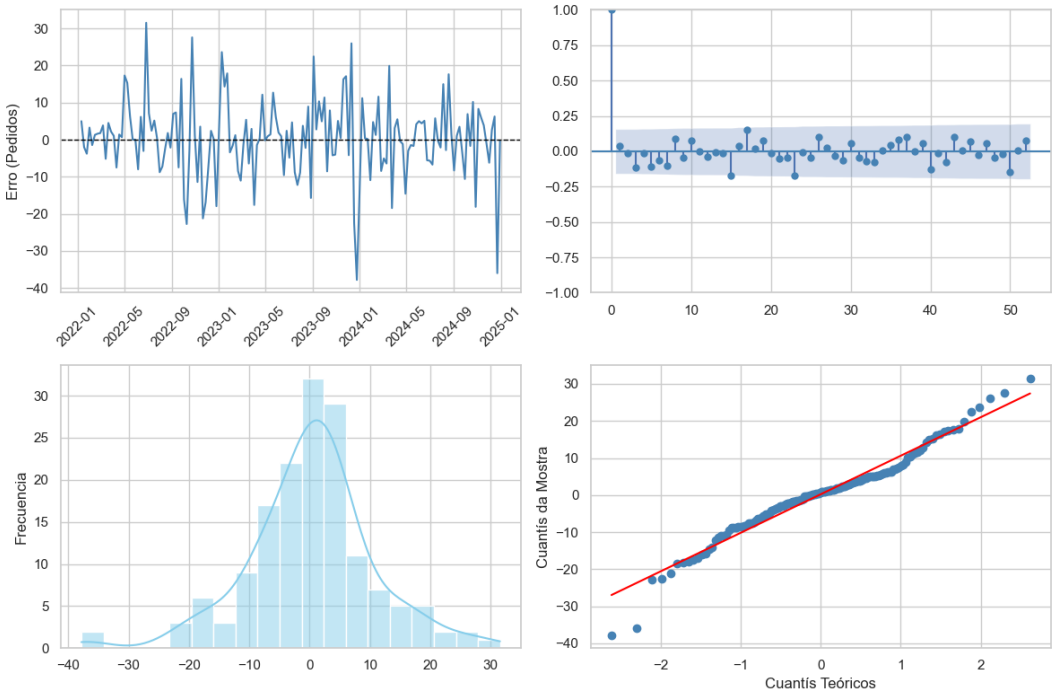


Figura 4.7: Diagnose de residuos para o modelo de demanda do clúster 1: evolución temporal para o contraste de homocedasticidade (superior esquerda), correlograma *fas* para verificar a incorrelación (superior dereita), e histograma xunto co gráfico Q-Q para a validación da normalidade (inferiores).

homocedasticidade (p-valor = 0.30) descartan a presenza de autocorrelación e heterocedasticidade. Tampouco se rexeita a hipótese de media cero (p-valor = 0.781). Porén, tal e como se intuía nas gráficas, os residuos non seguen unha distribución normal. En particular, rexéitase dita hipótese tanto co contraste de Jarque-Bera (p-valor < 0.001) coma co test de Shapiro-Wilk (p-valor = 0.001). Non obstante, recórdese que dita hipótese non resultaba eliminatoria, isto é, o modelo así axustado é correcto, mais se se quixeran construír intervalos de predición sería preciso empregar técnicas Bootstrap.

Contraste	**P-valor**
Media 0 (t-test)	0.781
Incorrelación (Ljung-Box)	0.640
Homocedasticidade (test H)	0.300
Normalidade (Shapiro-Wilk)	0.001
Normalidade (Jarque-Bera)	< 0.001

Cadro 4.7: Resultados dos contrastes de hipóteses sobre os residuos do modelo. O modelo ARIMA$(0, 1, 1)$ resulta válido e con residuos non gaussianos.

Predición

Finalmente avalíase a capacidade preditiva do modelo. Recórdese que para construír as predicións aplicábase a técnica do *Rolling-Window*.

Mediante a Figura 4.8 compárase a serie real (azul) coa predición do modelo (laranxa). Visualmente o axuste asemella considerablemente preciso. O modelo ARIMA$(0, 1, 1)$ con dous regresores externos, é capaz de captar a caída nos tres primeiros meses e a posterior estabilización do mercado.

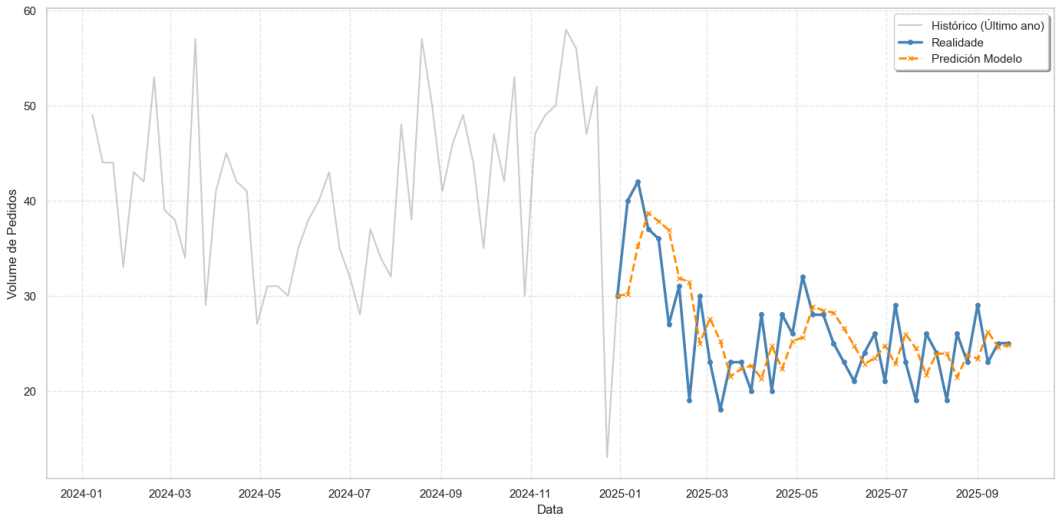


Figura 4.8: Gráfica comparativa da serie real (azul) na mostra de test (2025) fronte ás predicións do ARIMA$(0, 1, 1)$ realizadas mediante a técnica *rolling window* a horizonte $h = 1$ (laranxa).

Para cuantificar o rendemento do modelo na mostra de test (2025), no Cadro 4.8 resúmense os valores das métricas calculadas.

Métrica	**Valor**
MAE	3.88
$RMSE$	4.85
$MAPE$	15.57 %
$MASE$	0.45

Cadro 4.8: Avaliación da capacidade predictiva do modelo ARIMA$(0, 1, 1)$ para a serie do clúster 1: métricas de erro calculadas sobre o conxunto de test (38 semanas). RMSE e MAE mídense en envíos semanais. MAPE é medida porcentual. MASE < 1 indica mellor rendemento ca o preditor naïve.

Mediante o MAE e o RMSE obsérvase que o modelo comete erros en 4 pedidos semanais, o cal a nivel loxístico non supón un problema demasiado preocupante. Porén, o MAPE neste caso é lixeiramente superior ao obtido na serie global (15.57 %). Isto débese á escala da serie, estando agora entre 20 e 40 pedidos semanais, co cal erros en 4 unidades supoñen unha porporción máis considerable. Non

obstante, obténse un gran resultado a niveis do MASE (0.45), mellorándose de forma moi significativa a capacidade preditiva do modelo naïve.

En definitiva, os resultados obtidos para o segmento de transporte, *retail* e electrónica demostran un desempeño preditivo satisfactorio e de alto valor estratéxico para a empresa. Lógrase unha modelización precisa malia a volatilidade intrínseca deste clúster, caracterizado pola predominancia de contratos puntuais (*spot*) e a ausencia de acordos de volume a longo prazo con Trucksters. A capacidade do modelo para anticipar a tendencia e as flutuacións semanais nestas condicións valida a utilidade da segmentación previa.

4.2.3. Modelo para o Clúster 2

A continuación, abórdase a modelización dos clientes pertencentes aos sectores de paquetería e froita (Clúster 2).

Análise de intervención

Iníciase o axuste caracterizando o impacto do Black Friday. Como se adiantou na análise exploratoria (Capítulo 3), este clúster presenta a maior sensibilidade á *peak season*. A Figura 4.9 evidencia un claro efecto transitorio: nos exercicios de 2022 e 2024, o incremento da demanda comeza de forma visible dúas semanas antes do evento, seguido dunha corrección posterior.

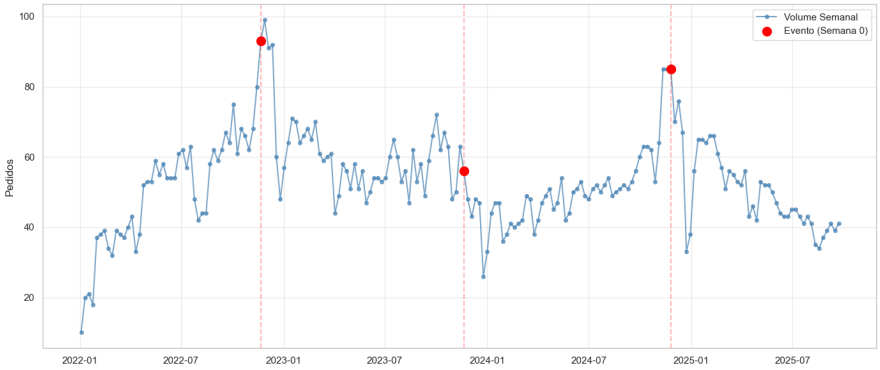


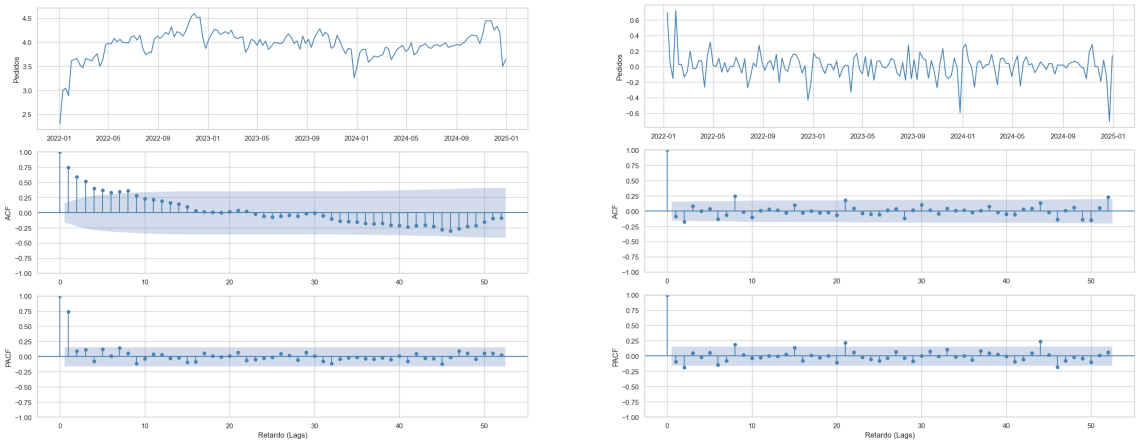
Figura 4.9: Análise de intervención na serie de demanda dos sectores de paquetería e froita: Impacto do Black Friday. As semanas centrais do evento márcanse en vermello.

Baseándose neste comportamento, defínense tres variables *dummy* de intervención:

1. Unha variable de anticipación activa nas dúas semanas previas $(t-2, t-1)$.
2. Unha variable impulso para a semana central (t).
3. Unha variable de corrección para a semana posterior $(t+1)$.

Selección e axuste do modelo

O seguinte paso é a identificación da estrutura estocástica. Unha análise preliminar sobre a serie bruta revelou problemas de heterocedasticidade nos residuos, o que invalidaba o suposto de varianza constante. Por conseguinte, procederase a aplicar unha transformación logarítmica (ec. 2.10, $\lambda = 0$) para estabilizar a varianza antes de continuar co proceso de identificación. Os detalles do estudo da serie orixinal pódense consultar no código (véxase Airas, 2026).



(a) Serie transformada

(b) Serie transformada diferenciada

Figura 4.10: Gráfico secuencial, *fas* e *fap* da serie de demanda do clúster 2 tras aplicar a transformación logarítmica. Represéntase á esquerda a serie orixinal e á dereita a serie tras unha diferenza regular ($d = 1$).

A Figura 4.10 representa o gráfico secuencial e os correlogramas para a serie transformada (Figura (4.10a)). Novamente obsérvase falta de estacionariedade debido a tendencia regular (cambio do nivel medio no gráfico secuencial e caída lenta na *fas*). Na serie diferenciada (Figura (4.10b)) corríxese a tendencia presente. Centrándose nos correspondentes correlogramas, tanto no *fas* coma no *fap* detéctanse picos significativos nos segundos retardos, o que leva a suxerir como posibles modelos candidatos para a serie transformada un ARIMA$(2, 1, 0)$ ou ben un ARIMA$(0, 1, 2)$.

Co obxectivo de identificar o modelo óptimo, calcúlase aquel con menor AIC. En particular, o modelo seleccionado polo algoritmo coincide co ARIMA$(2, 1, 0)$ proposto pola inspección visual (Cadro 4.9).

Modelo	**AIC**
$ARIMA(0, 1, 2)$	-118.64
$ARIMA(2, 1, 0)$	**-119.33**

Cadro 4.9: Comparativa dos valores AIC para os modelos candidatos na serie de demanda do clúster 2. En negro destácase o modelo seleccionado en termos do AIC óptimo e tendo en conta o criterio de parsimonia.

Para o modelo seleccionado aplícase o algoritmo de detección de atípicos. Neste caso captúranse tres *outliers* nas semanas: **31/01/2022**, **25/12/2023** e **23/12/2024**. O impacto destas observacións engádese ao modelo mediante variables *dummy* que se inclúen como regresores externos.

Tras identificar todas as compoñentes do modelo, axústase o correspondente ARIMA$(2, 1, 0)$ con seis variables exóxenas. Estímanse os coeficientes por máxima verosimilitude e elimínanse, recursivamente, aqueles parámetros non significativos.

No Cadro 4.10 recóllense as estimacións dos parámetros xunto co erro asociado ás mesmas, así como o p-valor correspondente ao contraste de significación. Á vista da análise de intervención previa, resulta

Parámetro	Estimación	Erro Estándar	Estatístico z	P-valor
$O_{31/01/22}$	0.384	0.036	10.650	< 0.001
$O_{23/12/24}$	-0.403	0.045	-9.016	< 0.001
$O_{25/12/23}$	-0.379	0.073	-5.215	< 0.001
ϕ_2 (AR_2)	-0.213	0.093	-2.277	0.023
σ^2 (Varianza)	0.018	0.002	10.148	< 0.001

Cadro 4.10: Resumo da estimación dos parámetros para o modelo ARIMA$(2, 1, 0)$ con variables de intervención. Inclúense os coeficientes asociados aos atípicos detectados, todos eles significativos cun nivel de significación do 5 %. Non se detectou efecto significativo do Black Friday.

destacable que ningún dos regresores que capturaban o efecto do Black Friday resultase significativo. Este feito choca coa interpretación que se fixo do impacto da *peak season* neste clúster. Porén, este fenómeno explícase pola interacción de dous factores derivados do axuste do modelo. Por unha banda, debido á natureza da transformación logarítmica, os picos do Black Friday comprímense, reducindo a diferenza co comportamento habitual da serie. Por outra banda, a compoñente autorregresiva de segunda orde (ϕ_2) captura eficazmente a inercia da serie. A dinámica de aceleración da demanda é absorbida polo modelo, facendo redundante a variable exóxena, xa que o propio histórico recente permite anticipar o cambio de nivel.

En conclusión, o modelo é capaz de modelar o impacto do Black Friday sen a necesidade de engadir regresores externos.

Outra particularidade estrutural deste modelo é a selección dunha compoñente autorregresiva de segunda orde (ϕ_2) significativa, mentres que o termo de primeira orde (ϕ_1) foi excluído. Estatisticamente, isto é coherente coa *fas* observada na Figura 4.10, onde se apreciaba un pico illado no retardo 2.

Diagnose do modelo

Unha vez estimados os parámetros, a seguinte etapa consiste en contrastar a validez do modelo así axustado.

Comézase cunha interpretación gráfica do comportamento dos residuos mediante a Figura 4.11. Á vista das gráficas, non se observa incumprimento da hipótese sobre a inesgadez dos residuos, nin problemas de heterocedasticidade (a serie foi transformada co obxectivo de corrixir isto). Tampouco se detectan problemas de autocorrelación nos residuos (*fas*).

En canto á hipótese de normalidade, obsérvase certa asimetría (histograma e gráfico de cuantís) o cal nos permite intuír a non gaussianidade dos residuos.

Co fin de confirmar a interpretación visual, no Cadro 4.11 resúmense os resultados dos distintos contrastes estatísticos.

Así, o diagnóstico dos residuos trala transformación logarítmica amosa un p-valor de 0.05 para o test de homocedasticidade. Se ben este resultado se atopa no límite de significación estatística, considérase aceptable dado o bo comportamento do modelo noutras métricas (contraste de Ljung-Box (p-valor =

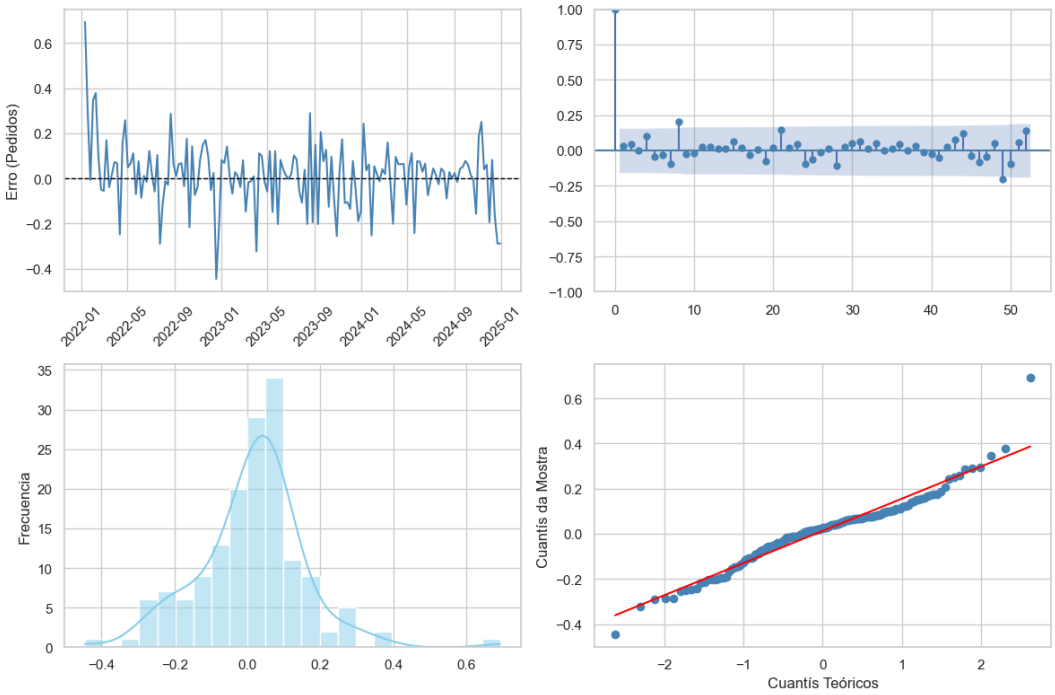


Figura 4.11: Diagnose de residuos para o modelo de demanda do clúster 2: evolución temporal para o contraste de homocedasticidade (superior esquerda), correlograma *fas* para verificar a incorrelación (superior dereita), e histograma xunto co gráfico Q-Q para a validación da normalidade (inferiores).

Contraste	**P-valor**
Media 0 (t-test)	0.276
Incorrelación (Ljung-Box)	0.700
Homocedasticidade (test H)	0.050
Normalidade (Shapiro-Wilks)	0.020
Normalidade (Jarque-Bera)	< 0.001

Cadro 4.11: Resultados dos contrastes de hipóteses sobre os residuos do modelo. O modelo ARIMA$(2, 1, 0)$ resulta válido e con residuos non gaussianos.

0.700) e contraste de media 0 (p-valor $= 0.276$)). Daquela asúmese a homocedasticidade, entendendo que o modelo captura adecuadamente a dinámica principal da serie temporal.

Por outra banda, tal e como se adiantaba mediante a interpretación gráfica dos residuos, rexéitase a hipótese de normalidade tanto co test de Jarque-Bera (p-valor < 0.001) coma co test de Shapiro-Wilks (p-valor $= 0.020$). Así e todo, o incumprimento desta hipótese non invalida o modelo axustado.

Predición

Tras a validación do modelo, avalíase o rendemento preditivo do mesmo. A Figura 4.12 representa unha comparación entre as predicións calculadas polo modelo ARIMA$(2,1,0)$ (liña laranxa) xunto cos valores da serie observada (liña azul). Á vista da gráfica, o axuste asemella captar con precisión o comportamento da serie durante a mostra de test. En particular, detecta a tendencia crecente inicial, así como a posterior caída do nivel medio.

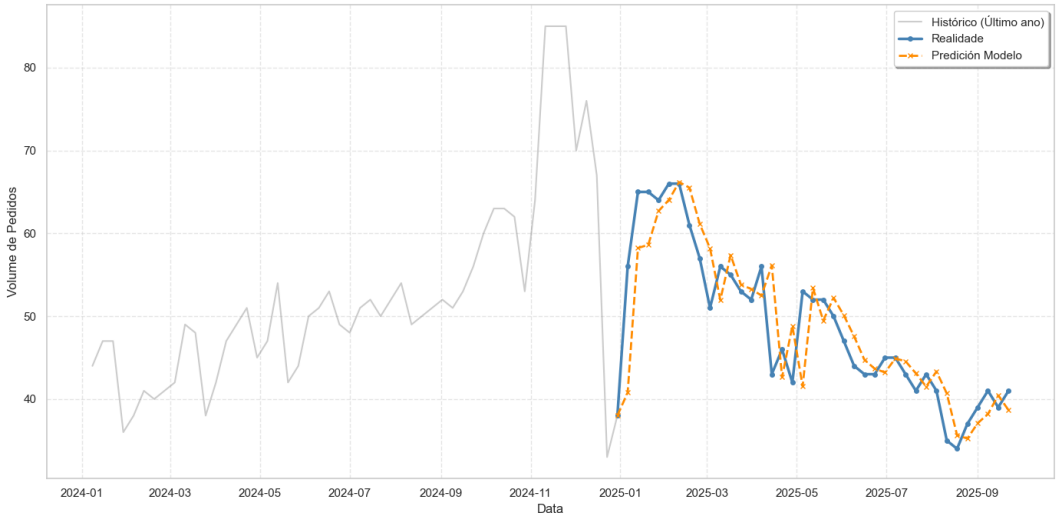


Figura 4.12: Gráfica comparativa da serie real (azul) na mostra de test (2025) fronte ás predicións do ARIMA$(2,1,0)$ realizadas mediante a técnica *rolling window* a horizonte $h = 1$ (laranxa).

Co obxectivo de xustificar o bo rendemento do modelo, o Cadro 4.12 recolle as métricas do erro das predicións calculadas. En particular, obsérvanse valores claramente baixos do MAE (3.60) e do RMSE (4.94), que, recórdese, medían a precisión volume. Tendo en conta ademais a demanda semanal de pedidos, erros medios próximos a 4 envíos semanais supoñen unha proporción pouco significativa, dando lugar a un MAPE $= 7.38\,\%$. Por último, é importante salientar a mellora preditiva que presenta o modelo axustado con respecto ao preditor naïve, sendo o MASE $= 0.63$.

Métrica	**Valor**
MAE	3.60
$RMSE$	4.94
$MAPE$	$7.38\,\%$
$MASE$	0.63

Cadro 4.12: Avaliación da capacidade predictiva do modelo ARIMA$(2,1,0)$ para a serie do clúster 2: métricas de erro calculadas sobre o conxunto de test (38 semanas). RMSE e MAE mídense en envíos semanais. MAPE é medida porcentual. MASE < 1 indica mellor rendemento ca o preditor naïve.

En conclusión, a metodoloxía Box-Jenkins ofrece resultados robustos para os sectores de froita e

paquetería. O modelo final ARIMA$(2,1,0)$ demostrou ser capaz de interiorizar a estacionalidade da *peak season* a través da súa propia estrutura dinámica e da transformación logarítmica, sen depender de regresores externos complexos. Isto simplifica a operativa de predición futura sen sacrificar precisión (MAPE $\approx 7\,\%$).

4.2.4. Modelo para o Clúster 3

Para rematar coa metodoloxía Box-Jenkins, repítese o proceso sobre os clientes que conforman o último clúster (alimentación e industria).

Análise de intervención

A análise exploratoria do Capítulo 3 amosou un efecto marxinal do Black Friday sobre o comportamento da serie temporal dos clientes de alimentación e industria. En efecto, a Figura 4.13 reforza a análise previa, destacando en vermello as semanas correspondentes co evento nos anos de mostra. Porén, co fin de captar posibles impactos que non se están a interpretar de forma visual, decidiuse construír 3 variables *dummy* seguindo a metodoloxía dos modelos anteriores. Así e todo, neste caso tan só se captura o efecto na semana previa e posterior (ademais do evento central).

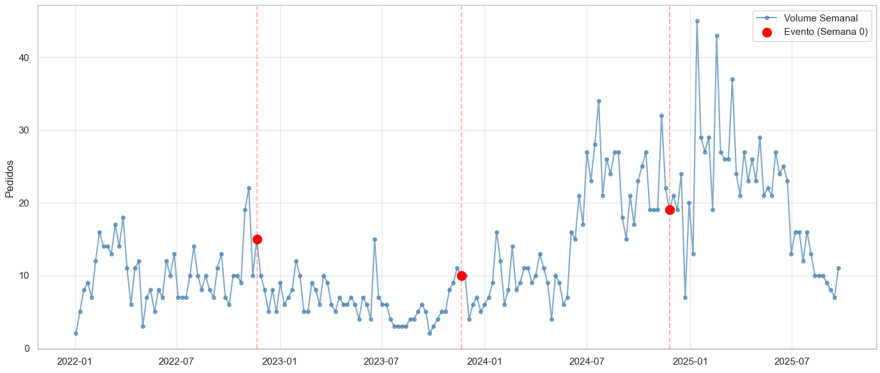


Figura 4.13: Análise de intervención na serie de demanda dos sectores de alimentación e industria: Impacto do Black Friday. As semanas centrais do evento márcanse en vermello.

Cómpre sinalar que esta práctica non supón un sobreaxuste posto que, en caso de confirmarse a ausencia de impacto da *peak season*, as variables regresoras corrrespondentes non resultarán significativas e serán eliminadas do modelo.

Selección e axuste do modelo

A seguinte etapa consiste na identificación da estrutura estocástica do modelo. A Figura 4.14 representa o gráfico secuencial e os correlogramas da serie orixinal. A inspección visual da evolución temporal revela unha certa estrutura heterocedástica. En particular, a volatilidade da serie non é constante, senón que presenta unha correlación positiva co nivel da media: os períodos de maior demanda (2024) presentan unha dispersión significativamente superior aos períodos de menor nivel.

Co obxectivo de estabilizar a varianza, aplicouse unha transformación logarítmica á serie (ec. 2.10, $\lambda = 0$). Repetindo a análise gráfica sobre a serie así transformada (Figura 4.15), detéctase falta de estacionariedade debido á tendencia observada no gráfico secuencial e na lenta caída da autocorrelación simple (Figura (4.15a)). Este feito corríxese aplicando unha diferenza regular ($d = 1$). Centrando o estudo nos gráficos da serie diferenciada resultante (Figura (4.15b)), non se observan problemas da estacionariedade.

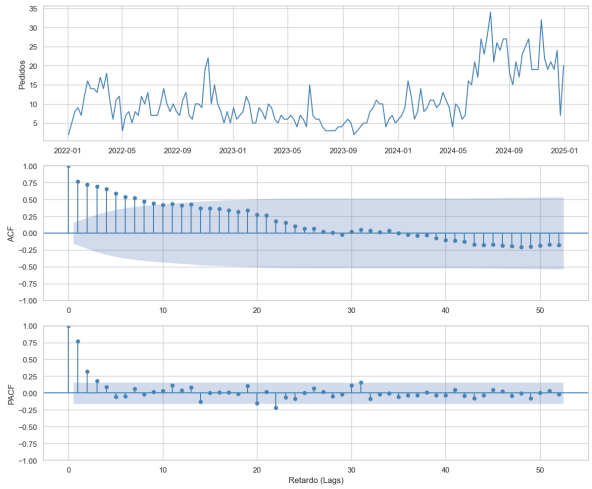
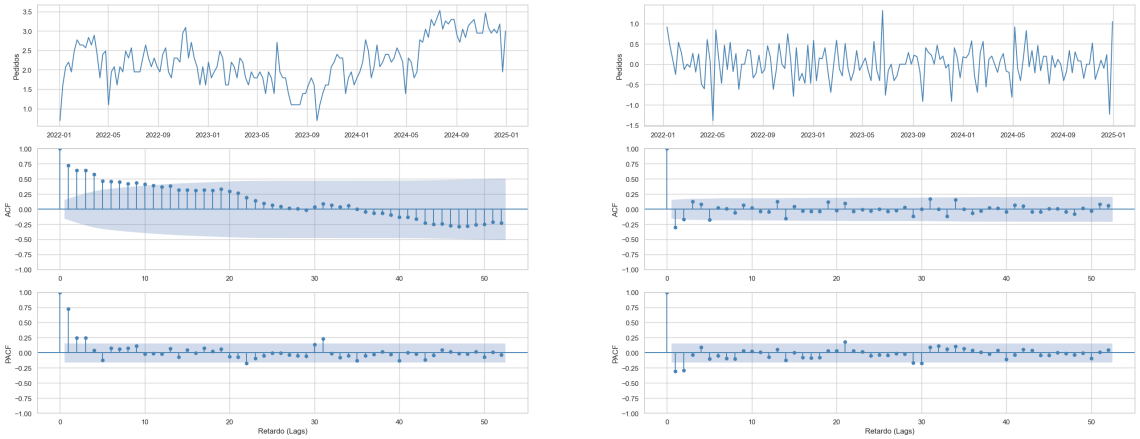


Figura 4.14: Gráfico secuencial, *fas* e *fap* da serie de demanda do clúster 3. Represéntase á esquerda a serie orixinal e á dereita a serie tras unha diferenza regular ($d = 1$). Obsérvanse problemas de heterocedasticidade.

En particular, atendendo aos autocorrelogramas, tanto no *fas* coma no *fap* existen picos significativos nos segundos retardos, o que leva a suxerir como posibles modelos para a serie transformada un ARIMA$(0, 1, 2)$ ou ben un ARIMA$(2, 1, 0)$.



(a) Serie transformada

(b) Serie transformada diferenciada

Figura 4.15: Gráfico secuencial, *fas* e *fap* da serie de demanda do clúster 3 tras aplicar a transformación logarítmica. Represéntase á esquerda a serie orixinal e á dereita a serie tras unha diferenza regular ($d = 1$).

Co obxectivo de identificar o modelo óptimo, calcúlase aquel que minimice o criterio de información de Akaike (AIC). O Cadro 4.13 recolle unha comparación dos valores do correspondente criterio para os modelos candidatos. En consecuencia, selecciónase o ARIMA$(2, 1, 0)$.

Para o modelo seleccionado, aplícase o algoritmo de detección de *outliers*, sen que se detectase, neste caso, ningunha observación atípica.

Polo tanto, a seguinte etapa será axustar o modelo identificado. Novamente, procédese á estimación

Modelo	AIC
$ARIMA(0,1,2)$	151.79
$ARIMA(2,1,0)$	**149.92**

Cadro 4.13: Comparativa dos valores AIC para os modelos candidatos na serie de demanda do clúster 3. En negro destácase o modelo seleccionado en termos do AIC óptimo e tendo en conta o criterio de parsimonia.

dos coeficientes por máxima verosimilitude e, de forma recursiva, elimínanse aqueles parámetros non significativos.

Parámetro	Estimación	Erro Estándar	Estatístico z	P-valor
ϕ_1 (AR_1)	-0.429	0.084	-5.139	< 0.001
ϕ_2 (AR_2)	-0.311	0.082	-3.807	< 0.001
σ^2 (Varianza)	0.137	0.013	10.404	< 0.001

Cadro 4.14: Resumo da estimación dos parámetros para o modelo ARIMA$(2,1,0)$. Non se detectou efecto significativo do Black Friday a un nivel de significación do 5 %. Tampouco se detectaron atípicos.

O Cadro 4.14 recolle as estimacións dos parámetros, o erro estándar de ditas estimacións e o p-valor do correspondente contraste de significación. En particular, a exclusión por falta de significación estatística de todos os regresores asociados á *peak season* confirma a hipótese formulada na análise exploratoria. A diferenza dos sectores orientados ao consumo final, a loxística de alimentación e industria responde a ciclos de produción e abastecemento estables, resultando inmune ás flutuacións inducidas por eventos do calendario comercial.

Por outra banda, a natureza negativa dos dous parámetros autorregresivos (ϕ_1 e ϕ_2) reflicte un comportamento oscilatorio con tendencia á autocorrección. No contexto loxístico de Trucksters, isto tradúcese nun mecanismo de compensación proporcional: un incremento porcentual na demanda durante unha semana determinada tende a ser seguido por axustes á baixa nas semanas posteriores para estabilizar o nivel medio.

Diagnose do modelo

A continuación, estúdase a validez do modelo axustado na etapa anterior. A Figura 4.16 permite facer un primeiro diagnóstico gráfico. Na primeira gráfica represéntase a evolución temporal dos residuos, onde non se observan problemas de heterocedasticidade nin incumprimento da hipótese de media cero dos residuos. O correlograma (*fas*) non mostra autocorrelacións significativas para ningún retardo, descartando problemas na hipótese de incorrelación. Pola contra, os gráficos inferiores (histograma e Q-Q plot) amosan a presenza de colas pesadas, o cal suxire a ausencia de normalidade nos residuos.

Co fin de reforzar a interpretación gráfica, o Cadro 4.15 recolle os resultados dos distintos contrastes aplicados sobre os residuos.

En particular, os contrastes de Ljung-Box (p-valor = 0.660) e de homocedasticidade (p-valor =

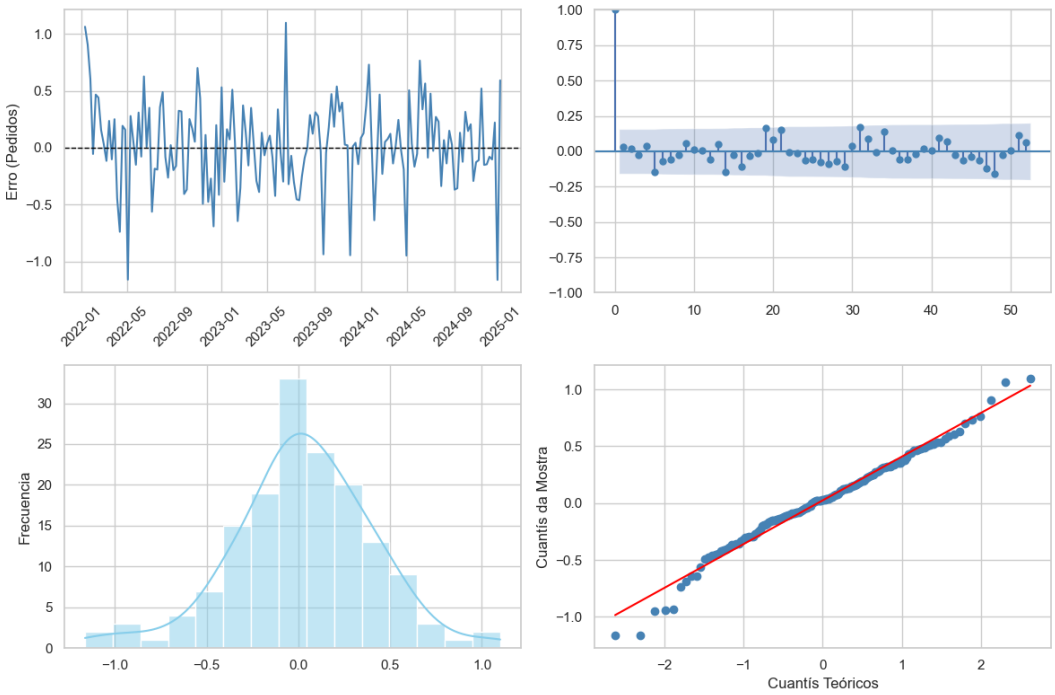


Figura 4.16: Diagnose de residuos para o modelo de demanda do clúster 3: evolución temporal para o contraste de homocedasticidade (superior esquerda), correlograma *fas* para verificar a incorrelación (superior dereita), e histograma xunto co gráfico Q-Q para a validación da normalidade (inferiores).

Contraste	**P-valor**
Media 0 (t-test)	0.431
Incorrelación (Ljung-Box)	0.660
Homocedasticidade (test H)	0.720
Normalidade (Shapiro-Wilk)	0.046
Normalidade (Jarque-Bera)	< 0.001

Cadro 4.15: Resultados dos contrastes de hipóteses sobre os residuos do modelo. O modelo ARIMA$(2, 1, 0)$ sen regresores externos resulta válido e con residuos non gaussianos.

0.720), xunto co test de media cero (p-valor $= 0.431$) validan as hipóteses fundamentais do modelo axustado.

Por outra banda, tal e como se intuía pola interpretación gráfica, confírmase a falta de normalidade, rexeitándose dita hipótese de forma contundente co test de Jarque-Bera (p-valor < 0.001), mentres que o test de Shapiro-Wilk (p-valor $= 0.046$) sitúase no límite da rexión de rexeitamento.

Predición

Unha vez confirmada a validez do modelo axustado, mídese o rendemento preditivo do mesmo. A Figura 4.17 compara a serie real (azul) coa predición do modelo ARIMA$(2, 1, 0)$ (laranxa).

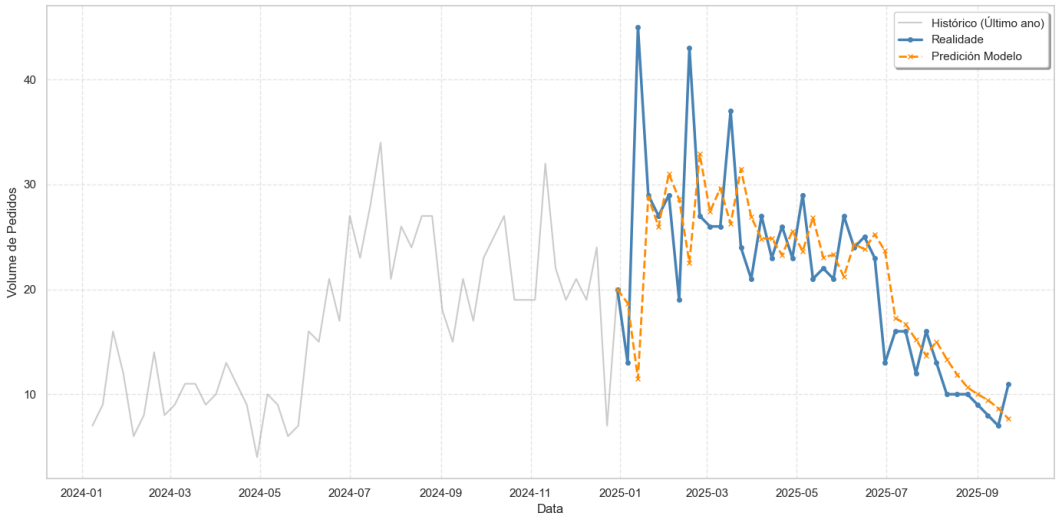


Figura 4.17: Gráfica comparativa da serie real (azul) na mostra de test (2025) fronte ás predicións do ARIMA$(2, 1, 0)$ realizadas mediante a técnica *rolling window* a horizonte $h = 1$ (laranxa).

Á vista da gráfica, obsérvase un rendemento considerablemente inferior aos acadados nos anteriores clústeres. Como se sinalou na análise exploratoria (Seccíon 3.3), o primeiro trimestre da mostra de test presenta unha alta volatilidade con picos de demanda considerables. O modelo, pola súa natureza, tende a suavizar estas oscilacións, subestimando a magnitude dos picos iniciais.

Este baixo rendemento que se amosa na gráfica confírmase nas métricas do erro (Cadro 4.16). Os valores do MAE (4.59) e do RMSE (7.65) son elevados en relación ao volume medio da serie, resultando nun MAPE de 20.73 %. Ademais, a métrica máis preocupante resulta o MASE (1.42), por acadar un valor superior a 1, o cal indica que o modelo se comporta peor ca o preditor naïve (de horizonte 1).

Métrica	**Valor**
MAE	4.59
$RMSE$	7.65
$MAPE$	20.73 %
$MASE$	1.42

Cadro 4.16: Avaliación da capacidade predictiva do modelo ARIMA$(2, 1, 0)$ para a serie do clúster 3: métricas de erro calculadas sobre o conxunto de test (38 semanas). RMSE e MAE mídense en envíos semanais. MAPE é medida porcentual. MASE > 1 indica peor rendemento ca o preditor naïve.

Así e todo, a inspección visual suxire que o erro se concentra nese primeiro trimestre de alta inestabilidade. Para avaliar a capacidade do modelo en condicións de mercado máis estables, recalculáronse

as métricas excluíndo este período transitorio (Cadro 4.17).

Métrica	Valor
MAE	2.68
$RMSE$	3.46
$MAPE$	17.21 %
$MASE$	0.83

Cadro 4.17: Avaliación da capacidade predictiva do modelo ARIMA$(2, 1, 0)$ para a serie do clúster 3, recortando o primeiro trimestre: métricas de erro calculadas sobre o conxunto de test (38 semanas). RMSE e MAE mídense en envíos semanais. MAPE é medida porcentual. MASE < 1 indica mellor rendemento ca o preditor naïve.

Daquela, obsérvase unha mellora drástica no rendemento. O MAE redúcese practicamente á metade (2.68), e, o máis relevante, o MASE descende ata 0.83, validando que o modelo achega melloras preditivas ao modelo "inxenuo" (naïve) cando a serie non presenta volatilidade extrema. Isto permite concluír que o modelo é útil para a planificación en escenarios estándar, aínda que presenta limitacións ante cambios bruscos de nivel a curto prazo.

A aplicación da metodoloxía Box-Jenkins aos tres clústeres de clientes permitiu asentar unha liña base robusta para a predición da demanda en Trucksters. A través da identificación, estimación e diagnose, extráense tres conclusións sobre a natureza das series tratadas:

- **Heteroxeneidade estrutural:** A análise confirmou que cada segmento de negocio posúe unha dinámica estocástica diferenciada. Mentres os clientes do clúster 1 puideron modelarse directamente, os outros sectores precisaron de transformacións de estabilización da varianza (logarítmicas) para satisfacer a hipótese de homocedasticidade.
- **Impacto do Black Friday:** Observáronse diferenzas no efecto da *peak season* sobre os distintos clústeres, confirmando a análise do Capítulo 3. En particular, para os sectores de transporte, *retail* e electrónica foi preciso introducir variables exóxenas para modelar o impacto do Black Friday. Pola súa parte, no cluster 2 (froita e paquetería), o modelo era quen de captar o efecto mediante as súas compoñentes estocásticas, mentres que para os clientes do último clúster non se apreciaba impacto significativo da *peak season*.
- **Limitacións ante volatilidade:** Se ben os modelos ARIMA demostraron unha gran capacidade de axuste en períodos estables, a predición do clúster 3 evidenciou as limitacións dos modelos clásicos para anticipar cambios bruscos de nivel ou picos de volatilidade non cíclicos.

Estes resultados, a pesar de ser moi positivos, suxiren a necesidade de explorar metodoloxías alternativas que poidan ofrecer unha maior flexibilidade no tratamento de tendencias non lineais e efectos de calendario complexos.

4.3. Modelos *Prophet*

Unha vez avaliada a capacidade preditiva dos modelos clásicos (ARIMA), neste apartado explórase unha aproximación baseada en Modelos Aditivos Xeneralizados (GAM): o modelo *Prophet*.

Recórdese que, ao contrario da metodoloxía de Box-Jenkins, a cal se centra en estudar a estrutura de autocorrelación dos datos para proxectar a serie cara ao futuro, *Prophet* formula o problema de predición como un exercicio de axuste de curvas. Daquela, axustaranse as distintas compoñentes do modelo explicadas na Sección 2.3.3 tanto para o caso global, coma para cada clúster por separado.

Tendo presentes consideracións previas explicadas na Sección 4.1, abórdase o axuste dos distintos modelos.

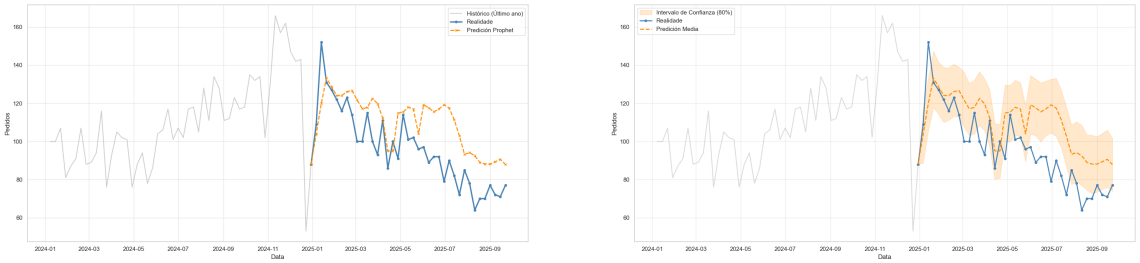
4.3.1. Modelo Global

Comézase estudando a serie agregada de todos os clientes. Retomando a análise previa, no axuste ARIMA resultaron significativas 5 variables exóxenas: 3 delas encargadas de captar o impacto do Black Friday, mentres que as restantes modelaban os atípicos detectados (25/12/2023 e 23/12/2024). Dado que non se detectaron problemas de heterocedasticidade na serie orixinal, configúrase o modelo *Prophet* considerando estacionalidade aditiva.

Procedeuse á selección automática de hiperparámetros, incluíndo no modelo tanto as variables *dummy* coma os festivos correspondentes. Tras aplicar este procedemento, o modelo óptimo obtivo un RMSE = 19.42 na fase de validación.

Unha vez axustado o modelo, realizáronse as predicións para o conxunto de test (2025) mediante a técnica de *Rolling Window* descrita no deseño experimental. A Figura 4.18 compara a serie real (azul) coa predición do modelo (laranxa). Na primeira representación (Figura (4.18a)), obsérvase un axuste visual aceptable: o modelo *Prophet* captura con precisión a tendencia e estacionalidade durante a primeira metade do ano. Porén, o algoritmo non é capaz de anticipar a caída estrutural da demanda rexistrada nos últimos meses da mostra.

Ademais, este algoritmo destaca pola súa capacidade para xerar intervalos de incerteza sen suposicións de normalidade estritas. Na Figura (4.18b) represéntanse ditos intervalos para un nivel de confianza do 80 %. Obsérvase como, na etapa final do ano, os datos reais caen sistematicamente por debaixo do límite inferior do intervalo, confirmando estatisticamente que o cambio de tendencia é unha anomalía que o modelo non pode explicar coa información histórica dispoñible.



(a) Predicións

(b) Intervalos de predición do 80 %

Figura 4.18: Validación mediante *Rolling Window* para a demanda global: comparación no conxunto de test (38 semanas) entre a serie real (azul) e as predicións do modelo *Prophet* (laranxa). Á dereita móstrase o detalle das predicións cos seus intervalos de confianza ao 80 %.

Co obxectivo de cuantificar o rendemento preditivo, calculáronse as métricas de erro recollidas no Cadro 4.18. O MAE e o RMSE presentan valores próximos (16.19 e 19.19 respectivamente), o que indica un erro medio absoluto próximo aos 17 envíos semanais. Pola súa parte, o MAPE sitúase nun 18.53 %, un resultado aceptable aínda que con marxe de mellora dado o nesgo final. Por último, obtense un MASE de 0.72. Este valor inferior á unidade indica que o modelo *Prophet* supera en rendemento ao

preditor inxenuo estacional (*Seasonal Naïve* con m=52), validando a súa utilidade fronte a simplemente replicar o comportamento do ano anterior.

Métrica	**Valor**
MAE	16.19
$RMSE$	19.19
$MAPE$	18.53 %
$MASE$	0.72

Cadro 4.18: Avaliación da capacidade predictiva do modelo *Prophet* para a serie de demanda global: métricas de erro calculadas sobre o conxunto de test (38 semanas). RMSE e MAE mídense en envíos semanais. MAPE é medida porcentual. MASE < 1 indica mellor rendemento ca o preditor naïve.

Outra vantaxe do modelo *Prophet* resulta a interpretabilidade das súas compoñentes. A Figura 4.19 desagrega a serie temporal nas súas partes constituíntes, permitindo unha interpretación detallada dos factores que impulsan a demanda global en Trucksters.

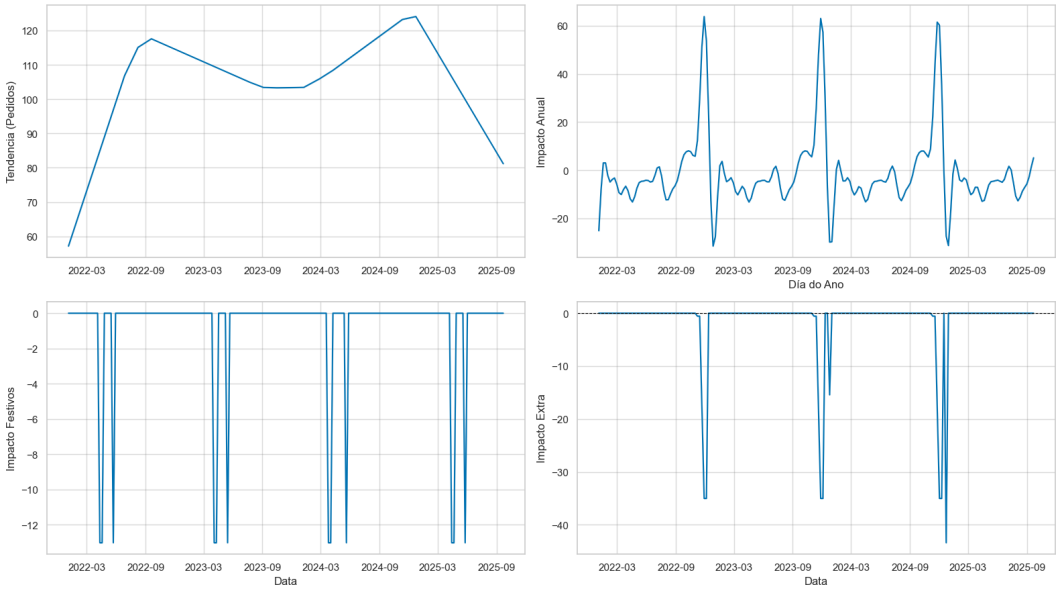


Figura 4.19: Descomposición das compoñentes do modelo *Prophet* para a demanda global. A disposición amosa a tendencia de fondo (superior esquerda), o patrón estacional semanal (superior dereita), o impacto dos días festivos (inferior esquerda) e o efecto dos regresores exóxenos incluídos (inferior dereita).

A primeira gráfica (cuadrante superior esquerdo) representa a compoñente da tendencia $g(t)$. Esta captura a evolución estrutural do negocio a longo prazo, illada de efectos estacionais. A curva mostra tres etapas diferenciadas:

- Unha primeira etapa de crecemento case lineal durante o 2022, característica da fase de recuperación do mercado.
- Durante o 2023 e comezos do 2024 estabilízase o volume nunha demanda relativamente elevada, representando a consolidación no mercado de Trucksters.
- No 2025 obsérvase un cambio de inflexión cara á baixa.

O segundo gráfico (superior dereita) representa a compoñente estacional do modelo $s(t)$. A diferenza da tendencia, este patrón repítese de forma case idéntica ano tras ano. En particular, obsérvase de forma clara a *peak season* mediante unha escalada progresiva durante o outono que culmina en novembro, seguida dunha caída brusca durante o Nadal.

O panel inferior esquerdo representa o impacto das festividades consideradas polo modelo $h(t)$. Como se intuía, o efecto é estritamente negativo. O modelo aprendeu correctamente que, en datas sinaladas a actividade detense, reducindo volume á predición base de forma puntual.

Por último, represéntase o impacto dos regresores inlcuídos no modelo. Obsérvanse picos extremos negativos, que se corresponden cos *outliers* detectados mediante o ARIMA. É importante notar que, ao modelar estes eventos extremos neste panel separado, evítase que "contaminen" a tendencia ou a estacionalidade habitual. O modelo entende que esas caídas drásticas non son parte do comportamento normal do mercado, senón excepcións que deben subtraerse especificamente neses días.

Tras esta primeira aproximación global, abordarase o axuste do modelo *Prophet* para os datos desagregados por clúster.

4.3.2. Modelo para o Clúster 1

En primeiro lugar, estúdase a serie dos clientes de transporte, *retail* e electrónica. Recuperando a análise ARIMA, tan só resultaron significativas dúas variables exóxenas: a primeira modelaba o impacto na demanda do Black Friday, mentres que a restante captaba o impacto na semana posterior ao evento. Nesta serie tampouco se detectaron problemas de heterocedasticidade, co cal o modelo *Prophet* configurouse con estacionalidade aditiva.

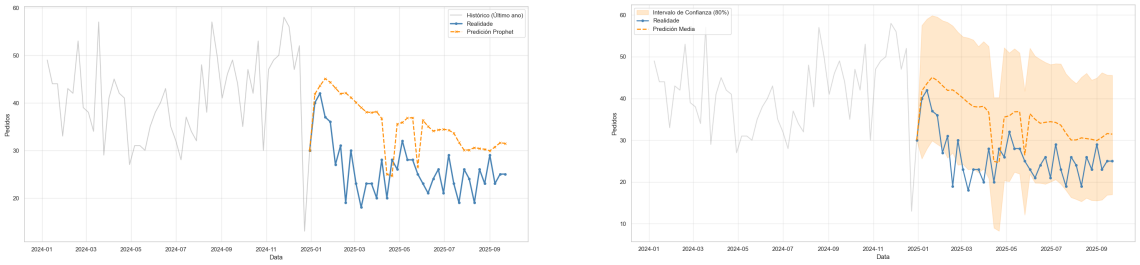
Na selección automática dos hiperparámetros inclúense as dúas variables exóxenas indicadas e os festivos considerados pola empresa. O modelo resultante da validación cruzada presentou un RMSE de 13.91.

Tras axustar o modelo, realízanse as predicións sobre a mostra de test. A Figura 4.20 presenta a comparación gráfica da serie real (azul) coa predición (laranxa). Na Figura (4.20a) obsérvase que, aínda que o modelo captura correctamente a dinámica das flutuacións e o repunte previo ao verán, existe unha sobreestimación sistemática do nivel da demanda. O modelo proxecta un volume de pedidos inferior ao do ano anterior (liña gris), captando a tendencia negativa, pero non con tanta intensidade como a caída real que experimentou o mercado no primeiro trimestre.

Esta sobreestimación confírmase na gráfica (4.20b), que visualiza os intervalos de predición ao 80 % de confianza. A serie real sitúase frecuentemente no límite inferior da banda de confianza, chegando a caer por debaixo da mesma en varios momentos do primeiro trimestre. Isto suxire que a contracción da demanda neste clúster foi máis significativa do que a historia recente da serie permitía anticipar.

Para cuantificar o rendemento, calculáronse as métricas de erro (Cadro 4.19). Os valores do MAE (9.46) e do RMSE (10.87) son elevados en relación ao volume medio da serie, reflectíndose nun MAPE do 40.44 %. Non obstante, o MASE (0.62) ofrece unha lectura moi positiva.

O feito de que o *MASE* sexa significativamente menor a 1 indica que o modelo *Prophet* é moito máis preciso que o preditor naïve estacional. Isto explícase pola forte caída interanual da demanda: mentres que o preditor naïve proxectaría os altos volumes de 2024 (erros moi grandes), *Prophet* foi capaz de



(a) Predicións (b) Intervalos de predición do 80 %

Figura 4.20: Validación mediante *Rolling Window* para a demanda no clúster 1: comparación no conxunto de test (38 semanas) entre a serie real (azul) e as predicións do modelo *Prophet* (laranxa). Á dereita móstrase o detalle das predicións cos seus intervalos de confianza ao 80 %.

anticipar o cambio de tendencia á baixa, aínda que fose de forma conservadora. En conecuencia, malia o erro absoluto, o modelo achega valor preditivo real fronte á simple repetición do ciclo anterior.

Métrica	**Valor**
MAE	9.46
$RMSE$	10.87
$MAPE$	40.44 %
$MASE$	0.62

Cadro 4.19: Avaliación da capacidade predictiva do modelo *Prophet* para a serie do clúster 1: métricas de erro calculadas sobre o conxunto de test (38 semanas). RMSE e MAE mídense en envíos semanais. MAPE é medida porcentual. MASE < 1 indica mellor rendemento ca o preditor naïve.

Por último, interprétanse as distintas compoñentes que conforman o modelo. A Figura 4.21 desagrega a serie temporal do clúster 1 nas súas catro compoñentes. A descomposición revela patróns moi específicos que validan a natureza comercial dos sectores que compoñen este grupo.

A primeira gráfica (superior esquerda) representa a compoñente da tendencia $g(t)$. A evolución mostra un comportamento de "V invertida" moi marcado. Até mediados do 2023 obsérvase un rápido crecemento lineal, onde o volume base pasou de 30 a máis de 42 unidades semanais. A serie acada o seu máximo estrutural durante o terceiro trimestre do 2023 e, a partires deste máximo, a tendencia sofre unha caída constante e simétrica á subida anterior. O modelo proxecta que, para finais de 2025, a demanda estrutural (sen contar estacionalidade) caerá por baixo dos niveis iniciais de 2022. Esta compoñente explica por que o modelo de predición insistía en baixar as proxeccións, malia que a inercia histórica fose alta.

No seguinte panel (superior dereito), represéntase a compoñente da estacionalidade $s(t)$. A diferenza da suavidade observada no modelo global, a estacionalidade do clúster 1 presenta unha alta volatilidade de frecuencia. O patrón reflicte a dinámica propia do sector *retail*, suxeito a múltiples campañas comerciais curtas e agresivas ao longo do ano, máis alá dunha única época de pico. As oscilacións de amplitude indican que a estacionalidade engade unha variabilidade considerable á operativa semanal.

Na parte inferior esquerda da figura, comézase vendo a representación gráfica da compoñente $h(t)$ encargada das festividades. O efecto destas datas mantense consistente coa lóxica do negocio: caídas drásticas da actividade (picos negativos de ata 9 pedidos) que representan o peche operativo en días non laborables.

Por último, obsérvase o impacto dos regresores externos incluídos no modelo. Neste caso, a diferenza do modelo global, tan só se incluían as variables correspondentes á semana do Black Friday e a posterior. Así, se no caso anterior se observaban caídas significativas (*outliers*), aquí preséncianse picos positivos extremos. Ditos impulsos, que engaden aproximadamente 20 pedidos adicionais á demanda base, corresponden ao impacto da *peak season*. A magnitude é considerable: un incremento de 20 pedidos sobre unha base de 40 supón un aumento puntual do 50 % da carga de traballo.

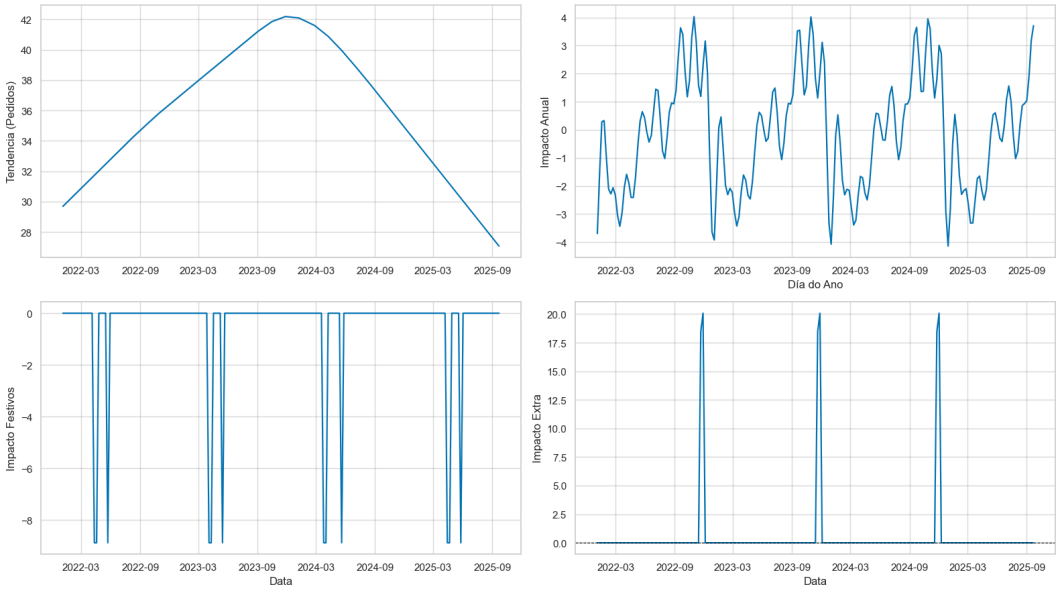


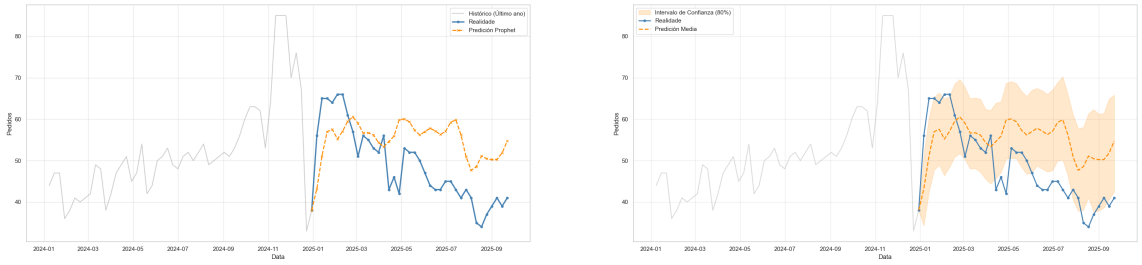
Figura 4.21: Descomposición das compoñentes do modelo *Prophet* para a demanda no clúster 1. A disposición amosa a tendencia de fondo (superior esquerda), o patrón estacional semanal (superior dereita), o impacto dos días festivos (inferior esquerda) e o efecto dos regresores exóxenos incluídos (inferior dereita).

4.3.3. Modelo para o Clúster 2

A continuación modélase o caso dos clientes de paquetería e froita. Partindo da análise ARIMA da sección previa, neste caso só resultaban significativas tres variables exóxenas correspondentes aos *outliers* detectados (31/01/2022, 25/12/2023 e 23/12/2024). Ademais, neste clúster foi necesario aplicar unha transformación logarítmica aos datos para corrixir problemas de heterocedasticidade, polo que se configurará o modelo *Prophet* con estacionalidade multiplicativa.

Así, inclúense estes regresores externos e as festividades na selección automática dos hiperparámetros. O modelo óptimo resultante da validación cruzada presenta un RMSE de 8.64.

Unha vez axustado o modelo, realízanse as predicións sobre a mostra de test (2025). Mediante a Figura 4.22 represéntase a comparación visual da serie real (azul) coa predición (laranxa). Na Figura (4.22a) obsérvase un comportamento dual: mentres que o modelo capta de forma acertada a dinámica do primeiro trimestre, non é quen de axustar a caída estrutural visible nos meses posteriores. O modelo, guiado pola inercia histórica, proxecta un volume de pedidos sostido similar ao do ano anterior (liña



(a) Predicións (b) Intervalos de predición do 80 %

Figura 4.22: Validación mediante *Rolling Window* para a demanda no clúster 2: comparación no conxunto de test (38 semanas) entre a serie real (azul) e as predicións do modelo *Prophet* (laranxa). Á dereita móstrase o detalle das predicións cos seus intervalos de confianza ao 80 %.

gris), sobreestimando a demanda real que experimentaron os sectores de paquetería e froita a partir do segundo trimestre.

Este desaxuste confírmase dramaticamente na Figura (4.22b), que representa os intervalos de predición ao 80 % de confianza. Aínda que nun comezo a serie real se mantén dentro das bandas, rapidamente rompe o límite inferior das mesmas cara á metade do ano. Isto suxire que a caída do mercado neste clúster non foi un simple desvío aleatorio, senón un cambio de réxime estrutural que a historia previa da serie non permitía anticipar.

Co obxectivo de avaliar numericamente o rendemento preditivo, no Cadro 4.20 resúmense as distintas métricas do erro calculadas. Tendo en conta o nivel medio da serie durante o 2025, os valores do MAE (9.61) e do RMSE (10.70) considéranse moderados. Obsérvase de feito un MAPE do 21.34 %, sendo este un resultado aceptable para o mercado.

O dato máis salientable, con todo, é o MASE de 0.66. Este valor inferior a 1 indica que, a pesar de que o modelo non proxecta a caída con exactitude, *Prophet* é considerablemente máis preciso que o preditor naïve estacional. Isto débese a que o modelo suaviza o ruído histórico, mentres que o método inxenuo tería replicado os picos aínda máis altos de 2024, xerando un erro global superior.

Métrica	**Valor**
MAE	9.61
$RMSE$	10.70
$MAPE$	21.34 %
$MASE$	0.66

Cadro 4.20: Avaliación da capacidade predictiva do modelo *Prophet* para a serie do clúster 2: métricas de erro calculadas sobre o conxunto de test (38 semanas). RMSE e MAE mídense en envíos semanais. MAPE é medida porcentual. MASE < 1 indica mellor rendemento ca o preditor naïve.

Para rematar con este clúster, estúdanse as distintas compoñentes do modelo axustado, recollidas na Figura 4.23. A diferenza dos casos anteriores, este modelo configurouse con estacionalidade multiplicativa, o que implica que os efectos estacionais e os eventos non suman unha cantidade fixa de pedidos,

senón que actúan como factores que amplifican ou reducen a tendencia base en termos porcentuais.

Na representación da tendencia (superior esquerda), obsérvase un perfil de crecemento e esgotamento moi definido. A serie comeza cunha forte pendente positiva, pasando dun nivel base de 44 envíos semanais a case superar os 60 e acadando o pico estrutural a comezos do 2023. Dende entón, a tendencia describe unha liña descendente cotinua e suave. Para o horizonte de predición, o modelo estima que a demanda base caerá ata os 49 pedidos.

Na segunda gráfica, represéntase a compoñente estacional $s(t)$, confirmándose a elevada estacionalidade dos sectores de paquetería e froita. Ao ser multiplicativo, o eixo OY móstrase en porcentaxes, revelando flutuacións extremas. A demanda aumenta nun 40 % nos picos e cae preto dun 25 % nos vales respecto á tendencia. Isto implica que a carga de traballo case se duplica entre o momento máis baixo e o máis alto do ano. Ademais, obsérvase un patrón bimodal claro. Un pico forte no verán (xullo-agosto), posiblemente impulsado polas campañas de froita de verán, e outro pico cara a final de ano.

No terceiro gráfico (inferior esquerdo), móstrase o impacto dos festivos. Aínda que estas datas presentan un coeficiente negativo, a súa magnitude neste modelo resulta marxinal (apenas superando o 0.35 %). Isto explícase pola potencia da compoñente estacional. O modelo *Prophet* integrou as grandes caídas de actividade (peches por festivos) directamente na curva de estacionalidade anual, deixando á compoñente $h(t)$ unicamente un rol de axuste residual.

Por último, represéntase o impacto dos *outliers* incluídos no modelo. De forma análoga, obsérvase un impacto porcentual negativo moi reducido (arredor do 0.10 %). Novamente, isto suxire que a estrutura multiplicativa e a estacionalidade anual foron suficientes para capturar a dinámica destas datas sen necesidade de asignar un peso elevado a estes regresores externos, a diferenza do que ocorría nos modelos ARIMA onde non se detectaba estacionalidade.

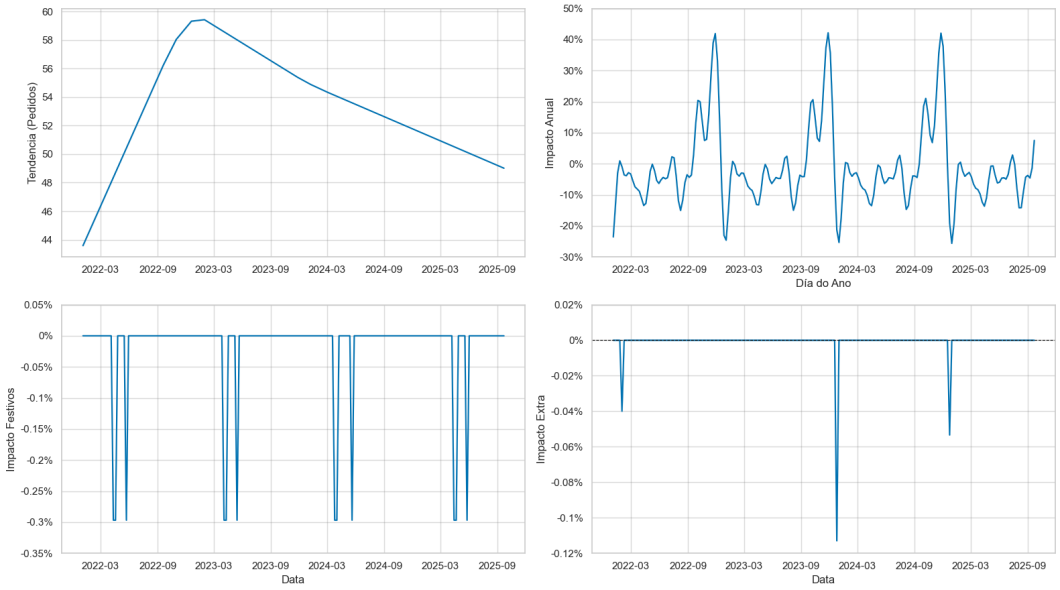


Figura 4.23: Descomposición das compoñentes do modelo *Prophet* para a demanda no clúster 2. A disposición amosa a tendencia de fondo (superior esquerda), o patrón estacional semanal (superior dereita), o impacto dos días festivos (inferior esquerda) e o efecto dos regresores exóxenos incluídos (inferior dereita).

4.3.4. Modelo para o Clúster 3

Para rematar, estúdase a serie temporal dos clientes de alimentación e industria. Recuperando a análise ARIMA, neste caso ningún dos regresores externos considerados resultou significativo, amosándose un efecto residual da *peak season* e sen detectar ningún *outlier*. Ademais, aplicouse a transformación logarítmica sobre os datos, co fin de corrixir problemas de heterocedasticidade. Daquela, o modelo *Prophet* neste clúster configurarase con estacionalidade multiplicativa.

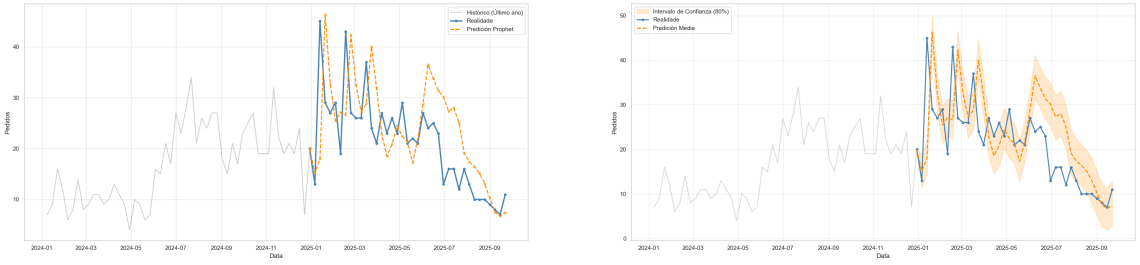
Así pois, tendo en conta os correspondentes festivos, procedeuse á selección automática dos hiperparámetros, resultando un modelo que presenta un RMSE na validación cruzada de 5.78.

Unha vez axustado o modelo, realízanse as predicións sobre a mostra de test. Na Figura 4.24 represéntase a comparación visual da serie real (azul) fronte a predición (laranxa).

A Figura (4.24a) permite apreciar a natureza suavizadora do modelo *Prophet*. Durante o primeiro trimestre, o modelo demostra unha boa capacidade para detectar a sincronía temporal dos picos de demanda (axusta correctamente os momentos de subida e baixada).

Pola contra, a partir do segundo trimestre, cando a demanda real sofre unha caída severa, obsérvase unha clara tendencia á sobreestimación. O modelo, influenciado pola inercia histórica, proxecta un descenso máis suave ca o real, mantendo a predición (liña laranxa) sistematicamente por riba dos valores observados (liña azul) durante os meses centrais do ano. Non é ata o último trimestre cando ambas series volven converxer.

Esta lectura confírmase na gráfica dos intervalos de predición (Figura (4.24b)). Mentres que no inicio do ano a serie real escapa ocasionalmente polo límite superior (picos non captados), durante a fase central sitúase frecuentemente rozando o límite inferior da banda de confianza, confirmando que a caída do mercado foi máis profunda do que o rango de incerteza do modelo estimaba probable.



(a) Predicións

(b) Intervalos de predición do 80 %

Figura 4.24: Validación mediante *Rolling Window* para a demanda no clúster 3: comparación no conxunto de test (38 semanas) entre a serie real (azul) e as predicións do modelo *Prophet* (laranxa). Á dereita móstrase o detalle das predicións cos seus intervalos de confianza ao 80 %.

Finalmente, o Cadro 4.21 resume o desempeño numérico. O MAPE sitúase nun 35.84 %, un valor penalizado pola sobreestimación do val central. O dato máis relevante é o MASE de 0.97. Ao situarse lixeiramente por debaixo de 1, indica que o modelo mellora, aínda que de forma axustada, o rendemento do preditor naïve estacional. Isto suxire que, a pesar das dificultades para replicar a amplitude exacta da volatilidade, *Prophet* aporta valor ao anticipar correctamente os puntos de xiro e a dirección da tendencia, superando a un modelo "inxenuo" que simplemente repetiría os niveis do ano anterior.

Así e todo, é importante recordar que a alta volatilidade deste clúster resultaba nun rendemento todavía peor na metodoloxía de Box-Jenkins. Cómpre matizar que, mentres que a avaliación do modelo ARIMA se realizou fronte ao preditor naïve básico ($m = 1$), obtendo un MASE de 1.42, a avaliación de *Prophet* realizouse contra un competidor máis robusto para este tipo de datos: o naïve estacional ($m = 52$).

Métrica	Valor
MAE	7.34
$RMSE$	9.50
$MAPE$	35.84 %
$MASE$	0.97

Cadro 4.21: Avaliación da capacidade predictiva do modelo *Prophet* para a serie do clúster 3: métricas de erro calculadas sobre o conxunto de test (38 semanas). RMSE e MAE mídense en envíos semanais. MAPE é medida porcentual. MASE < 1 indica mellor rendemento ca o preditor naïve.

Neste contexto, o feito de que o *Prophet* lograse un MASE inferior a 1 (0.97), mentres que ARIMA non logrou mellorar nin sequera á predición inxenua simple, evidencia a superioridade da aproximación de compoñentes para capturar a complexa dinámica dos clientes deste clúster.

Para rematar con esta análise, estúdanse as compoñentes do modelo *Prophet* construído para os clientes de alimentación e industria. A Figura 4.25 desagrega a serie temporal nas súas partes constituíntes. É importante lembrar que, ao igual que no clúster 2, aquí tamén se aplicou unha estacionalidade multiplicativa, polo que as variacións estacionais exprésanse en porcentaxes sobre a tendencia base.

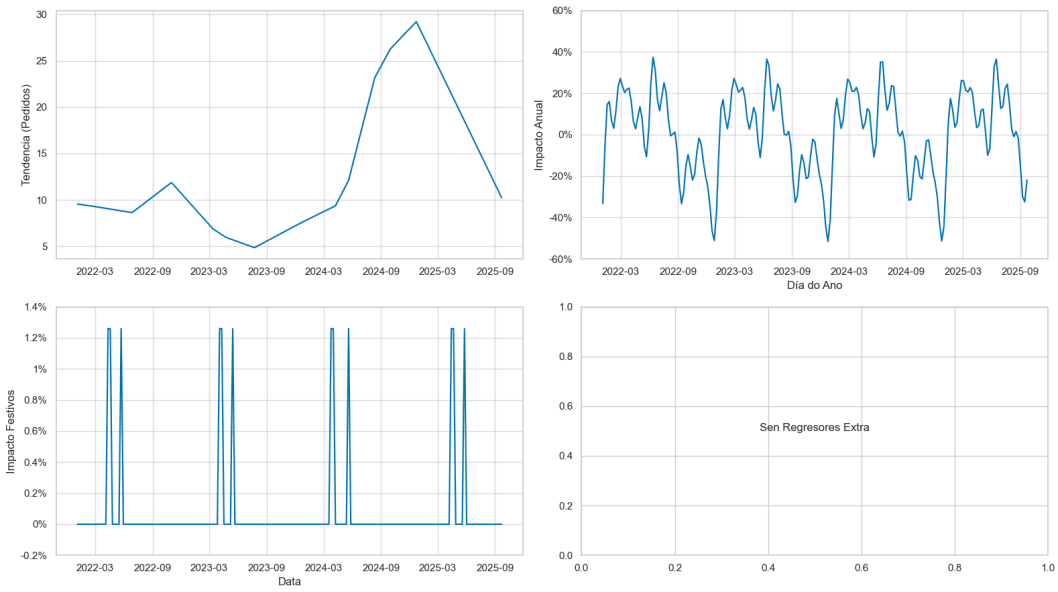


Figura 4.25: Descomposición das compoñentes do modelo *Prophet* para a demanda no clúster 3. A disposición amosa a tendencia de fondo (superior esquerda), o patrón estacional semanal (superior dereita), o impacto dos días festivos (inferior esquerda) e o efecto dos regresores exóxenos incluídos (inferior dereita). Neste caso non hai regresores externos.

A primeira gráfica (superior esquerda) representa a tendencia do modelo ($g(t)$). A curva de tendencia explica a forma da predición final. Durante os dous primeiros anos, a demanda mantívose en niveis baixos (entre 5 e 10 pedidos), tocando un mínimo estrutural a mediados de 2023. A partir de

finais de 2023, obsérvase un crecemento vertical e explosivo, triplicando o volume base ata rozar os 30 pedidos a mediados de 2024. Así, o modelo detecta que ese crecemento non era sostible e proxecta unha corrección igual de agresiva para 2025, devolvendo a demanda aos niveis de 2022 (arredor de 10 pedidos). Esta compoñente é a responsable de que a predición (liña laranxa na gráfica) teña esa pendente negativa tan marcada.

Con respecto á compoñente estacional ($s(t)$), na segunda gráfica reflíctese a natureza irregular do sector industrial e de alimentación. As flutuacións son extremas, oscilando entre un incremento do 40 % e unha diminución do 50 % respecto á tendencia. Isto indica que, á marxe do volume anual, a carga de traballo semanal é moi inestable. A diferenza das ondas suaves dos outros clústeres, aquí obsérvanse "dentes de serra" moi frecuentes.

Por outra banda, sobre a compoñente que capta o impacto dos festivos ($h(t)$), chama a atención que o efecto estimado para os festivos é lixeiramente positivo (arredor do 1.2 %). Non obstante, ao comparar a escala desta compoñente coa da estacionalidade, conclúese que o impacto dos festivos é absolutamente marxinal e desprezable. O modelo confirma que, para este tipo de clientes industriais, o calendario laboral ten moita menos relevancia operativa que os seus propios ciclos de produción.

Por último, tal e como se indicou previamente, neste modelo non se incluíron regresores externos, posto que non resultaron significativos durante a análise feita mediante os modelos ARIMA. Por iso o último gráfico da Figura 4.25 mostra unha representación baleira.

En síntese, a implementación da metodoloxía *Prophet* permite extraer conclusións fundamentais sobre a capacidade preditiva e a estrutura da demanda en Trucksters:

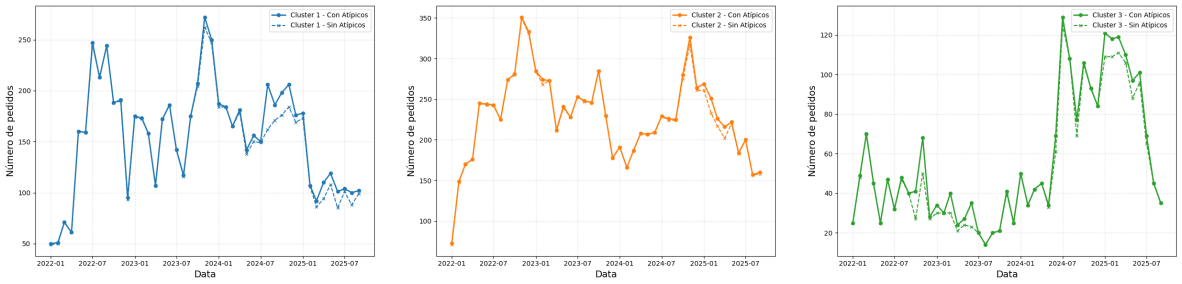
- **Superioridade estatística fronte ao modelo inxenuo:** O indicador máis robusto do éxito desta aproximación é o valor do MASE, que se situou sistematicamente por debaixo da unidade en tódalas series ($0.62 < \text{MASE} < 0.97$). Isto confirma que, malia as perturbacións da serie, o modelo baseado en compoñentes é significativamente máis eficaz que a simple réplica do ciclo anual anterior (*Seasonal Naïve*), logrando capturar a tendencia de fondo e as estacionalidades complexas que a metodoloxía ARIMA non sempre acadaba a procesar.
- **Interpretabilidade e valor de negocio:** A gran vantaxe competitiva de *Prophet* reside na transparencia da súa descomposición. A capacidade de illar a tendencia $g(t)$ dos efectos estacionais $s(t)$ e das intervencións exóxenas $h(t)$ permite á empresa diferenciar entre o crecemento estrutural do negocio e as flutuacións puntuais de mercado.
- **Dificultade ante o cambio de réxime en 2025:** A tendencia á sobreestimación detectada na fase final da mostra de test non se debe a unha deficiencia do algoritmo, senón a un cambio de réxime estrutural na demanda. Como se analizou, a nova política comercial orientada á rendibilidade xerou unha contracción no volume que a historia previa da serie non permitía anticipar.

Tras levar a cabo os distintos axustes durante a presente sección, farase o propio sobre o conxunto de datos reducido resultante da análise de atípicos da Sección 3.4.

4.4. Análise de sensibilidade: Aplicación a datos sen atípicos

Durante a presente sección estudarase o impacto dos atípicos detectados no Capítulo 3. A metodoloxía a seguir consistirá en repetir os procedementos desenvoltos durante as seccións previas aplicados ao conxunto de datos reducido.

Levando a cabo unha primeira análise gráfica do impacto dos *outliers* sobre as series temporais de demanda, á vista da Figura 4.26 é de esperar que as principais diferenzas con respecto aos modelos construídos nas Seccións 4.2 e 4.3 se acaden nos clientes de *retail*, transporte e electrónica, por ser dito clúster onde se observan maiores diferenzas no volume de pedidos (Figura (4.26a)).



(a) Cluster 1: Pedidos (b) Cluster 2: Pedidos (c) Cluster 3: Pedidos

Figura 4.26: Impacto da eliminación de atípicos no volume de pedidos mensuais por clúster (2022-2025). O maior impacto obsérvase na serie do clúster 3 (verde), correspondente aos sectores de alimentación e industria.

Así e todo, o obxectivo consistirá en contrastar unha posible mellora dos modelos ao traballar con datos máis estables.

Cómpre sinalar que, dada a extensa explicación levada a cabo dos modelos axustados sobre a base de datos orixinais, nesta sección tan só se presentará un resumo dos resultados acadados. Se o lector desexa afondar no axuste dos distintos modelos pode consultar o código empregado (Airas, 2026).

4.4.1. Modelos ARIMA

Para o axuste da metodoloxía de Box-Jenkins, empregarase a información obtida da análise de intervención levada a cabo para cada unha das series na Sección 4.2. Para cada un dos modelos, explicitarase a estimación dos parámetros tras a detección de *outliers* e a eliminación de termos non significativos no modelo seleccionado, así como unha comparación da precisión das predicións calculadas con respecto aos resultados obtidos sobre os datos orixinais.

Modelo Global

Para a serie global, o impacto do Black Friday modelárase mediante tres variables *dummy* que abarcan un espazo temporal de 5 semanas. Unha vez seleccionado o modelo óptimo en termos do AIC, tras aplicar o algoritmo de detección de *outliers* detectáronse dúas observacións atípicas que coinciden coas achadas no modelo orixinal (**24/12/2023** e **23/12/2024**). Así, os coeficientes estimados do modelo resultante (unha vez eliminados os termos non significativos) pódense consultar no Cadro 4.22.

En comparación co modelo estimado para os datos orixinais (Cadro 4.2) obsérvanse resultados certamente similares. En particular, os parámetros estimados presentan valores lixeiramente inferiores neste caso, mais cunha interpretación análoga á explicada no modelo orixinal.

Tras validar o modelo así axustado, proceso no que non se afondará nesta sección, procedeuse a avaliar a capacidade preditiva do mesmo. Os resultados das métricas do erro das predicións recóllense no Cadro 4.23, que permite ademais comparar os resultados cos dos datos orixinais. En efecto, obsérvase unha lixeira mellora, reducíndose nun envío semanal de media o erro cometido. Esta melloría reflíctese nun menor erro porcentual (MAPE = 7.81 %) así como no MASE que se reduce de 0.60 nos datos orixinais a 0.54 sobre a base de datos reducida.

Os resultados acadados reforzan o bo rendemento preditivo que presenta a metodoloxía de Box-Jenkins sobre a serie de demanda global, confirmando un impacto positivo da eliminación de observacións atípicas.

Parámetro	Estimación	Erro Estándar	Estatístico z	P-valor
BF_{pre}	16.003	5.913	2.706	0.007
BF_{peak}	27.263	13.454	2.026	0.043
BF_{post}	14.755	5.202	2.836	0.005
$\text{O}_{23/12/24}$	-62.457	8.863	-7.047	< 0.001
$\text{O}_{25/12/23}$	-41.833	9.271	-4.512	< 0.001
θ_1 (MA_1)	-0.527	0.082	-6.427	< 0.001
σ^2 (Varianza)	210.745	22.421	9.400	< 0.001

Cadro 4.22: Estimación dos parámetros do modelo ARIMA$(0, 1, 1)$ con regresores para a serie global depurada. Modelo final tras a eliminación de termos non significativos. A estimación dos parámetros mídese en envíos/semana.

Métrica	Serie Orixinal	Serie Sen Atípicos
MAE	7.90	7.07
$RMSE$	11.15	9.84
$MAPE$	8.18 %	7.81 %
$MASE$	0.60	0.54

Cadro 4.23: Comparativa do rendemento preditivo na serie global: Impacto da eliminación de *outliers*.

Seguindo a estrutura da Sección 4.2, repetirase o procedemento para cada clúster por separado, contrastando as posibles diferencias do impacto dos *outliers* en función do tipo de cliente.

Modelo para o Clúster 1

Para os clientes do clúster 1 (transporte, *retail* e electrónica), o impacto da *peak season* tan só abranguía un intervalo temporal de tres semanas, sendo o efecto máis reducido ca no modelo global. Aplicando o procedemento de selección do modelo en termos do AIC e tras facer o propio co algoritmo de detección de *outliers*, detectouse como atípica a observación do **25/12/2023**, contrastando coa ausencia de *outliers* no caso orixinal. Os parámetros estimados do modelo axustado unha vez eliminados os coeficientes non significativos recóllense no Cadro 4.24.

En contraste co modelo estimado para os datos orixinais (Cadro 4.6), a principal diferenza reside na incorporación dun parámetro adicional para estimar o efecto do *outlier* detectado. En particular, o atípico presenta un impacto negativo de 24 envíos na semana do 25/12/2023. O resto de parámetros presentan valores similares ao modelo orixinal, sendo a súa interpretación análoga.

Parámetro	Estimación	Erro Estándar	Estatístico z	P-valor
BF_{peak}	9.706	3.692	2.629	0.009
BF_{post}	11.361	4.460	2.547	0.011
$\text{O}_{25/12/23}$	-24.096	5.343	-4.510	< 0.001
θ_1 (MA_1)	-0.482	0.069	-7.032	< 0.001
σ^2 (Varianza)	98.575	8.826	11.168	< 0.001

Cadro 4.24: Estimación dos parámetros do modelo ARIMA$(0, 1, 1)$ con regresores para a serie do clúster 1 depurada. Modelo final tras a eliminación de termos non significativos. A estimación dos parámetros mídese en envíos/semana.

Posteriormente, tralo proceso de diagnose, avaliouse a capacidade preditiva do modelo sobre a mostra de test. No Cadro 4.25 recóllense os resultados das métricas do erro tanto para a serie orixinal coma para a serie axustada sobre os datos sen atípicos. En particular, a precisión do modelo resulta moi similar ao caso orixinal, reducíndose apenas o erro absoluto medio en 0.18 envíos semanais. Porén, esta lixeira disminución non se proxecta en termos do erro porcentual, sendo neste caso o MAPE = 16.28 %, indicando que erros de 3.70 envíos semanais supoñen unha proporción maior da demanda ca os 3.88 envíos semanais na serie orixinal. O aumento nesta métrica explícase pola diminución no volume de demanda semanal neste clúster ao eliminar os atípicos, feito que xa se sinalou na Figura 4.26. Así e todo, en termos do MASE non se aprecian diferenzas, acadándose en ambos casos un resultado que mellora de forma moi significativa a predición naïve (0.45).

Métrica	Serie Orixinal	Serie Sen Atípicos
MAE	3.88	3.70
$RMSE$	4.85	4.82
$MAPE$	15.57 %	16.28 %
$MASE$	0.45	0.45

Cadro 4.25: Comparativa do rendemento preditivo na serie do clúster 1: Impacto da eliminación de *outliers*.

En conclusión, a pesar de que os clientes de *retail*, transporte e electrónica acusaban un maior impacto ao eliminar as observacións atípicas da súa serie de demanda, o modelo ARIMA axustado sobre estes datos non consegue mellorar os resultados obtidos no axuste sobre a serie orixinal.

Modelo para o Clúster 2

Os clientes dos sectores de paquetería e froita presentaban unha maior sensibilidade á *peak season*, modelándose o efecto da mesma dende as dúas semanas previas ao Black Friday até a semana posterior

ao evento. Da mesma forma ca na serie orixinal, foi preciso aplicar unha transformación logarítmica (ec. 2.10, $\lambda = 0$) para estabilizar a varianza.

Solucionado o problema da heterocedasticidade, aplicouse o precedemento de selección do modelo óptimo en termos do AIC. Posteriormente, o algoritmo de detección de *outliers* marcou as mesmas observacións atípicas ca no modelo orixinal: **31/01/2022**, **25/12/2023** e **23/12/2024**. Así, os coeficientes estimados do modelo resultante (trala eliminación dos termos non significativos) recóllense no Cadro 4.26.

O modelo presenta características similares ao axustado na serie orixinal, destacando a ausencia de significación dos coeficientes que modelaban o impacto do Black Friday. Ademais, a compoñente autorregresiva de segunda orde (ϕ_2) resulta significativa a pesar de non incluírse o termo de primeira orde (ϕ_1). As estimacións dos parámetros restantes apenas difiren coas do modelo axustado sobre a serie orixinal, co cal a súa interpretación xa foi abordada.

Parámetro	**Estimación**	**Erro Estándar**	**Estatístico z**	**P-valor**
$O_{31/01/22}$	0.388	0.037	10.604	< 0.001
$O_{23/12/24}$	-0.384	0.047	-8.134	< 0.001
$O_{25/12/23}$	-0.368	0.076	-4.817	< 0.001
ϕ_2 (AR_2)	-0.202	0.092	-2.191	0.028
σ^2 (Varianza)	0.018	0.002	10.199	< 0.001

Cadro 4.26: Estimación dos parámetros do modelo ARIMA$(2, 1, 0)$ con regresores para a serie do clúster 2 transformada. Modelo final tras a eliminación de termos non significativos.

Tras isto, e unha vez avaliado o modelo axustado, mediuse a calidade das predicións sobre os datos do 2025 (mostra de test). O Cadro 4.27 resume os resultados acadados sobre a serie orixinal e sobre a correspondente serie modificada. Neste caso a precisión do modelo é certamente positiva, mellorando o erro absoluto medio (pasa de 3.60 a 3.07 envíos semanais). A diferenza do que sucedía no clúster 1, esta mellora en termos absolutos supón unha diminución do erro porcentual, representando o erro cometido un 6.49 % do volume medio semanal. En consecuencia, tamén se produce unha mellora considerable en termos do MASE, reducíndose do 0.63 acadado sobre a serie orixinal a 0.55 neste caso.

A pesar de que o clúster 2 amosaba menos sensibilidade á eliminación de pedidos atípicos, os resultados acadados para os clientes de paquetería e froita presentan unha mellora considerable sobre os da serie orixinal.

Modelo para o Clúster 3

Para finalizar coa metodoloxía clásica de Box-Jenkins, estúdase o impacto dos atípicos sobre os sectores de alimentación e industria. Recórdese que, na análise de intervención previa ao axuste do modelo para a serie orixinal, observouse un efecto marxinal do Black Friday, modelándose o impacto tan só dende a semana previa ao evento até a posterior ao mesmo. Seguindo o procedemento aplicado sobre a serie orixinal, foi preciso estabilizar a varianza da serie mediante unha transformación logarítmica (ec. 2.10, $\lambda = 0$).

Tras solucionarse o problema de heterocedasticidade presente na serie temporal, seleccionouse o

Métrica	Serie Orixinal	Serie Sen Atípicos
MAE	3.60	3.07
$RMSE$	4.94	4.32
$MAPE$	7.38 %	6.49 %
$MASE$	0.63	0.55

Cadro 4.27: Comparativa do rendemento preditivo na serie do clúster 2: Impacto da eliminación de *outliers*.

modelo óptimo en termos do AIC. Ao aplicar o algoritmo de detección de *outliers* identificouse unha observación atípica: **02/05/2022**. Isto difire co caso orixinal, onde o algoritmo non detectaba ningún atípico. Ademais, chama a atención que o *outlier* sexa no mes de maio, pois durante os modelos axustados ao longo do capítulo, os atípicos indentificados correspondíanse na súa maioría con semanas de decembro ou xaneiro, reflectindo o complexo comportamento do mercado posterior ao Black Friday.

Parámetro	Estimación	Erro Estándar	Estatístico z	P-valor
$O_{02/05/22}$	-1.088	0.358	-3.037	0.002
θ_1 (MA_1)	-0.457	0.088	-5.219	< 0.001
θ_2 (MA_2)	-0.216	0.085	-2.539	0.011
σ^2 (Varianza)	0.113	0.012	9.607	< 0.001

Cadro 4.28: Estimación dos parámetros do modelo ARIMA$(0, 1, 2)$ con regresores para a serie do clúster 3 transformada. Modelo final tras a eliminación de termos non significativos.

As estimacións do modelo resultante tras eliminar aqueles termos non significativos amósanse no Cadro 4.28. Ademais da inclusión do *outlier* non presente no modelo axustado na serie orixinal, obsérvase un achado moi relevante: mentres que para serie orixinal se identificou un ARIMA$(2, 1, 0)$ (Cadro 4.14), na serie depurada axústase un proceso ARIMA$(0, 1, 2)$. Este cambio ten unha interpretación estatística clara: a ausencia de *outliers* na serie orixinal estaba a introducir unha falsa estrutura de autocorrelación. Isto implica que, en condicións normais de operación, a demanda non presenta unha dependencia forte do nivel de pedidos das semanas previas.

Para interpretar os coeficientes estimados é preciso recordar que se está a traballar coa serie transformada.

Así, observando o Cadro 4.28, os parámetros da parte de medias móbiles son negativos, reflectindo a dinámica de corrección das innovacións no modelo. Pola súa parte, o coeficiente asociado ao *outlier* (-1.088) indica unha caída do 66 % $((1 - e^{-1.088}) \cdot 100)$ no volume de pedidos polo efecto do atípico.

Posteriormente, tras superar o modelo a etapa de diagnose, avaliouse o calidade das predicións sobre a mostra de test. No Cadro 4.29 resúmense as métricas do erro cometido tanto na serie orixinal

coma na serie depurada. En particular, o rendemento neste clúster resultaba moi baixo, empeorándose inclusive a predición do modelo "inxenuo". Obsérvase así que, tras depurar a serie, os resultados non melloran significativamente. O MAE apenas diminúe (4.33 fronte ao 4.59 orixinal), representando de feito un erro porcentual maior (MAPE = 22.01 %). Ademais, o resultado en termos do MASE continúa sendo moi negativo, empeorándo de forma considerable ao preditor naïve.

Porén, convén recordar a gran dificultade que presentaba o estudo da serie de demanda para os clientes deste clúster: a alta volatilidade durante o primeiro trimestre do 2025. A presenza destes picos de demanda complica a capacidade preditiva das aproximación de Box-Jenkins. En particular, a eliminación de atípicos apenas reduciu a volatilidade existente (Figura 4.26), derivando nun rendemento preditivo similar sobre a serie depurada.

Métrica	**Serie Orixinal**	**Serie Sen Atípicos**
MAE	4.59	4.33
$RMSE$	7.65	6.67
$MAPE$	20.73 %	22.01 %
$MASE$	1.42	1.44

Cadro 4.29: Comparativa do rendemento preditivo (ARIMA$(2,1,0)$ vs ARIMA$(0,1,2)$) na Serie do clúster 3: Impacto da eliminación de *outliers*.

Aplicando o mesmo razonamento empregado durante o estudo dos resultados sobre a serie orixinal (Sección 4.2), recalculáronse as métricas excluíndo o primeiro trimestre da mostra de test. Así, no Cadro 4.30 amósase unha mellora drástica no rendemento. Porén, os resultados conseguidos sobre a serie orixinal son lixeiramente mellores neste caso: o erro absoluto medio diminúe ata 2.68 envíos fronte aos 2.82 envíos semanais na serie modificada, mentres que en termos porcentuais os resultados seguen sen ser particularmente bos (17.21 % e 19.88 % respectivamente). Non obstante, a diferenza máis positiva atópase no MASE, mellorando neste caso os modelos en ambas series respecto ao preditor naïve (0.83 e 0.94) aínda que de forma minoritaria.

Métrica	**Serie Orixinal**	**Serie Sen Atípicos**
MAE	2.68	2.82
$RMSE$	3.46	3.53
$MAPE$	17.21 %	19.88 %
$MASE$	0.83	0.94

Cadro 4.30: Comparativa do rendemento preditivo ($ARIMA(2,1,0)$ vs $ARIMA(0,1,2)$) na Serie do clúster 3 recortada: Impacto da eliminación de *outliers*.

En conclusión, na serie de demanda dos clientes de alimentación e industria, a eliminación de envíos atípicos non supuxo unha mellora en termos preditivos. En efecto, os resultados amosados polo modelo

axustado sobre a serie depurada resultaron peores ca os obtidos sobre a serie de pedidos orixinais. Estes resultados suxiren que, para series altamente volátiles e con picos concentrados, é recomendable explorar modelos alternativos ou máis robustos que poidan captar mellor a dinámica irregular da demanda.

Antes de sacar conclusións sobre o impacto dos *outliers* neste traballo, débese repetir estudo sobre os modelos *Prophet*, comparando os resultados cos obtidos na Sección 4.3.

4.4.2. Modelos *Prophet*

Para que os resultados sexan comparables, continuarase coa metodoloxía explicada na Sección 4.1, aproveitando a información estrutural aportada polos modelos ARIMA á hora de aplicar a metodoloxía *Prophet*. En particular, tanto os atípicos coma as correspondentes variables *dummy* que resultaron significativas no axuste de Box-Jenkins incluiranse como variables exóxenas no modelo *Prophet*.

Tras isto, a selección dos hiperparámetros do modelo levarase a cabo cabo mediante o procedemento de *Grid Search* empregando *rolling origin cross-validation*. Ademais, a estrutura de estacionalidade seleccionouse en función das características da varianza estudadas no ARIMA. Así, se o modelo resultaba heterocedástico, considerouse unha estacionalidade multiplicativa, mentres que no caso de non precisar a serie unha corrección da varianza, mantívose a estacionalidade aditiva.

Finalmente, avaliarase o rendemento preditivo do modelo axustado sobre a serie sen *outliers* empregando as mesmas métricas dos modelos ARIMA, cunha diferenza clave en termos do MASE. Esta variación fundaméntase na diferenza de construción dos modelos: mentres que os ARIMA non incluían compoñente estacional, esta resulta elemental no axuste *Prophet*, provocando que o modelo "inxenuo" co que se comparan os resultados sexa o naïve estacional durante a presente sección.

Modelo Global

Unha vez máis, iníciase a análise mediante o estudo da serie global. Recuperando o axuste ARIMA previo, incluiranse 5 variables exóxenas no modelo *Prophet*: dúas modelan o efecto dos atípicos detectados (25/12/2023 e 23/12/2024), mentres que as outras tres reflicten o impacto da *peak season* sobre a serie. Tras incluír no modelo tanto as variables *dummy* coma os correspondentes festivos (Apéndice B), seleccionáronse os hiperparámetros acadando o modelo óptimo un RMSE= 19.15 na fase de validación.

Unha vez axustado o modelo, mediuse a súa calidade preditiva sobre a mostra de test, obtendo os resultados que se recollen no Cadro 4.31. En comparación cos resultados acadados sobre a serie orixinal, prodúcese unha lixeira mellora: en termos de erro absoluto, o modelo equivócase en case 15 envíos semanais (14.90) fronte aos 16.19 na serie completa. Isto proxéctase nunha diminución do 1 % a nivel porcentual, sendo o erro cometido inferior ao 18. %. Non obstante, en relación á comparación co preditor naïve estacional, apenas se aprecian diferenzas (0.72 fronte a 0.71).

Os resultados obtidos reafirman o bo comportamento que amosa a metodoloxía *Prophet* sobre a serie da demanda global, mostrando un impacto positivo da eliminación dos *outliers* nunha escala similar á presenciada nos modelos ARIMA.

A contiuación, procederase do mesmo xeito para cada un dos grupos de clientes cos que se está a traballar por separado.

Modelo para o Clúster 1

Os clientes dos sectores de *retail*, transporte e electrónica tan só presentaban tres variables exóxenas no axuste de Box-Jenkins: un atípico na semana do 25/12/2023, e dúas variables para capturar o efecto do Black Friday durante a propia semana do evento e a posterior. Procedeuse daquela á selección automática dos hiperparámetros do axuste incluíndo estas varibles *dummy* e os correspondentes festivos. O modelo óptimo presentou un RMSE= 13.65 na fase de validación.

Métrica	Serie Orixinal	Serie Sen Atípicos
MAE	16.19	14.90
$RMSE$	19.19	17.43
$MAPE$	18.53 %	17.57 %
$MASE$	0.72	0.71

Cadro 4.31: Comparativa do rendemento preditivo (*Prophet*) na serie global: Impacto da eliminación de *outliers*.

Tras axustar o modelo resultante, avaliáronse as predicións sobre a mostra do 2025, mostrándose os respectivos resultados no Cadro 4.32. Así, comparando as métricas coas acadadas no axuste da serie orixinal, a mellora resulta considerablemente positiva, diminuíndose todas as medidas calculadas case no 50 %. En particular, o erro absoluto medio cometido é inferior aos 5 envíos semanais, sendo este un gran resultado.

Ademais, en termos do MASE pásase dun 0.62 cometido na serie orixinal a un 0.30 sobre a serie depurada. Porén, a nivel porcentual, aínda que se reduce drásticamente o valor, o erro cometido segue representando unha propoción considerable do volume de envíos (20.24 %).

Métrica	Serie Orixinal	Serie Sen Atípicos
MAE	9.46	4.50
$RMSE$	10.87	5.72
$MAPE$	40.44 %	20.24 %
$MASE$	0.62	0.30

Cadro 4.32: Comparativa do rendemento preditivo do modelo *Prophet* na serie do clúster 1: Impacto da eliminación de *outliers*.

Así, a diferencia do observado no axuste ARIMA, o impacto da eliminación dos atípicos na serie de demanda do clúster 1 resulta certamente positivo no modelo *Prophet*. En especial, cómpre sinalar o valor acadado para o MASE, sendo neste caso de 0.30. Isto demostra a gran mellora do *Prophet* fronte ao preditor naïve, captando correctamente a caída na demanda semanal apreciada na serie.

Modelo para o Clúster 2

Para os clientes do clúster 2 (paquetería e froita) non resultaba significativa ningunha das variables *dummy* que modelaban o impacto da *peak season* sobre a serie temporal da demanda. Porén, identificábanse tres atípicos: 31/01/2022, 25/12/2023 e 23/12/2024. Tras incluír estas tres variables no modelo, aplicouse o algoritmo de selección automática de hiperparámetros, acadando o modelo óptimo un RMSE= 8.13.

Os resultados dos erros cometidos polas predicións do modelo resultante amósanse no Cadro 4.33. A comparativa das métricas obtidas na serie orixinal fronte aos correspondentes valores na serie sen atípicos mostran unha mellora lixeira: redúcese o erro absoluto medio de envíos semanais (baixa de 9.61 a 8.46), supoñendo unha diminución dun 2 % no MAPE, sendo o erro cometido inferior ao 20 %. Consecuentemente, tamén se aprecia unha redución considerable do MASE, reducíndose do 0.66 na serie orixinal a 0.59 neste caso.

Métrica	**Serie Orixinal**	**Serie Sen Atípicos**
MAE	9.61	8.46
$RMSE$	10.70	9.57
$MAPE$	21.34 %	19.21 %
$MASE$	0.66	0.59

Cadro 4.33: Comparativa do rendemento preditivo do modelo *Prophet* na serie do clúster 2: Impacto da eliminación de *outliers*.

En conclusión, o impacto da eliminación dos atípicos resultou positivo nos sectores de paquetería e froita, mais sendo dita mellora moito menor ca no clúster anterior. Así e todo, o comportamento aseméllase bastante ao observado para este clúster na metodoloxía de Box-Jenkins.

Modelo para o Clúster 3

Finalmente, procedese ao estudo da serie dos clientes de alimentación e industria. No axuste correspondente ARIMA tan só resultaba significativo a variable *dummy* que modelaba o atípico detectado mediante o algoritmo na semana do 02/05/2022.

Tras configurar o modelo coa estacionalidade multiplicativa e a correspondente variable exóxena, aplicouse o algoritmo de selección automática de hiperparámetros. O modelo óptimo seleccionado presentou un RMSE = 6.75 na fase de validación.

Unha vez axustado o modelo, calculáronse as predicións sobre a mostra de test, comparando os resultados co valor da serie real. Así, o Cadro 4.34 permite contrastar as métricas calculadas sobre a serie orixinal fronte as acadadas co axuste na serie depurada de atípicos. Á vista dos resultados, a eliminación dos atípicos presentou un impacto positivo sobre a calidade preditiva dos modelos. O erro absoluto medio é inferior aos 7 envíos semanais, representando un erro porcentual de case o 36 %. Adicionalmente, o MASE diminúese minimamente de 0.97 na serie orixinal a 0.91 neste caso.

A pesar de que os resultados conseguidos para os clientes de alimentación e industria non resulten tan positivos coma os dos outros clústeres, en comparación aos acadados mediante a metodoloxía clásica (Cadro 4.29), estes resultan considerablemente mellores en termos do MASE. Así, a gran volatilidade observada no volume semanal durante o primeiro trimestre do 2025 resulta case imposible de capturar por parte dos modelos, derivando en predicións que apenas melloran aos correspondentes preditores naïve.

Non obstante, en relación ao impacto dos *outliers*, a metodoloxía *Prophet* amosou unha mellora ao aplicarse sobre a serie depurada, contrastando co comportamento no ARIMA.

Na seguinte sección levarase a cabo unha comparación dos resultados obtidos neste capítulo, contrastando o rendemento da metodoloxía clásica (Box-Jenkins) fronte aos modelos *Prophet*.

Métrica	Serie Orixinal	Serie Sen Atípicos
MAE	7.34	6.76
$RMSE$	9.50	8.81
$MAPE$	35.84 %	35.99 %
$MASE$	0.97	0.91

Cadro 4.34: Comparativa do rendemento preditivo do modelo *Prophet* na serie do clúster 3: Impacto da eliminación de *outliers*.

4.5. Comparación: ARIMA vs *Prophet*

Os resultados conseguidos ao longo do presente capítulo amosaron rendementos dispares dos modelos nas distintas series de demanda. O obxectivo desta sección é comparar as distintas metodoloxías e interpretar o seu rendemento en cada caso.

Modelo Global. Na serie de demanda global, a aproximación ARIMA$(0, 1, 1)$ axustada presenta mellores resultados. En particular, o modelo de Box-Jenkins erra en menos de 8 envíos semanais en termos absolutos, mentres que *Prophet* presenta un MAE = 16.19. Á vista das gráficas das predicións (Figuras 4.4 e 4.18), o modelo *Prophet* non é capaz de anticipar a caída estrutural da demanda que se produce durante o último trimestre da mostra de test.

Clúster 1. Na serie orixinal dos clientes de *retail*, transporte e electrónica a metodoloxía de Box-Jenkins presenta maior precisión. O MAE cometido molo ARIMA$(0, 1, 1)$ axustado reduce á metade o valor da métrica no *Prophet*. Observando a gráfica das predicións deste último modelo (Figura 4.20), o *Prophet* presenta nesgo positivo de predición, sobreestimando de forma sistemática o nivel de demanda. A causa reside na propia estrutura do modelo, que captura a elevada tendencia da serie na mostra de adestramento.

Clúster 2. No estudo da demanda dos sectores de paquetería e froita, unha vez máis a metodoloxía ARIMA amosa maior calidade preditiva. Así, o erro cometido en erros porcentuais resulta inferior ao 8 %, fronte ao 21 % que comete *Prophet*. Visualizando as gráficas das predicións (Figuras 4.12 e 4.22), o *Prophet* volve sobreestimar o nivel de demanda semanal, especialmente a partir do mes de maio da mostra de test.

Clúster 3. A serie de demanda dos clientes do clúster 3 (alimentación e industria) amosaba unha alta volatilidade durante o primeiro trimestre. Estes picos de demanda afectaron ao rendemento do ARIMA$(2, 1, 0)$ axustado, que non foi capaz de capturar a magnitude das oscilacións (Figura 4.17). Por conseguinte, a metodoloxía de Box-Jenkins infrapredeciu o volume de pedidos semanais durante o primeiro trimestre. En particular, o axuste non mellora a considerar o volume da última semana observada como predición, presentando un MASE = 1.42.

Pola contra, *Prophet* si era capaz de detectar os picos de demanda (Figura 4.24). Porén, ante a caída da serie a partir do segundo trimestre, o modelo tende á sobrepredición. Así e todo, a pesar dos malos resultados en termos porcentuais (MAPE = 35.84 %), *Prophet* consegue mellorar ao preditor naïve estacional. Este feito reflicte o gran cambio que sofre a serie de demanda no devandito clúster.

Impacto de *outliers*. En relación a análise de sensibilidade, os modelos amosan comportamentos certamente distintos. Mentres que o ARIMA apenas mellora a súa precisión tras eliminar os *outliers*, *Prophet* consegue reducir os erros de predición ao axustarse sobre a serie depurada. Especialmente salientables os resultados no clúster 1, onde a eliminación de *outliers* reduce á metade as métricas do erro.

En vista dos resultados obtidos, especialmente nas series con alta volatilidade ou presenza de *outliers*, faise evidente a necesidade de explorar modelos alternativos ou híbridos que combinen a robustez fronte a cambios abruptos con capacidade de capturar tendencias estruturais, co fin de mellorar a precisión das predicións no contexto complexo do transporte de mercadorías.

No seguinte capítulo profundizarase nas conclusións do proxecto, e exporanse posibles variantes de estudo que se poderían aplicar ao problema da previsión de demanda no sector do transporte.

Capítulo 5

Conclusións e liñas futuras

Neste último capítulo preséntanse as conclusións xerais derivadas da investigación, sintetizando os achados obtidos durante a modelización da demanda de transporte internacional. Así mesmo, propóñense novas vías de investigación e melloras metodolóxicas que poderían dar continuidade a este traballo.

5.1. Conclusións

O obxectivo principal deste Traballo Fin de Máster foi desenvolver un sistema de previsión de demanda fiable para optimizar a planificación loxística de Trucksters. Trala análise exploratoria, o preprocesado de datos e a comparativa entre metodoloxías clásicas (ARIMA) e baseadas en modelos aditivos (*Prophet*) sobre series segmentadas (clústeres), extráense as seguintes conclusións:

Eficacia da segmentación de clientes

A decisión estratéxica de dividir a carteira de clientes en clústeres baseados no comportamento (transporte/*retail*/electrónica, paquetería/froita e alimentación/industria) resultou fundamental. O estudo da serie global pode funcionar como unha primeira aproximación da carga prevista nas vindeiras semanas, pero a segmentación dos clientes resulta fundamental para optimizar planificación nas distintas rutas.

A análise confirmou que cada grupo de clientes posúe dinámicas estocásticas diferenciadas, así como diferente sensibilidade a eventos especiais e festivos (*peak season*).

Superioridade de ARIMA en series con cambio de tendencia

Para a serie global e os clústeres 1 e 2, a metodoloxía clásica (ARIMA) demostrou un rendemento superior, tanto en termos absolutos (MAE) coma relativos (MAPE). En efecto, observouse que os modelos *Prophet* tenden a sobreestimar sistematicamente a demanda cando a serie sofre unha contracción estrutural ao final da mostra. A inercia da tendencia aprendida durante a fase de adestramento (etapa de crecemento da empresa) impediu a *Prophet* adaptarse de forma correcta á estratexia de mercado adaptada por Trucksters durante o ano 2025. Pola contra, os modelos ARIMA, grazas á súa estrutura de diferenciación e memoria a curto prazo, axustáronse mellor a este novo escenario. No clúster 1, por exemplo, o ARIMA logrou reducir o erro absoluto medio á metade con respecto a *Prophet*, consolidándose como a ferramenta máis robusta para sectores estables.

Utilidade de *Prophet* en escenarios de alta volatilidade

O clúster 3 (alimentación e industria) representou o maior desafío de modelización debido á súa heterocedasticidade e picos abruptos. Neste escenario, o modelo ARIMA fracasou na súa tentativa de superar un preditor básico (MASE = 1.42), demostrando a rixidez dos modelos lineais ante series moi irregulares. Pola contra, *Prophet* logrou superar a barreira de utilidade preditiva (MASE < 1), grazas á súa flexibilidade para modelar cambios de punto de inflexión e á xestión da estacionalidade multiplicativa. Aínda que o seu erro porcentual segue sendo elevado (MAPE $> 35\,\%$), demostrou ser a alternativa máis viable cando o comportamento do cliente carece de inercia clara.

O impacto crítico dos atípicos

A análise de sensibilidade revelou que a xestión de *outliers* é determinante, pero o seu impacto varía segundo o modelo. Mentres que a metodoloxía de Box-Jenkins resultou máis ríxida ante a limpeza de datos, *Prophet* beneficiouse enormemente da depuración da serie, especialmente no clúster 1, onde a eliminación de atípicos reduciu o erro á metade. Isto suxire que *Prophet* é unha ferramenta potente cando se dispón dun coñecemento de dominio que permita limpar ou parametrizar correctamente as anomalías históricas.

Asimetría do risco na predición: Sobrestimación vs. Subestimación

Máis aló das métricas estatísticas puras (RMSE ou MAPE), a avaliación dos modelos debe considerar o impacto operativo dos erros nun contexto de negocio real. No sector do transporte internacional, as consecuencias de errar na predición non son simétricas:

- O risco da subestimación: Cando o modelo infrapredí a demanda (como ocorreu co ARIMA no clúster 3), a empresa enfróntase a unha rotura de servizo. A falta de capacidade para cubrir os envíos comprometidos pode derivar en penalizacións contractuais e, o máis crítico, nun dano irreparable á relación co cliente e á reputación de Trucksters.
- O custo da sobrestimación: Pola contra, a tendencia á sobrepredición observada fundamentalmente no *Prophet* (clústeres 1 e 2) implica un custo económico directo por recursos ociosos (camións ou condutores parados). Non obstante, este escenario garante o nivel de servizo e protexe a relación comercial.

Baixo esta óptica, aínda que os modelos ARIMA foron estatisticamente máis precisos en sectores estables, o nesgo positivo de *Prophet* podería considerarse "máis seguro" estratexicamente en escenarios de incerteza, ao priorizar a cobertura do servizo fronte á eficiencia de custos inmediata. Ademais deste feito, outra das vantaxes da metodoloxía consistía na construción de intervalos de confianza sen a necesidade de asuncións de normalidade nos residuos, aportando información adicional sobre a demanda prevista.

Limitacións pola lonxitude da serie histórica

Finalmente, cómpre sinalar unha limitación técnica relevante que condicionou o rendemento dos modelos baseados en compoñentes. Tal e como se indicaba ao iniciarse este estudo (Sección 1.2), a base de datos dispoñible abrangue o período 2022-2025, o que supón apenas tres anos completos de historia para o adestramento. Así, esta lonxitude resulta insuficiente para que algoritmos como *Prophet* desenvolvan todo o seu potencial. Para estimar con robustez compoñentes complexas, como a estacionalidade anual ou os cambios de tendencia a longo prazo, estes modelos benefícianse de series máis extensas que permitan distinguir patróns cíclicos recorrentes de anomalías puntuais. En consecuencia, a escaseza de datos históricos explica parte das dificultades atopadas para modelar a estacionalidade nos clústeres máis irregulares e reforza a idea de que os modelos mellorarán as súas prestacións a medida que a empresa acumule máis histórico de operacións.

En resumo, conclúese que non existe un "modelo único" óptimo para toda a operativa de Trucksters, nin en termos estatísticos nin de negocio. A solución ideal pasa por un enfoque híbrido e estratéxico: empregar modelos ARIMA para os fluxos de carga consolidados e estables (clústeres 1 e 2) onde se busca a máxima eficiencia, e reservar aproximacións tipo *Prophet* para os segmentos máis volátiles ou industriais (clúster 3), asumindo o seu nesgo positivo como unha marxe de seguridade operativa ante a incerteza.

5.2. Liñas futuras de estudo

Dada a complexidade inherente á predición de demanda no sector loxístico, ábrense diversas vías para profundar nesta investigación e mellorar a precisión dos modelos actuais.

En relación á complexidade dos axustes, unha das limitacións dos modelos empregados durante o traballo é que non consideran o estado de saturación do mercado europeo de transporte. Unha liña de mellora inmediata sería incorporar a información de datos externos en tempo real, especificamente o Barómetro de Transporte de Timocom (2025). Como se explicou na Sección 1.1, esta fonte proporciona o ratio diaria entre ofertas de carga e espazo de camións dispoñible en Europa. A incorporación desta métrica como regresor exóxeno permitiría ao modelo anticipar picos de demanda non estacionais e axustar as predicións baseándose na lei da oferta e a demanda a nivel continental, máis aló da carteira propia de Trucksters.

Por outra banda, aínda que este traballo abordou a previsión de demanda, a dinámica comercial depende intrinsecamente do prezo. Unha evolución natural do proxecto sería incorporar a información dos prezos de venda (cliente) e de compra (provedor). Non obstante, dado que o prezo futuro é unha incógnita necesaria para predicir a demanda futura proponse o uso de procesos autorregresivos vectoriales (VAR), que son unha xeralización dos procesos autorregresivos univariantes (AR). Esta aproximación permitiría predicir ambas variables de forma conxunta, capturando a elasticidade prezo-demanda e, o máis importante, permitindo á empresa optimizar non só o volume de envíos, senón a marxe comercial esperada (diferenza entre prezo de cliente e custo de provedor).

Por último, ante a alta volatilidade na serie do clúster 3, poderíase explorar o uso de modelos máis complexos como as Redes Neuronais, en especial as redes LSTMN (*Long-shot-term memory neutral networks*) (véxase Fernández Casal et al., 2024). Este tipo de metodoloxías amosa mellores resultados á hora de captar estruturas moi complexas, pagando o prezo de resultar modelos moito menos interpretables (caixas negras).

Apéndice A

Variables de interese

Este apéndice detalla a natureza funcional e estatística das variables contidas na base de datos orixinal empregada neste proxecto. Para facilitar a súa comprensión, as variables agrupáronse segundo a dimensión da operativa loxística que representan.

Identificación e Temporalidade

- **id_trucksters**: Identificador único e inequívoco asignado a cada pedido no sistema de xestión da empresa.
- **route_start_date**: Marca temporal que rexistra o inicio do envío (data e hora). Esta variable constitúe o eixo cronolóxico fundamental para a construción das series temporais deste traballo.
- **lane_zipcode_clean**: Variable sintética que identifica a ruta específica mediante a combinación de cliente, nodo de orixe e nodo de destino.

Xeografía e Operativa

- **direction**: Clasificación do fluxo de transporte en catro categorías:
  - *Exportación*: Envío con orixe en España e destino internacional.
  - *Importación*: Envío con orixe internacional e destino España.
  - *Cabotaxe nacional*: Traxectos realizados integramente dentro do territorio español.
  - *Cabotaxe internacional*: Envíos entre dous países estranxeiros.
- **origin_country / destination_country**: Países de carga e descarga, respectivamente.
- **origin_zipcode / destination_zipcode**: Códigos postais que delimitan con precisión os puntos xeográficos da operativa.
- **loaded_km**: Distancia percorrida polo vehículo transportando a mercadoría. É a métrica base para o cálculo da eficiencia e dos prezos unitarios.
- **empty_km**: Distancia percorrida sen carga (traxectos de posicionamento).

Dimensión Económica e Contractual

- **client_name_complete**: Razón social do cliente que solicita o servizo.
- **contract**: Variable categórica que define a modalidade comercial:
  - *Fix* (Primario): Clientes con contratos de longa duración e volumes de carga pactados. Adoitan presentar prezos máis estables e independentes das flutuacións inmediatas do mercado.
  - *Spot* (Secundario): Clientes puntuais sen compromiso de volume. Os prezos destes servizos son máis elevados e volátiles, xa que se axustan á dinámica de oferta e demanda no momento da contratación.
- **sector**: Sector industrial do cliente. A base de datos comprende 7 categorías: alimentación, electrónica, froita, industria, paquetería, transporte e *retail*. Os rexistros de transporte corresponden a outras empresas loxísticas que, ante picos de demanda ou falta de flota, subcontratan os seus servizos a Trucksters.
- **client_price**: Importe total en euros facturado ao cliente polo envío.
- **po_price_loaded / po_price_empty**: Custos operativos pagos ao provedor (colaborador externo) polos quilómetros percorridos con carga e sen ela, respectivamente.

Apéndice B

Calendario de festividades e eventos especiais

A actividade loxística internacional está estreitamente vinculada ao calendario laboral dos países onde opera a rede de Trucksters (principalmente España, Francia, Alemaña, Bélxica e Polonia). Este apéndice detalla a selección de festividades empregadas na modelización, a súa orixe e o procedemento técnico para a súa integración nos modelos.

Selección de datas

As festividades foron seleccionadas en base á súa relevancia para o transporte por estrada na Unión Europea, considerando tanto o peche de centros loxísticos como as restricións á circulación de vehículos pesados. As datas consideradas no presente proxecto foron aportadas polo responsable da empresa, sendo estas as festividades que se teñen en conta en Trucksters á hora de organizar a estrutura loxística.

As datas consideradas como "críticas" pola súa influencia na demanda son:

- **Ciclo de Ano Novo**: 1 de xaneiro e vésperas.
- **Semana Santa**: Datas variables segundo o calendario lunar (marzo/abril).
- **Día do Traballador**: 1 de maio.
- **Día da Vitoria en Europa**: 8 de maio (especialmente relevante en rutas con Francia e Alemaña).
- **Ascensión e Asunción**: Datas variables e 15 de agosto, respectivamente.
- **Tódolos Santos**: 1 de novembro e a súa véspera (31 de outubro).
- **Ponte da Constitución e Inmaculada**: 6 e 8 de decembro (con forte impacto en toda a rede loxística peninsular).
- **Ciclo de Nadal**: 25 de decembro e días posteriores.

Adaptación á escala semanal

Dado que o presente traballo emprega datos agrupados semanalmente, a inclusión destas festividades require unha adaptación técnica previa. Ao non poder asignar o impacto a un día exacto, as

festividades trasládanse ao día luns da semana na que acontecen (día de referencia da serie).

Cómpre sinalar que as festividades que caen sempre na mesma semana do ano tenden a ser absorbidas pola compoñente estacional do modelo. Polo tanto, o esforzo de modelización mediante variables exóxenas centrouse naquelas festividades móbiles (como a Semana Santa) ou eventos comerciais de grande impacto como o Black Friday.

Bibliografía

Airas D (2026) Repositorio de código do TFM: Modelos de predición de demanda para Trucksters. GitHub. https://github.com/DiegoAiras/Trucksters. Accedido 15 de xaneiro de 2026.

Aneiros Pérez G (2024) Series de Tiempo. Material docente do Máster en Técnicas Estadísticas, Universidade da Coruña.

Basavaraju K, Valilai OF (2025) Developing a demand planning strategy for joint forecasting and employing analytical tool in an empirical case study. Discover Applied Sciences 7:345.

Boin R, Gavin R, Rau P, Stoffels J (2020) Getting the price right in logistics. McKinsey & Company. https://www.mckinsey.com/industries/logistics/our-insights/getting-the-price-right-in-logistics. Accedido 15 de xaneiro de 2026.

Box GEP, Jenkins GM, Reinsel GC, Ljung GM (2015) Time Series Analysis: Forecasting and Control (5ª Edición). John Wiley & Sons, Hoboken.

Cao Abad R, Fernández Casal R (2022) Técnicas de Remuestreo. Material docente do Máster en Técnicas Estadísticas, Universidade da Coruña.

Cryer JD, Chan KS (2008) Time Series Analysis: With Applications in R (2ª Edición). Springer, Nova York.

Ertürk MA (2024) Comparison of Time Series Forecasting for Intelligent Transportation Systems in Digital Twins. Electrica 24:375-384.

Fernández Casal R, Costa Bouzas J, Oviedo de la Fuente M (2024) Métodos predictivos de aprendizaje estadístico. Universidade da Coruña, Servizo de Publicacións.

Liu FT, Ting KM, Zhou ZH (2012) Isolation-based anomaly detection. ACM Transactions on Knowledge Discovery from Data 6:1-39.

López Pouso R (2019) Series de Fourier y ecuaciones en derivadas parciales: una introducción con Maple y ejercicios resueltos. Universidade de Santiago de Compostela, Servizo de Publicacións e Intercambio Científico.

Maersk (2025) 5 peak logistics periods to prepare for in 2026. https://www.maersk.com/it-it/logistics-explained/freight-seasons/2025/11/04/peak-periods-in-logistics-2026. Accedido 15 de xaneiro de 2026.

Pateiro López B, Sánchez Sellero C (2024) Análisis Multivariante. Material docente do Máster en Técnicas Estadísticas, Universidade de Santiago de Compostela.

Python Software Foundation (2026) Python Language Reference, version 3.x. https://www.python.org/. Accedido 15 de xaneiro de 2026.

Revelo Chacón ED, Mafla Bolaños IG (2025) Modelo de predicción de demanda para la optimización de la cadena de suministros. Visión Empresarial 15:112-135.

Shumway RH, Stoffer DS (2017) Time Series Analysis and Its Applications: With R Examples (4ª Edición). Springer, Cham.

Taylor SJ, Letham B (2018) Forecasting at scale. The American Statistician 72:37-45.

Timocom (2025) Barómetro de transporte: Ratio de cargas e camións en Europa. https://www.timocom.es. Accedido 15 de xaneiro de 2026.

Trucksters (2026) Sobre nosotros: Sistema de relevos y tecnología. https://www.trucksters.io/es/sobre-nosotros/. Accedido 15 de xaneiro de 2026.

Wu T, Deng J, Chu C, Luo J (2024) ARIMA-based Freight Forecasting and Network Optimization in E-commerce Logistics. Highlights in Business, Economics and Management 33:296-303.