Facultad de Ciencias Económicas y Empresariales
ICADE

El futuro de los centros comerciales: Análisis Predictivo de las Afluencias en Centros Comerciales mediante técnicas de Machine Learning

Autor: Juan Jiménez Pumar
Director: Carlos Miguel Vallez Fernández

MADRID | Junio de 2026

Resumen

El presente trabajo analiza la evolución de las afluencias en centros comerciales españoles mediante técnicas de análisis de datos y modelos de Machine Learning. A partir de un conjunto de datos real con 480.334 registros horarios correspondientes a doce centros comerciales durante el período 2020-2025, se implementan cuatro modelos predictivos: Naïve lag-7 (modelo de referencia), SARIMA, Prophet y Random Forest. El modelo Random Forest, entrenado con veinticuatro variables explicativas, incluyendo retardos temporales, efectos calendario y variables meteorológicas procedentes de AEMET, obtiene los mejores resultados (R^2=0,862; RMSE=4.432), superando al modelo de referencia en un 24,2%. A partir del modelo validado, se generan proyecciones a diez años (2026–2035) bajo tres escenarios: base (≈91 millones de visitas anuales), optimista (+1,5% anual acumulado) y pesimista (-1,0% anual acumulado). Los resultados apuntan a una fase de madurez pospandémica del sector, con una incertidumbre que se amplía hasta el ±12,8% en el horizonte de 2035.

Palabras clave: centros comerciales, afluencias, predicción, series temporales, Machine Learning, Random Forest, SARIMA, Prophet.

Abstract

This work analyses the evolution of footfall in Spanish shopping centres through data analytics and Machine Learning techniques. Drawing on a real-world dataset of 480,334 hourly records covering twelve shopping centres over the 2020–2025 period, four predictive models are implemented: Naïve lag-7 (baseline), SARIMA, Prophet and Random Forest.

The Random Forest model, trained on twenty-four explanatory variables including temporal lags, calendar effects and meteorological data from AEMET, delivers the best performance (R^2=0.862; RMSE=4,432), outperforming the baseline by 24.2%. Building on the validated model, ten-year projections (2026–2035) are generated under three scenarios: base (≈91 million annual visits), optimistic (+1.5% compound annual growth) and pessimistic (-1.0% compound annual decline). The results point to a post-pandemic maturity phase of the sector, with uncertainty widening to ±12,8% by 2035. The study contributes a reproducible pipeline (publicly available on GitHub) and a set of strategic recommendations for shopping centre management.

Keywords: shopping centres, footfall, forecasting, time series, Machine Learning, Random Forest, SARIMA, Prophet.

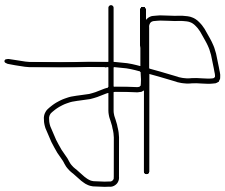
Declaración de Uso de Herramientas de Inteligencia Artificial Generativa

Por la presente, yo, Juan Jiménez Pumar, estudiante de la Facultad de Ciencias Económicas y Empresariales (ICADE) de la Universidad Pontificia Comillas, al presentar mi Trabajo Fin de Grado titulado "El futuro de los centros comerciales: Análisis Predictivo de las Afluencias en Centros Comerciales mediante técnicas de Machine Learning", declaro que he utilizado herramientas de Inteligencia Artificial Generativa únicamente en el contexto de las actividades descritas a continuación:

1. **Brainstorming de ideas de investigación:** Utilizado para idear y esbozar posibles áreas de investigación.
2. **Crítico:** Para encontrar contraargumentos a una tesis específica que pretendo defender.
3. **Referencias:** Usado conjuntamente con otras herramientas, como Science, para identificar referencias preliminares que luego he contrastado y validado.
4. **Interpretador de código:** Para realizar análisis de datos preliminares.
5. **Constructor de plantillas:** Para diseñar formatos específicos para secciones del trabajo.
6. **Corrector de estilo literario y de lenguaje:** Para mejorar la calidad lingüística y estilística del texto.
7. **Generador previo de diagramas de flujo y contenido:** Para esbozar diagramas iniciales.
8. **Sintetizador y divulgador de libros complicados:** Para resumir y comprender literatura compleja.
9. **Revisor:** Para recibir sugerencias sobre cómo mejorar y perfeccionar el trabajo con diferentes niveles de exigencia.
10. **Traductor:** Para traducir textos de un lenguaje a otro.

Afirmo que toda la información y contenido presentados en este trabajo son producto de mi investigación y esfuerzo individual, excepto donde se indique lo contrario y se hayan dado los créditos correspondientes. Asimismo, soy plenamente consciente de las implicaciones de presentar este trabajo y de las normas de integridad académica de la Universidad Pontificia Comillas.

Fecha: Junio de 2026

Firmado: 

Índice

1. Introducción

1.1. Contexto y justificación del estudio

En los últimos años, los centros comerciales han experimentado una profunda transformación como consecuencia de los cambios en los hábitos de consumo del consumidor, el auge del comercio electrónico y la creciente digitalización del sector del retail (Verhoef, Kannan & Inman, 2015). La consolidación de las estrategias omnicanal y la integración de las plataformas digitales en el proceso de decisión de compra han modificado la relación entre el consumidor y los comercios físicos. Ante esta nueva realidad, la afluencia de visitantes (comúnmente denominada *footfall* en la literatura sectorial) se ha convertido en uno de los principales indicadores para evaluar el rendimiento y la viabilidad de estos activos (Fildes et al., 2009; Retlife, 2024), al estar estrechamente relacionada con el volumen potencial de ventas, la capacidad de atracción y la retención de operadores.

Tras el impacto provocado por la pandemia del COVID-19, el sector de los centros comerciales en España inició a comienzos de 2021 una fase de recuperación progresiva, con un crecimiento tanto en ventas como en afluencias. No obstante, esta recuperación no ha sido homogénea: se observan diferentes patrones de comportamiento entre centros comerciales con diferentes ubicaciones geográficas o *tenant mix*, lo que sugiere un cambio estructural en las dinámicas de consumo. Informes recientes del sector (CBRE, 2025; AECC, 2024) confirman que, si bien las ventas y afluencias han recuperado e incluso superado los niveles prepandémicos en muchos activos, persiste una elevada heterogeneidad entre formatos y ubicaciones. Este fenómeno pone de manifiesto la necesidad de disponer de herramientas analíticas que permitan comprender con mayor precisión los patrones de comportamiento de los visitantes y anticipar su evolución futura en distintos escenarios (Makridakis, Spiliotis & Assimakopoulos, 2020; Petropoulos, Makridakis & Assimakopoulos, 2022).

Paralelamente, la literatura más reciente subraya el papel creciente del Machine Learning en problemas de forecasting de demanda en retail, especialmente cuando intervienen múltiples variables explicativas o relaciones no lineales (Giannopoulos, Dasaklis, Tsantilis & Patsakis, 2025). Los modelos basados en árboles, así como aproximaciones híbridas que combinan series temporales clásicas con aprendizaje automático, han demostrado en los últimos años una mejora consistente sobre los métodos tradicionales en contextos de alta estacionalidad y dependencia de factores exógenos.

En el caso específico de los centros comerciales, la capacidad de anticipar la evolución de las afluencias resulta estratégica por diversas razones. En primer lugar, facilita la planificación operativa, así como la asignación de recursos humanos y logísticos. En segundo lugar, permite optimizar campañas de marketing y promociones en función de las expectativas de afluencia. En tercer lugar, ayuda a la evaluación de proyectos de inversión, al proporcionar estimaciones más fundamentadas sobre flujos futuros de visitantes, lo que resulta especialmente relevante en un contexto en el que los ciclos de inversión y rotación de operadores requieren visibilidad a medio y largo plazo (CBRE, 2025).

El presente trabajo de fin de grado se propone abordar el estudio de las afluencias en centros comerciales a través de la aplicación de técnicas de análisis de datos al sector inmobiliario comercial. A partir de datos reales anonimizados de afluencias, se analiza su comportamiento histórico y se desarrollan modelos capaces de predecir su evolución en períodos temporales de medio y largo plazo. La combinación de técnicas estadísticas tradicionales y metodologías de Machine Learning permite evaluar su capacidad predictiva y su utilidad práctica en la gestión

de estos activos (Hastie, Tibshirani & Friedman, 2009). Con el fin de garantizar la reproducibilidad del análisis (principio cada vez más exigido en proyectos de data science), el código y los notebooks asociados al trabajo se publican en un repositorio público de GitHub (https://github.com/juanjpumar-collab/TFG-CC).

1.2. Objetivos generales y específicos

El objetivo general del presente trabajo consiste en analizar y predecir la evolución de las afluencias en centros comerciales mediante la aplicación de técnicas de análisis de datos y modelos predictivos, evaluando su capacidad explicativa y su utilidad práctica en la gestión estratégica de dichos activos.

A partir de este objetivo general se plantean los siguientes objetivos específicos:

- Analizar el comportamiento histórico de las afluencias con el fin de identificar patrones de tendencia, estacionalidad y posibles cambios estructurales.
- Garantizar un proceso de anonimización riguroso que preserve la estructura temporal y relativa de los datos sin comprometer la confidencialidad.
- Implementar modelos clásicos de predicción de series temporales, particularmente modelos ARIMA.
- Desarrollar modelos de Machine Learning supervisado orientados a la predicción de una variable continua.
- Comparar el rendimiento de los modelos mediante métricas cuantitativas de error.
- Interpretar los resultados desde una perspectiva estratégica aplicable a la gestión de centros comerciales.

Estos objetivos reflejan una combinación de rigor metodológico y aplicabilidad práctica, de

modo que el análisis no se limite a obtener resultados estadísticos, sino que pueda aportar valor real en la toma de decisiones.

1.3. Empresa que aporta los datos para el trabajo

El análisis realizado en este trabajo se basa en datos reales proporcionados por una empresa del sector inmobiliario comercial. Dicha empresa gestiona centros comerciales y dispone de registros históricos de afluencias estructurados temporalmente.

Con el fin de proteger la confidencialidad de la información y garantizar el cumplimiento de criterios éticos, los datos han sido sometidos a un proceso de anonimización previo a su utilización. Este proceso ha supuesto la eliminación de identificadores directos y la aplicación de transformaciones que preservan la estructura temporal y las relaciones entre las diferentes observaciones, evitando así la posibilidad de reconstruir los valores originales.

El uso de datos reales permite analizar el comportamiento de los modelos en contexto más cercano a la práctica profesional. No obstante, la anonimización garantiza que el análisis se realice sin comprometer información estratégica.

1.4. Metodología general

La metodología empleada se basa en el análisis cuantitativo de datos históricos y en la aplicación de distintos modelos predictivos, que posteriormente se comparan entre sí mediante métricas estandarizadas de error.

En una primera fase, se realiza la preparación y depuración de los datos. Esta etapa incluye la revisión de la consistencia temporal, la identificación de posibles *outliers* (valores atípicos), el tratamiento de valores cero asociados a fallos de sensores y la estructuración adecuada de la

información para su correcta modelización.

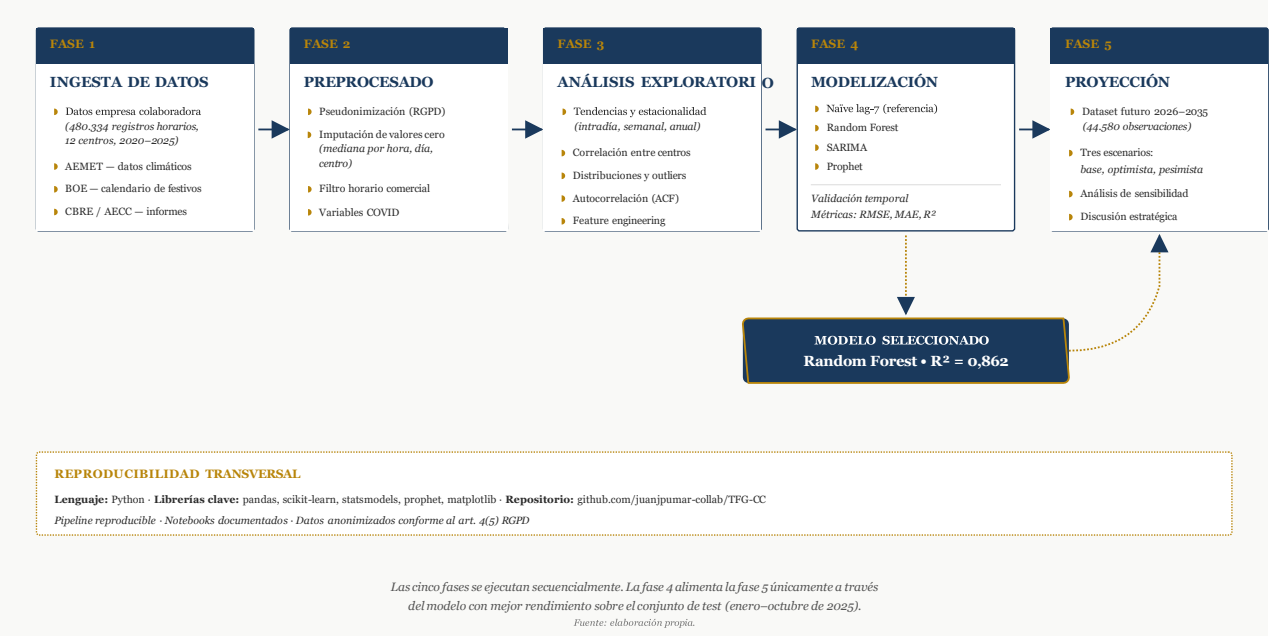
En una segunda fase, se lleva a cabo un análisis exploratorio de datos con el objetivo de comprender la dinámica de las afluencias, identificar patrones estacionales y analizar posibles tendencias.

Posteriormente, se implementan modelos de predicción de series temporales (SARIMA y Prophet) y modelos de Machine Learning supervisado (Random Forest), junto con un modelo de referencia (Naïve lag-7). Cada modelo es entrenado con datos históricos y evaluado mediante métricas de error (RMSE, MAE y R^2) que permiten comparar su capacidad predictiva.

A partir del modelo seleccionado se construyen proyecciones a diez años bajo tres escenarios (base, optimista y pesimista) y se realiza un análisis de sensibilidad sobre las variables externas. Por último, se discuten las implicaciones de los resultados para la gestión estratégica de los centros comerciales.

Todo el flujo metodológico (ingesta, limpieza, *feature engineering*, entrenamiento, evaluación y proyección) se ha implementado en Python y está disponible en un repositorio público de GitHub (https://github.com/juanjpumar-collab/TFG-CC), lo que garantiza la reproducibilidad del análisis. La Figura 0 sintetiza este flujo.

Figura 0. Flujo metodológico de la aplicación



Las cinco fases se ejecutan secuencialmente. La fase 4 alimenta la fase 5 únicamente a través del modelo con mejor rendimiento sobre el conjunto de test (enero–octubre de 2025).

Fuente: elaboración propia.

2. Marco Teórico

2.1. Conceptos clave de Data Analytics y del Machine Learning

El desarrollo de técnicas de análisis de datos ha transformado profundamente la manera en que las organizaciones abordan la toma de decisiones. El concepto de Data Analytics engloba un conjunto de métodos estadísticos, computacionales y matemáticos orientados a la extracción de información útil a partir de grandes volúmenes de datos. Desde una perspectiva conceptual, puede distinguirse entre análisis descriptivo, diagnóstico, predictivo y prescriptivo. Mientras que el análisis descriptivo se centra en explicar qué ha ocurrido y el diagnóstico en por qué ha ocurrido, el análisis predictivo busca estimar qué ocurrirá en el futuro, apoyándose en patrones identificados en datos históricos.

En el ámbito empresarial, el análisis predictivo ha adquirido especial relevancia como herramienta de apoyo a la decisión. Shmueli y Koppius (2011) distinguen entre modelos explicativos, cuyo objetivo principal es comprender relaciones causales entre variables, y modelos predictivos, orientados a maximizar la precisión en la estimación de valores futuros. En el contexto del presente trabajo, el interés se centra fundamentalmente en el enfoque predictivo, sin que ello excluya la posibilidad de interpretar los resultados obtenidos.

El Machine Learning es una rama de la inteligencia artificial que se caracteriza por utilizar algoritmos capaces de identificar patrones en los datos y realizar predicciones sin que sea necesario programar explícitamente todas las reglas. Hastie, Tibshirani y Friedman (2009) señalan que el objetivo principal del aprendizaje supervisado consiste en estimar una función f(X) que relacione un conjunto de variables explicativas X con una variable dependiente Y, minimizando una función de pérdida asociada al error de predicción.

En problemas de regresión, como el análisis de afluencias, donde la variable objetivo es

continua, el interés reside en estimar con la mayor precisión posible el valor esperado de la variable futura. El desempeño del modelo se evalúa generalmente mediante métricas como el error cuadrático medio (MSE) o su raíz (RMSE), que penalizan las desviaciones entre valores observados y predichos, así como el coeficiente de determinación R^2 y el error absoluto medio (MAE).

Dentro del Machine Learning supervisado, los modelos basados en árboles de decisión han demostrado un buen equilibrio entre capacidad predictiva e interpretabilidad. El algoritmo Random Forest, propuesto por Breiman (2001), constituye una extensión de los árboles individuales mediante la construcción de múltiples árboles sobre subconjuntos aleatorios de datos y variables. Este enfoque reduce la varianza del modelo y mejora su capacidad de generalización. La agregación de predicciones individuales permite mitigar el riesgo de sobreajuste, especialmente en contextos donde las relaciones entre variables pueden ser complejas o no lineales.

En el ámbito del forecasting empresarial, evidencia reciente muestra que los modelos de Machine Learning pueden superar a los métodos tradicionales cuando la estructura de los datos es no lineal o cuando intervienen múltiples variables explicativas (Makridakis, Spiliotis & Assimakopoulos, 2020; Petropoulos, Makridakis & Assimakopoulos, 2022). En el caso particular del retail, Giannopoulos et al. (2025) destacan que los modelos basados en árboles y los enfoques híbridos están desplazando progresivamente a los métodos estadísticos clásicos en problemas con estacionalidad múltiple y efectos calendario. No obstante, ningún modelo es universalmente superior: el rendimiento depende de las características del problema, de la calidad del *feature engineering* y del propio diseño experimental.

Conviene señalar, además, que la aportación de las variables exógenas a los modelos predictivos no siempre se manifiesta en una importancia porcentual elevada. En muchas aplicaciones de retail, los retardos temporales concentran la mayor parte de la importancia,

mientras que las variables externas (meteorología, calendario, eventos) actúan como factores de ajuste fino que, pese a su menor peso relativo, mejoran de forma medible las métricas de error (Fildes, Goodwin & Nikolopoulos, 2019). Este matiz es relevante de cara a la interpretación de resultados del presente trabajo.

Por todo ello, en este trabajo se comparan modelos clásicos de series temporales con modelos de Machine Learning, evaluando cuál ofrece mejores resultados en la predicción de afluencias.

2.2. Predicción de series temporales

Las series temporales representan secuencias de observaciones ordenadas cronológicamente, en las que el orden temporal desempeña un papel esencial en la estructura de dependencia entre datos. La teoría clásica de series temporales distingue tres componentes principales: tendencia, estacionalidad y componente irregular. La tendencia se refiere a la evolución a largo plazo de la variable; la estacionalidad recoge patrones recurrentes asociados a intervalos temporales específicos; y el componente irregular representa fluctuaciones no sistemáticas.

La metodología Box-Jenkins (Box & Jenkins, 2015; Box, Jenkins, Reinsel & Ljung, 2015) constituye uno de los enfoques fundamentales para la modelización de series temporales. Esta metodología se basa en la identificación, estimación y validación de modelos ARIMA (*AutoRegressive Integrated Moving Average*). Los modelos ARIMA combinan componentes autorregresivos (AR), de medias móviles (MA) y procesos de diferenciación (I) para lograr la estacionariedad de la serie. Su extensión estacional, SARIMA, incorpora además componentes específicos para capturar patrones periódicos, lo que la hace particularmente adecuada para series con estacionalidad semanal o anual.

Un modelo ARIMA(p,d,q) puede interpretarse como una combinación de p términos autorregresivos, d diferenciaciones y q términos de medias móviles. Su aplicación requiere la identificación adecuada de los parámetros, habitualmente mediante el análisis de las funciones de autocorrelación (ACF) y autocorrelación parcial (PACF) o procedimientos automatizados como *auto.arima*.

Hyndman y Athanasopoulos (2021) destacan que los modelos ARIMA resultan particularmente eficaces cuando la serie presenta patrones lineales bien definidos y dependencias temporales estables, y cuando sus parámetros se ajustan específicamente a la dinámica de cada serie. No obstante, su rendimiento puede verse limitado en tres escenarios: cuando la estructura de la serie incluye relaciones no lineales, cuando concurren cambios estructurales abruptos (como los inducidos por la pandemia del COVID-19) y cuando se utilizan parametrizaciones genéricas no optimizadas para la serie específica. Esta última limitación tiene implicaciones directas para el diseño experimental del presente trabajo, como se discutirá en el capítulo 4.

Como alternativa moderna a los modelos ARIMA, Prophet (Taylor & Letham, 2018) ha ganado popularidad por su capacidad para descomponer series con múltiples estacionalidades e incorporar de forma natural festivos y regresores externos, manteniendo a la vez una elevada interpretabilidad.

En el caso de las afluencias a centros comerciales, la presencia de múltiples estacionalidades (horaria, semanal y anual) plantea retos adicionales para la modelización. La correcta especificación de los componentes estacionales resulta crucial para evitar sesgos en la predicción.

2.3. Variables que pueden influir en las afluencias de los Centros Comerciales

El comportamiento de las afluencias en los centros comerciales está influenciado por una combinación de factores internos y externos. Desde una perspectiva temporal, se pueden identificar patrones recurrentes asociados a días de la semana, periodos vacacionales y eventos especiales. Estos efectos calendario pueden generar variaciones sistemáticas en el tráfico de visitantes.

En el caso de los centros comerciales, factores como la climatología pueden influir en la tendencia a visitar espacios físicos. Diferentes estudios han mostrado que variables meteorológicas, como temperatura o precipitaciones, pueden afectar la movilidad y a los comportamientos de compra.

Asimismo, variables macroeconómicas como el nivel de renta disponible o los índices de confianza del consumidor pueden influir en las decisiones de consumo y, por tanto, en la afluencia a centros comerciales. Aunque estas variables no siempre se integran directamente en modelos operativos, su influencia contextual ha de tomarse en consideración para el análisis interpretativo.

La inclusión de variables externas en modelos de predicción permite capturar parte de la variabilidad no explicada por la estructura temporal. No obstante, su incorporación debe realizarse con precaución para evitar sobreajustes y problemas de multicolinealidad.

En definitiva, la modelización de afluencias requiere un enfoque 360 que considere tanto la dependencia temporal interna como la posible influencia de factores contextuales externos.

3. Preparación y descripción de los datos

3.1. Fuentes de información

El conjunto de datos principal utilizado en este trabajo ha sido proporcionado por una empresa del sector inmobiliario comercial en España, que gestiona una cartera diversificada de centros comerciales distribuidos por distintas comunidades autónomas. Los datos están compuestos por registros históricos de afluencias, medidas mediante sensores de conteo de personas situados en los accesos de cada centro, una práctica consolidada en el sector para la monitorización del tráfico de visitantes (Fildes, Goodwin, Nikolopoulos & Lawrence, 2009).

La base de datos original contiene 480.334 registros correspondientes a 12 centros comerciales, con una granularidad horaria. El período temporal abarca desde el 1 de enero de 2020 hasta el 29 de octubre de 2025, lo que supone un horizonte de casi seis años que incluye tanto el impacto de la pandemia del COVID-19 como la fase de recuperación posterior.

Cada registro del conjunto de datos se compone de tres variables:

- DES_CC: identificador del centro comercial.
- FECHAHORA: marca temporal con precisión horaria (formato datetime).
- PERSONAS_ENTRAN: número de personas que acceden al centro en la franja horaria correspondiente.

Los 12 centros comerciales incluidos en el estudio presentan características heterogéneas en cuanto a tamaño, ubicación geográfica y perfil comercial (tenant mix), lo que nos permite un análisis comparado de las dinámicas de afluencia. El rango horario de los registros varía entre centros: mientras que algunos presentan datos exclusivamente en horario comercial (9:00–23:00), otros disponen de registros durante las 24 horas del día, lo que refleja diferencias en

los sistemas de medición o en los horarios de apertura.

Tabla 1. Descripción del conjunto de datos por centro comercial

Centro	Registros	Horas/día	Rango horario	Inicio datos
Alpha	31.890	15	9–23	Ene 2020
Beta	44.642	21	0–23	Ene 2020
Delta	38.319	18	0–23	Ene 2020
Epsilon	36.193	17	7–23	Ene 2020
Gamma	38.320	18	0–23	Ene 2020
Ipsilon	33.768	24	0–23	Dic 2021
Kappa	48.540	23	0–23	Ene 2020
Lambda	29.865	15	9–23	May 2020
Omega	46.835	22	0–23	Ene 2020
Sigma	38.319	18	0–23	Ene 2020
Tau	42.547	20	0–23	Ene 2020

Theta	51.096	24	0–23	Ene 2020

Adicionalmente, se contempla integrar variables externas que nos permitan enriquecer el poder explicativo y predictivo de los modelos. Las fuentes previstas para estas variables son:

- Datos meteorológicos: Agencia Estatal de Meteorología (AEMET), a través de su API de datos abiertos (AEMET OpenData). Se prevé incorporar variables como temperatura media, precipitación y horas de sol.
- Calendario de festivos: Boletín Oficial del Estado (BOE), que publica anualmente los festivos nacionales y autonómicos.
- Indicadores macroeconómicos: Instituto Nacional de Estadística (INE), para series como el Índice de Confianza del Consumidor o el Índice de Precios al Consumo.
- Informes sectoriales: CBRE Retail Index (CBRE, 2025) y datos de la Asociación Española de Centros y Parques Comerciales (AECC, 2024).

3.2. Anonimización y limpieza

Proceso de anonimización

Con el fin de proteger la confidencialidad de la información proporcionada por la empresa colaboradora, se ha aplicado un proceso de pseudonimización sobre el conjunto de datos, conforme a lo establecido en el artículo 4(5) del Reglamento General de Protección de Datos (RGPD, Reglamento UE 2016/679). Según dicho reglamento, la pseudonimización consiste en el tratamiento de los datos personales de tal forma que ya no puedan atribuirse a un interesado sin utilizar información adicional, siempre que dicha información se mantenga por separado y esté sujeta a medidas técnicas y organizativas que garanticen su seguridad.

En la práctica, el proceso se ha instrumentado mediante la sustitución de los nombres reales de los centros comerciales por identificadores basados en letras del alfabeto griego (Alpha, Beta, Delta, Epsilon, Gamma, Ipsilon, Kappa, Lambda, Omega, Sigma, Tau y Theta). La tabla de correspondencia entre nombres originales y pseudonimizados se ha almacenado de forma separada y no se incluye en los datos de trabajo, haciendo así imposible la identificación de los centros.

Esto nos permite preservar la estructura temporal y numérica de los datos, sin alterar los valores de afluencia ni las relaciones entre observaciones, lo que resulta clave a la hora de dar validez a los modelos predictivos. Además, permite realizar análisis comparativos entre centros sin necesidad de conocer su identidad real. El proceso de transformación se ha realizado mediante Power Query en Microsoft Excel.

Valores ausentes y valores cero

El análisis del conjunto de datos revela la existencia de 16.066 registros con valor cero en la variable PERSONAS_ENTRAN, lo que representa un 3,34% del total de observaciones. Estos registros, más que reflejar una afluencia nula real, se interpretan como ausencias de medición o períodos en los que el sistema de conteo no estuvo operativo, dado que resulta altamente improbable que un centro comercial no reciba absolutamente ningún visitante durante una hora completa de actividad comercial.

La distribución de estos valores cero no es homogénea entre centros comerciales:

Tabla 2. Distribución de valores cero por centro comercial

Centro	Valores cero	% del total CC
Alpha	3	0,01%

Beta	4.601	10,31%
Delta	1.308	3,41%
Epsilon	936	2,59%
Gamma	118	0,31%
Ipsilon	329	0,97%
Kappa	3.353	6,91%
Lambda	33	0,11%
Omega	447	0,95%
Sigma	181	0,47%
Tau	1.886	4,43%
Theta	2.871	5,62%

Se observa que los centros Beta, Kappa y Theta concentran la mayor proporción de valores cero (10,31%, 6,91% y 5,62% respectivamente). El análisis temporal de estos ceros muestra una mayor concentración en 2020 (4.110 ceros), particularmente en los meses de marzo a mayo, coincidentes con el confinamiento. No obstante, la persistencia de valores cero en años posteriores (3.131 en 2021, 1.814 en 2022, 1.861 en 2023, 2.225 en 2024 y 2.925 hasta octubre de 2025) nos confirma que el fenómeno no es exclusivamente pandémico, sino que responde a problemas recurrentes en los sensores de determinados centros.

Caso especial: centro Ipsilon

El centro Ipsilon presenta una particularidad relevante: no dispone de ningún registro con afluencia positiva hasta el 23 de diciembre de 2021. El año 2020 completo registra una afluencia total de cero, lo que indica que este centro no estaba incorporado al sistema de medición durante ese período. Sus series son significativamente más cortas (1.407 días frente a los 2.129 del período completo), aspecto que debe tenerse en cuenta en la modelización.

Heterogeneidad en los rangos horarios

Un hallazgo relevante durante la fase de limpieza es la heterogeneidad en el número de franjas horarias registradas por centro. Mientras que centros como Alpha y Lambda presentan registros exclusivamente en horario comercial (15 horas, de 9:00 a 23:00), otros como Theta e Ipsilon registran datos las 24 horas. El análisis de los registros fuera de horario comercial muestra valores medios muy reducidos: por ejemplo, Sigma registra una media de
42 personas/hora fuera de horario frente a 2.681 dentro (ratio del 1,6%). Para garantizar la comparabilidad entre centros, se aplicará un filtro a las franjas con actividad comercial significativa.

Consistencia temporal

Se ha verificado la ausencia total de registros duplicados en el conjunto de datos: no existe ningún par (DES_CC, FECHAHORA) repetido entre las 480.334 observaciones. La mayoría de los centros presentan registros para los 2.129 días que comprende el período de estudio, con las excepciones de Ipsilon (1.407 días), Lambda (1.991 días) y Alpha (2.126 días).

Estadísticas descriptivas básicas

Tras la revisión de integridad, se presentan las estadísticas descriptivas de la afluencia horaria por centro comercial, calculadas sobre registros con valor positivo:

Tabla 3. Estadísticas descriptivas de la afluencia horaria por centro (valores > 0)

Centro	Media	Desv. típ.	Mín.	Máx.
Alpha	2.183	1.169	1	6.667
Beta	909	885	1	5.864
Delta	971	953	1	7.146
Epsilon	1.109	799	1	4.820
Gamma	775	501	1	2.998
Ipsilon	643	649	1	4.105
Kappa	365	435	0	2.696
Lambda	2.127	1.486	1	9.325
Omega	601	752	1	6.227
Sigma	1.848	1.925	0	11.802
Tau	1.109	1.059	1	7.767
Theta	709	828	1	5.652

Se observan diferencias sustanciales entre centros. Alpha, Lambda y Sigma presentan las mayores afluencias medias horarias (superiores a 1.800 personas/hora), lo que nos sugiere una

mayor superficie comercial o una ubicación con mayor capacidad de atracción. En el extremo opuesto, Kappa registra la media más baja (365). La elevada desviación típica, frecuentemente del mismo orden de magnitud que la media, refleja la alta variabilidad horaria característica de este tipo de series temporales (Hyndman & Athanasopoulos, 2021).

Evolución agregada de las afluencias

Tabla 4. Afluencia total anual agregada (12 centros comerciales)

Año	Afluencia total	Registros	Observación
2020	57.819.803	74.966	Impacto COVID-19
2021	75.713.610	77.092	Recuperación progresiva
2022	93.253.837	85.773	Normalización
2023	96.576.098	85.774	Consolidación
2024	97.443.048	85.952	Estabilización
2025*	79.885.318	70.777	Datos hasta octubre

La serie agregada muestra con nitidez el impacto del COVID-19 en 2020 (57,8 millones). A partir de 2021 se inicia una recuperación progresiva que alcanza 93,3 millones en 2022, lo que supone un incremento del 61,3% respecto a 2020. Los años 2023 y 2024 muestran una estabilización en torno a los 96–97 millones, lo que nos sugiere que el sector ha alcanzado un nuevo nivel de equilibrio pospandémico. Este patrón es consistente con los datos publicados por la Asociación

Española de Centros y Parques Comerciales (AECC, 2024), que reportó crecimientos de hasta un 6% en ventas y afluencias durante el primer semestre de 2024.

Tratamiento del período COVID-19

El año 2020 presenta características atípicas derivadas del confinamiento. Se han creado dos variables indicadoras: COVID_confinamiento (14 marzo – 21 junio 2020, con 20.074 registros) y COVID_restricciones (22 junio 2020 – 9 mayo 2021, con 67.591 registros). En el presente trabajo se opta por incluir el período completo en el análisis exploratorio, incorporando dichas variables indicadoras, y se evaluará la sensibilidad de los modelos predictivos a su inclusión o exclusión

Estrategia de imputación

Para los registros con valor cero no asociados al período de confinamiento, se aplica una estrategia de imputación basada en la estructura temporal de los datos. Dado que estos ceros no responden a una ausencia real de afluencia sino a fallos en los sensores de conteo, se clasifican como datos ausentes de forma no aleatoria y requieren un tratamiento específico.

La técnica aplicada consiste en sustituir cada valor cero por la mediana de los registros positivos con idénticas condiciones: misma hora del día, mismo día de la semana y mismo centro comercial. Este enfoque aprovecha la doble estacionalidad intradía e intrasemanal propia de las series de afluencias y ofrece robustez frente a valores extremos. Siguiendo este procedimiento, se han imputado 13.605 valores exitosamente, quedando únicamente 2.461 registros con valor cero que sí corresponden a ausencia real de afluencia: el período de confinamiento estricto y el centro Ipsilon antes de su incorporación al sistema de medición.

3.3. Análisis Exploratorio

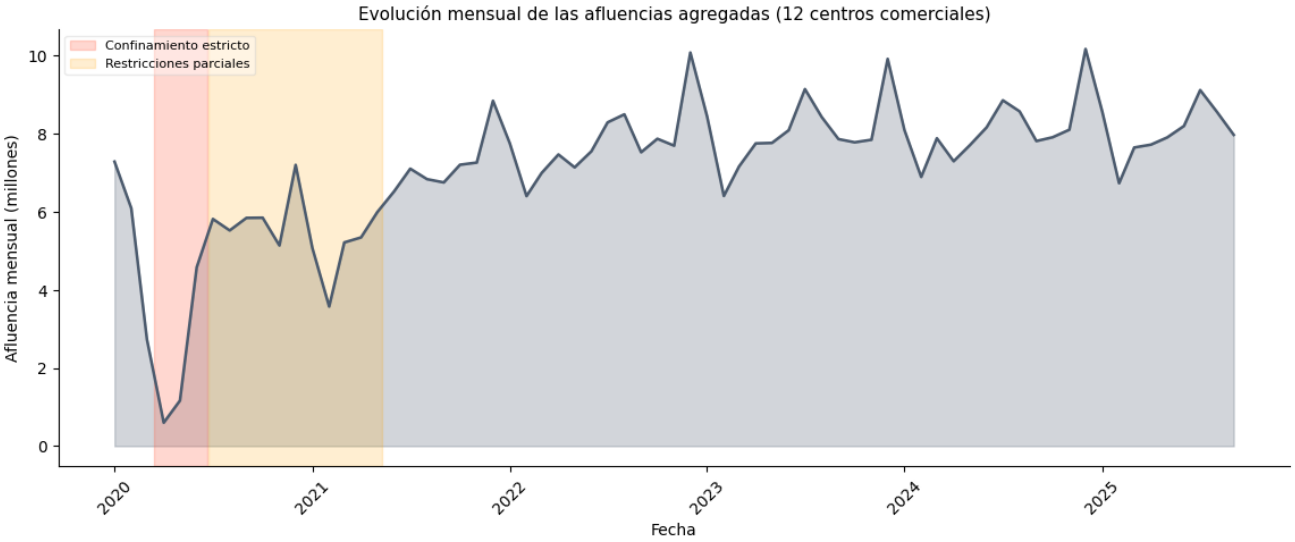
El análisis exploratorio de datos tiene como objetivo identificar los patrones fundamentales

que caracterizan la dinámica de las afluencias, así como detectar anomalías, relaciones entre variables y estructuras de dependencia temporal que resulten relevantes para la fase de modelización. Siguiendo las recomendaciones de Hyndman y Athanasopoulos (2021), el análisis se estructura en torno a tres dimensiones: tendencia, estacionalidad y relaciones entre centros. El diseño de las visualizaciones empleadas en este capítulo (gráficos de series temporales, diagramas de caja, mapas de calor y matrices de correlación) sigue los principios de comunicación gráfica propuestos por Wilke (2019) y Healy (2018), orientados a representar fielmente la estructura de los datos y facilitar su interpretación

Evolución temporal de las afluencias

La Figura 1 presenta la evolución mensual de las afluencias agregadas para los 12 centros comerciales. Se observan con claridad tres fases diferenciadas: la caída abrupta asociada al confinamiento de marzo–junio de 2020, un período de restricciones parciales con recuperación gradual, y la posterior normalización a partir de mediados de 2021.

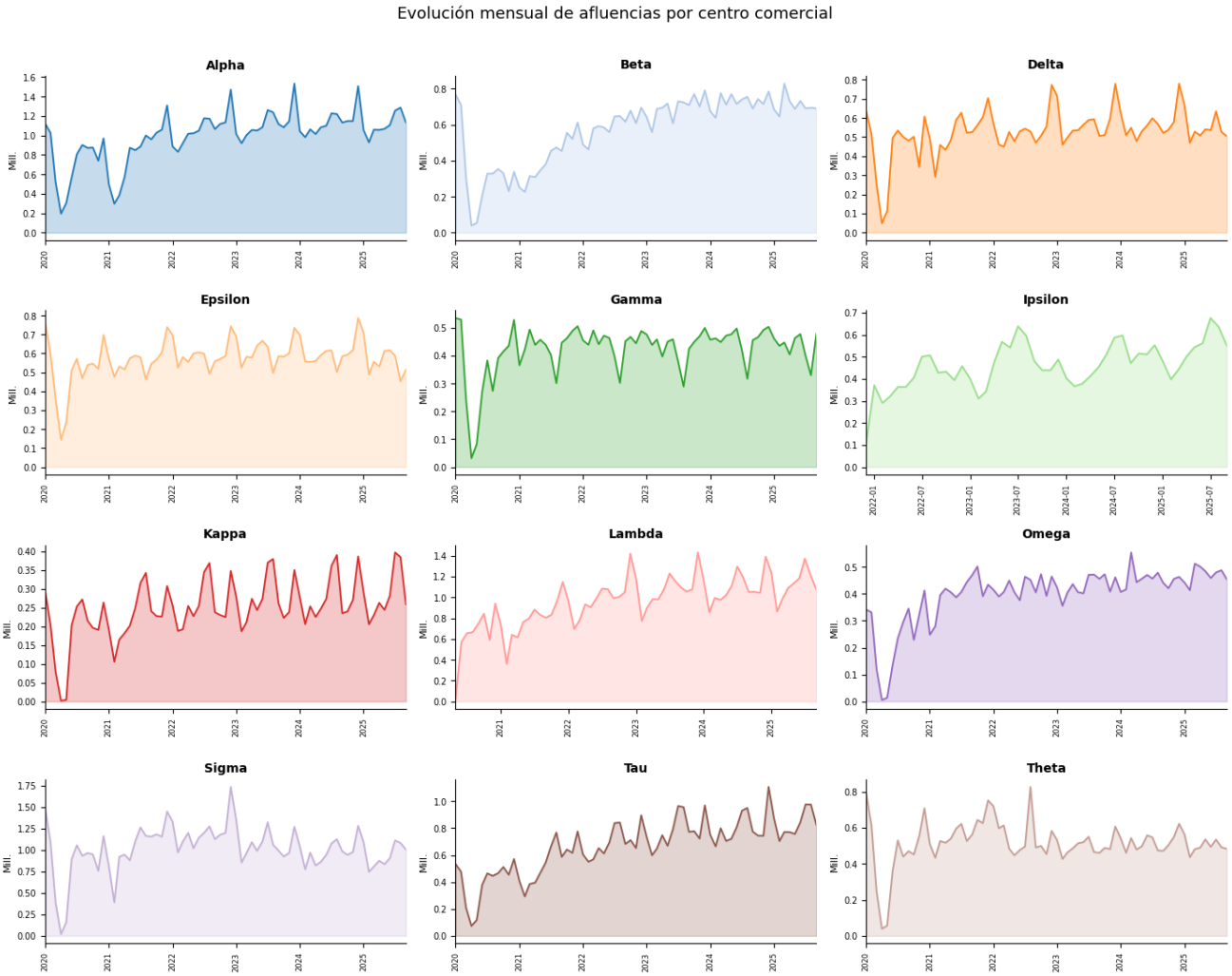
Figura 1. Evolución mensual de las afluencias agregadas (12 centros comerciales)



La desagregación por centro comercial (Figura 2) nos revela que la recuperación no ha sido homogénea. Centros como Lambda y Tau muestran una trayectoria de crecimiento sostenido que supera los niveles prepandémicos, mientras que otros como Theta y Sigma presentan

señales de estancamiento o retroceso a partir de 2023. Esta heterogeneidad es consistente con las diferencias en ubicación geográfica, perfil de tenant mix y grado de competencia local que caracterizan al sector (Ministerio de Agricultura, Alimentación y Medio Ambiente, 2013).

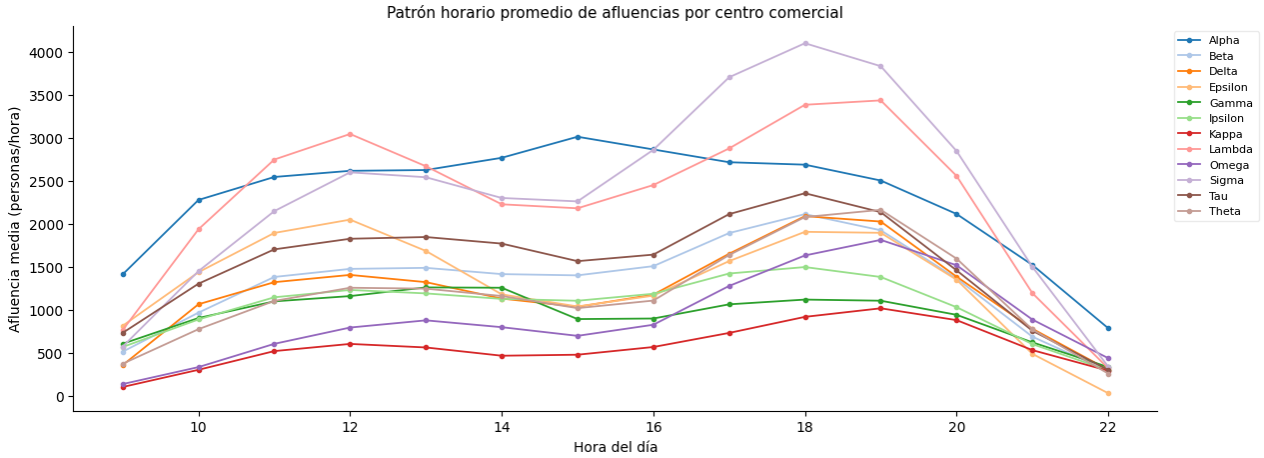
Figura 2. Evolución mensual de afluencias por centro comercial



Estacionalidad intradía: patrón horario

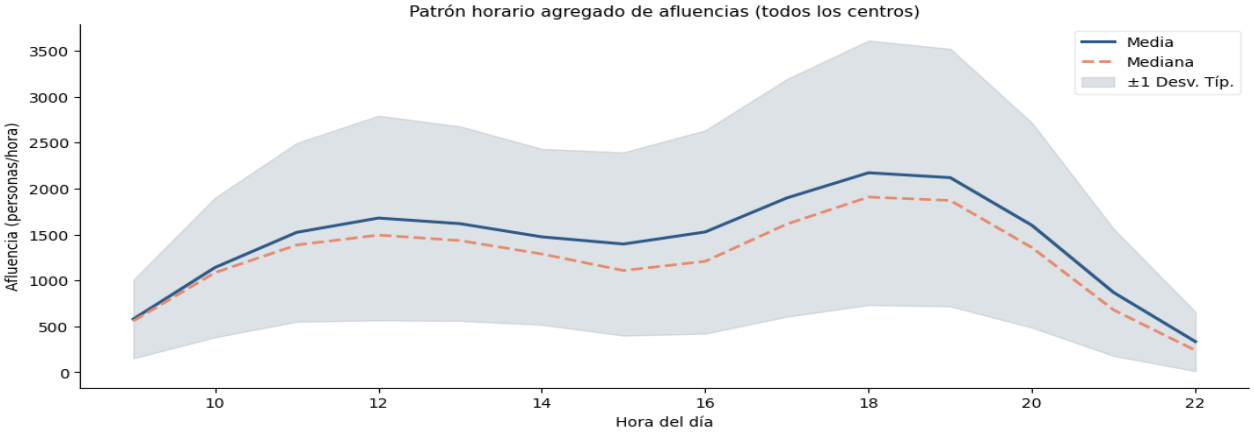
La Figura 3 presenta el patrón horario promedio de afluencias para cada centro, calculado sobre las franjas de actividad comercial (9:00–22:00). Se identifica un patrón común de doble pico: un primer incremento matutino con máximo entre las 12:00 y las 14:00, un descenso parcial en las horas centrales de la tarde, y un segundo pico entre las 17:00 y las 19:00, coincidente con el flujo de salida laboral. Este patrón es consistente con la literatura sobre comportamiento de compra en centros comerciales (Fildes, Goodwin & Nikolopoulos, 2019).

Figura 3. Patrón horario promedio de afluencias por centro comercial



La Figura 4 muestra el patrón horario agregado con bandas de variabilidad. La diferencia entre media y mediana nos revela una distribución asimétrica positiva, con colas hacia valores altos de afluencia.

Figura 4. Patrón horarilo agregado con bandas de variabilidad.

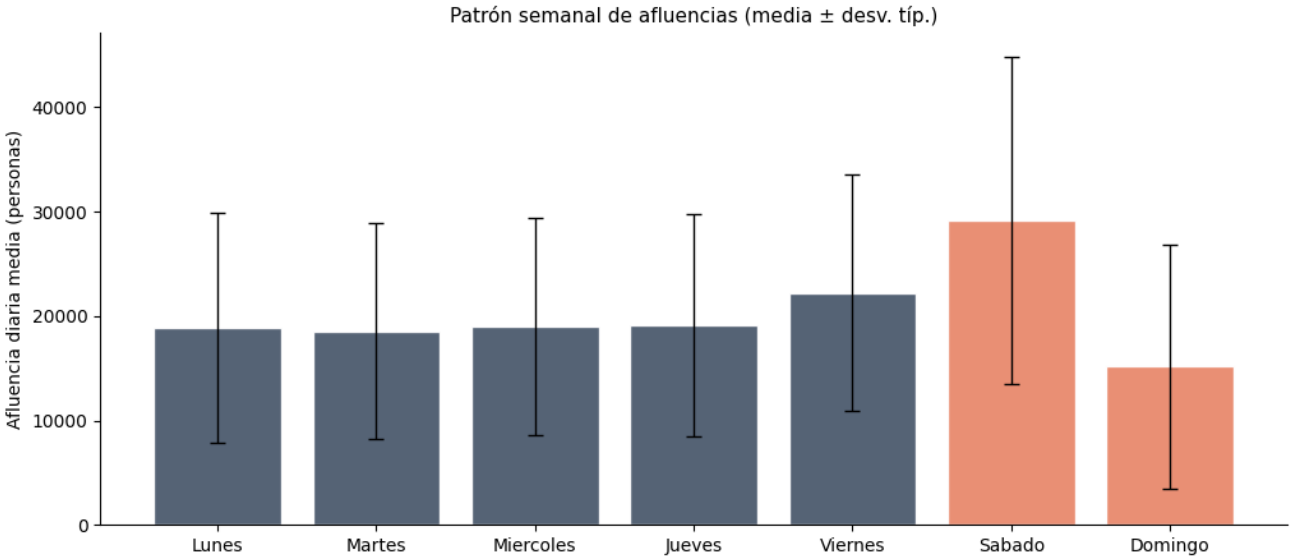


Estacionalidad semanal

La Figura 5 ilustra el patrón semanal de afluencias. Se observa un comportamiento diferenciado entre días laborables y fines de semana: los sábados registran sistemáticamente las mayores afluencias, seguidos de los viernes. Los domingos muestran niveles inferiores, probablemente asociados a la restricción de horarios de apertura vigente en algunas comunidades autónomas. Entre los días laborables, las afluencias se mantienen relativamente estables, con un ligero

incremento progresivo de lunes a viernes.

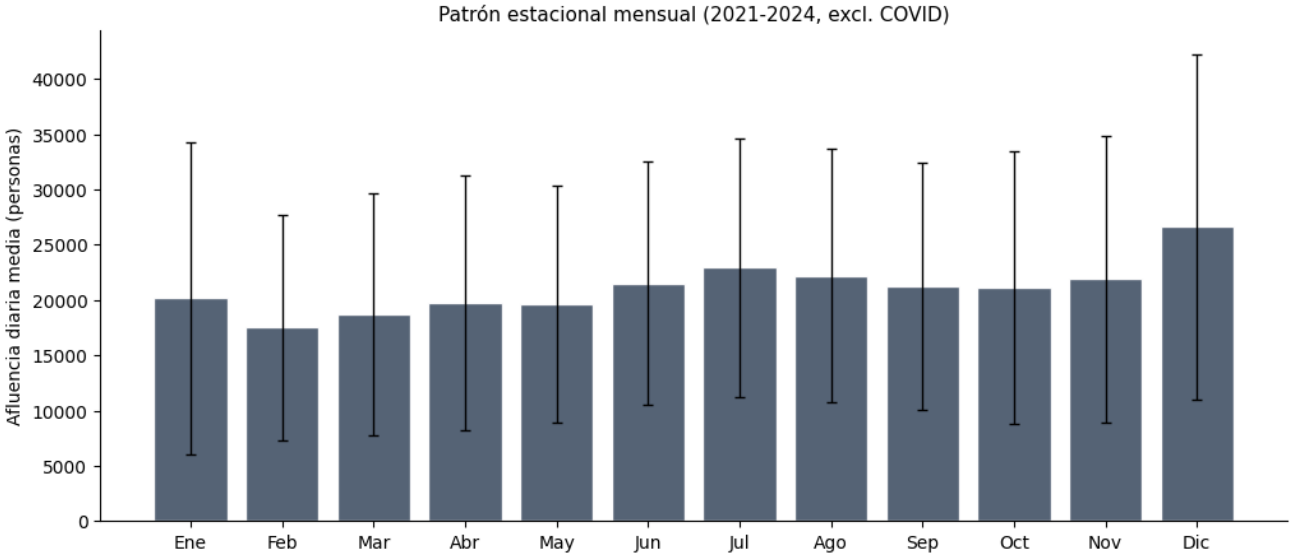
Figura 5. Patrón semanal de afluencias (media ± desviación típica)



Estacionalidad mensual

La Figura 6 presenta el patrón estacional mensual, calculado sobre los años completos no afectados por la pandemia (2021–2024). Se identifican picos de afluencia en diciembre (campaña navideña), marzo (Semana Santa) y julio (rebajas de verano). Los meses de enero y agosto presentan las menores afluencias, coincidiendo con el período postvacacional y la época estival.

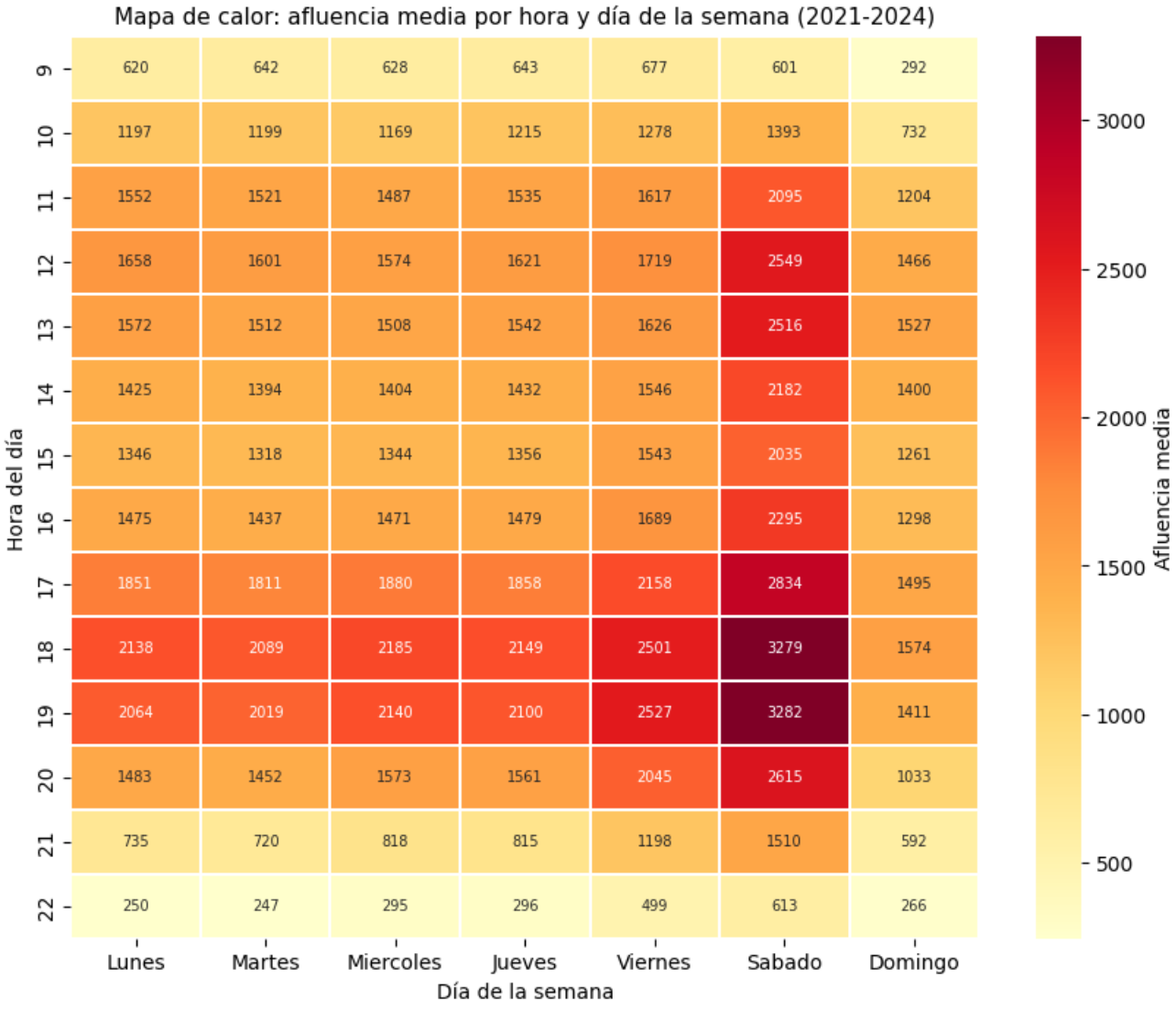
Figura 6. Patrón estacional mensual (2021–2024)



Mapa de calor: interacción hora × día de la semana

La Figura 7 combina las dos estacionalidades principales en un mapa de calor. Los sábados de 12:00 a 14:00 y de 17:00 a 19:00 concentran las mayores afluencias, alcanzando valores medios superiores a 1.800 personas/hora. Esta información resulta de especial interés para la planificación operativa y la asignación de recursos humanos y logísticos.

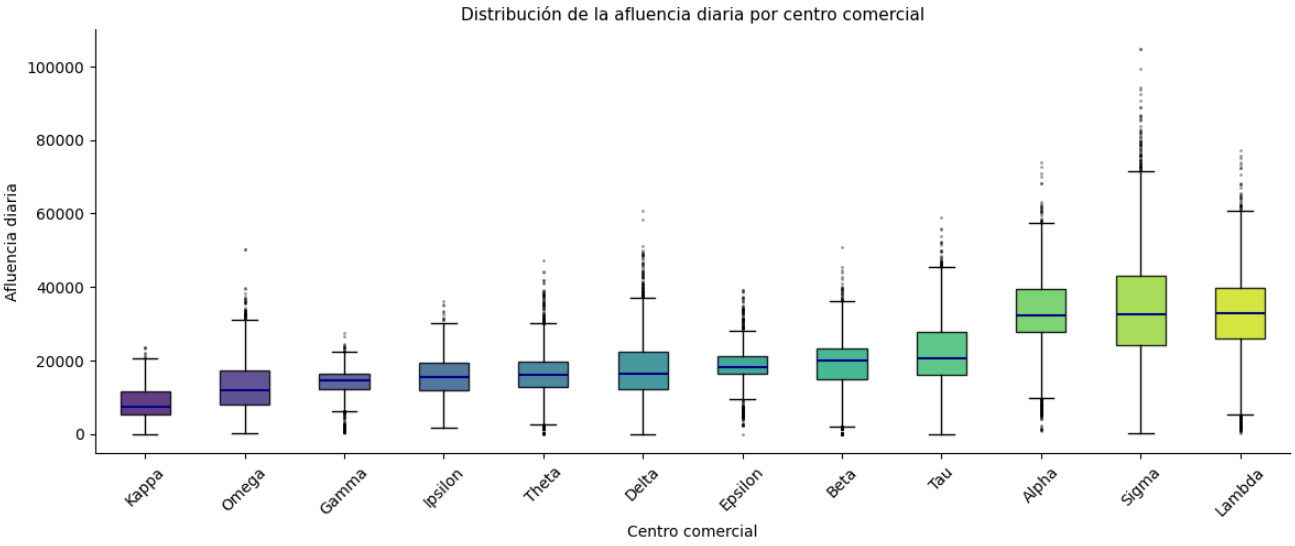
Figura 7. Mapa de calor: afluencia media por hora y día (2021–2024)



Distribución y dispersión de la afluencia

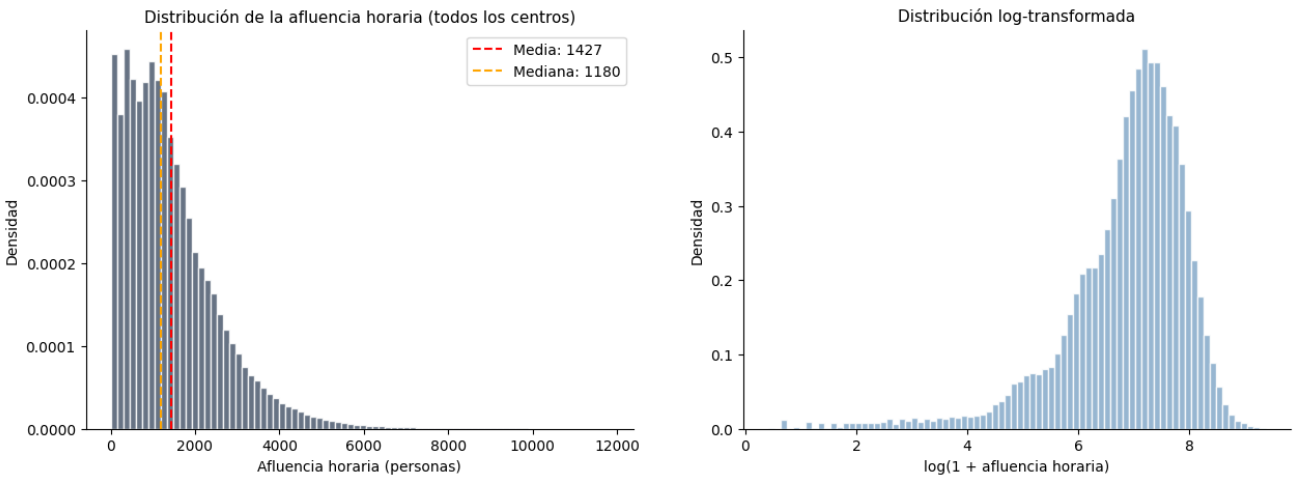
La Figura 8 presenta la distribución de la afluencia diaria por centro mediante diagramas de caja. La heterogeneidad entre centros es notable: Sigma y Alpha superan los 40.000 visitantes diarios de mediana, mientras que Kappa y Omega se sitúan por debajo de 10.000. La presencia de numerosos valores atípicos en la parte superior refleja el efecto de días con afluencias excepcionales, como festivos o períodos de rebajas.

Figura 8. Distribución de la afluencia diaria por centro comercial



El análisis de la distribución horaria (Figura 9) muestra una asimetría positiva marcada, con una media de 1.427 personas/hora y una mediana de 1.180. La transformación logarítmica aproxima la distribución a una forma más simétrica, lo que sugiere su utilidad potencial en la fase de modelización. Los tests de Shapiro-Wilk rechazan la hipótesis de normalidad en todos los centros ($p < 0{,}001$), resultado esperable dada la naturaleza estacional de los datos (Hastie, Tibshirani & Friedman, 2009).

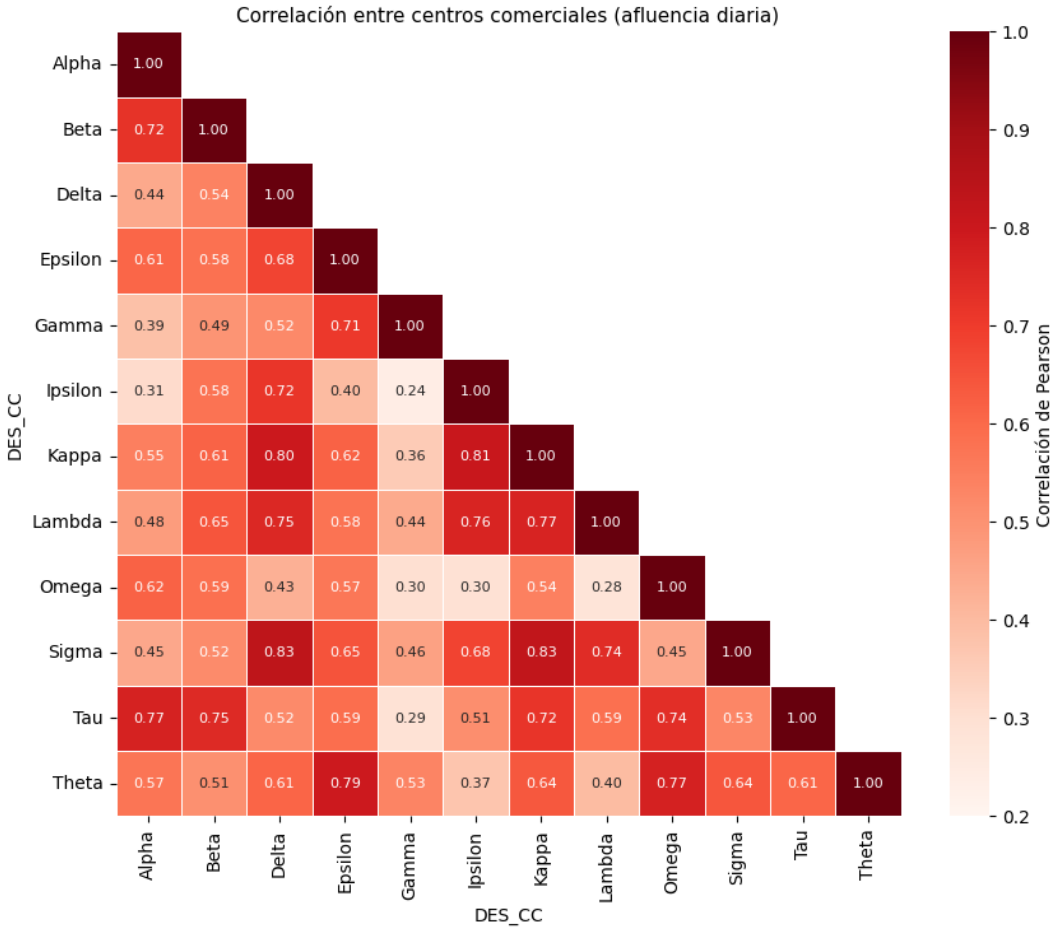
Figura 9. Distribución de la afluencia horaria y transformación logarítmica



Correlación entre centros comerciales

La Figura 10 presenta la matriz de correlación de Pearson entre las afluencias diarias de los 12 centros. Las correlaciones son predominantemente positivas y elevadas (rango 0,24–0,83), lo que nos indica que los centros comparten factores comunes que determinan las fluctuaciones en la afluencia, tales como festivos nacionales, condiciones meteorológicas o tendencias estacionales.

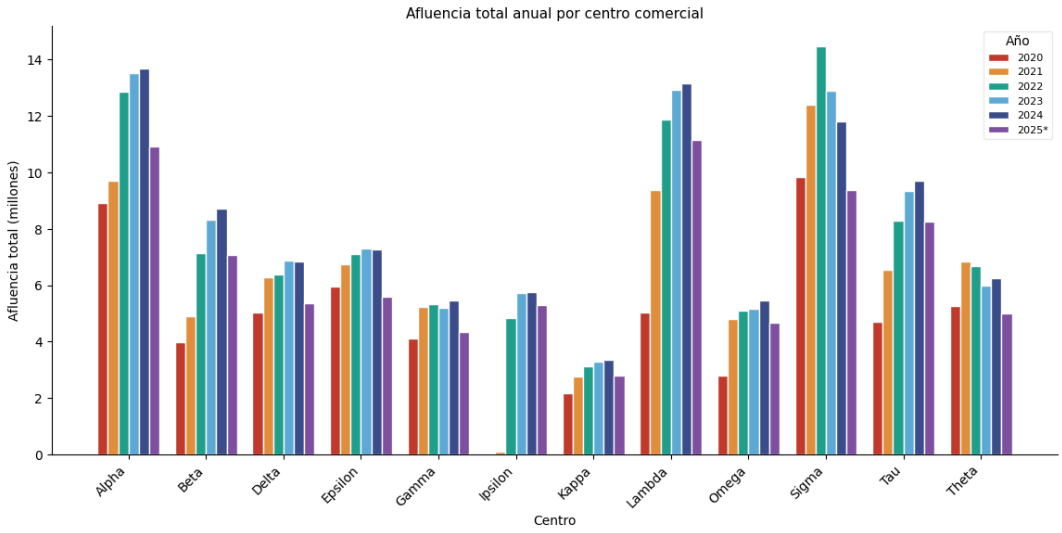
Figura 10. Matriz de correlación entre centros comerciales (afluencia diaria)



Comparación interanual

La Figura 11 muestra la afluencia total anual por centro. La magnitud de la caída en 2020 y la velocidad de recuperación varían significativamente. Lambda y Tau destacan por haber superado con claridad los niveles de 2020, mientras que Theta muestra una tendencia descendente desde 2022.

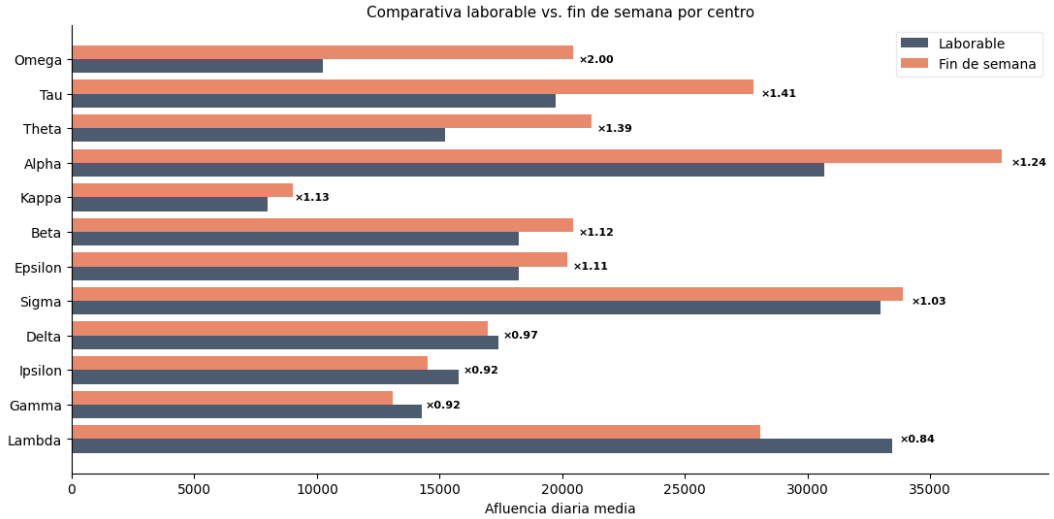
Figura 11. Afluencia total anual por centro comercial



Efecto fin de semana

La Figura 12 compara las afluencias medias en días laborables y fines de semana. El ratio fin de semana/laborable oscila entre 1,03 (Gamma) y 1,55 (Alpha). Esta variabilidad refleja diferencias en el perfil de visitantes: centros con mayor componente de ocio tienden a presentar ratios más elevados, mientras que aquellos situados en zonas residenciales muestran patrones más estables.

Figura 12. Comparativa laborable vs. fin de semana por centro

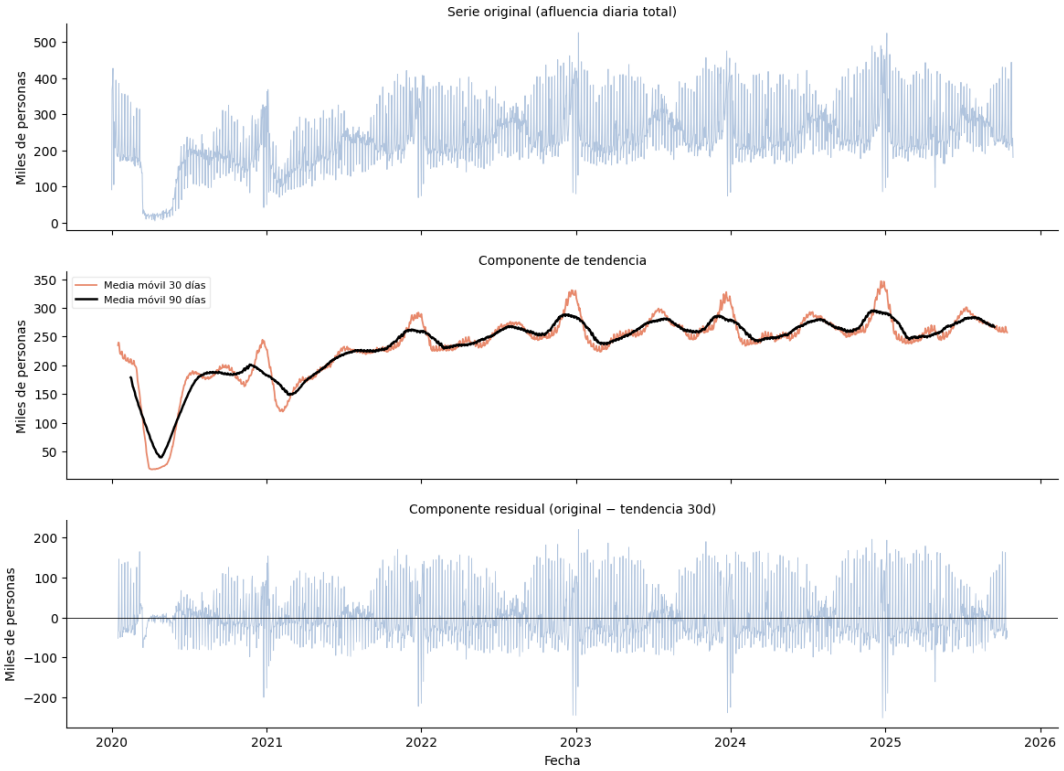


Descomposición de tendencia

La Figura 13 presenta una descomposición de la serie temporal agregada diaria, basada en medias móviles de 30 y 90 días. El componente de tendencia muestra con claridad la caída

pandémica, la recuperación en forma de V asimétrica durante 2021 y la posterior estabilización. El componente residual revela fluctuaciones estacionales regulares de alta frecuencia. Una descomposición formal mediante el método STL (Cleveland, Cleveland, McRae & Terpenning, 1990) podrá realizarse en la fase de modelización para una separación más rigurosa.

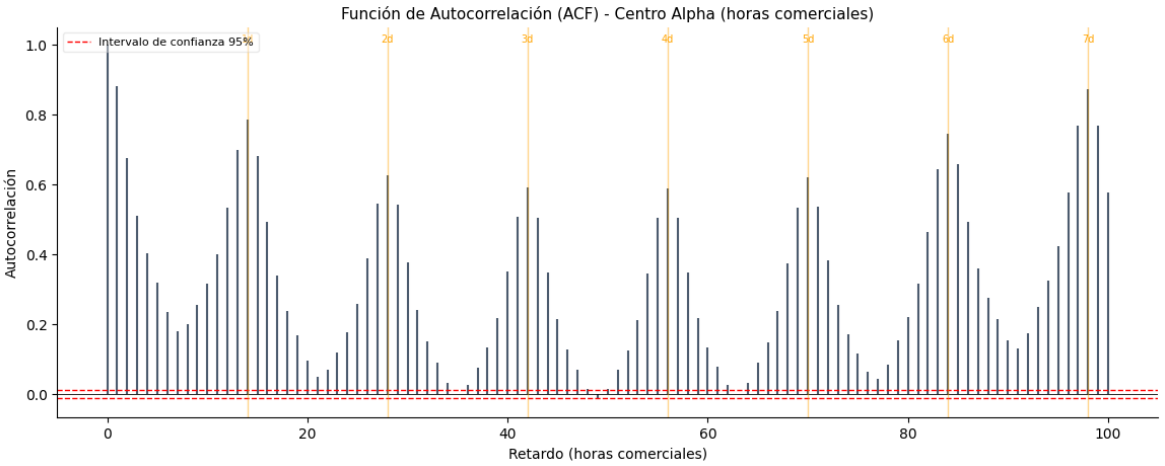
Figura 13. Descomposición de tendencia (serie diaria agregada)



Autocorrelación temporal

La Figura 14 muestra la función de autocorrelación (ACF) para el centro Alpha, calculada sobre las horas comerciales. Se observan picos significativos cada 14 retardos (un día de horas comerciales) y cada 98 retardos (una semana), lo que confirma cuantitativamente la existencia de estacionalidades intradía e intrasemanal. La lenta caída de la ACF sugiere asimismo la presencia de una componente de tendencia, lo que puede requerir diferenciación en los modelos ARIMA (Box, Jenkins, Reinsel & Ljung, 2015).

Figura 14. Función de autocorrelación (ACF) – Centro Alpha



Síntesis del análisis exploratorio

El análisis exploratorio nos ha permitido identificar los siguientes hallazgos clave que orientarán la fase de modelización predictiva:

- Las afluencias presentan una estructura de estacionalidad múltiple: intradía (patrón de doble pico), semanal (efecto fin de semana) y anual (picos navideños, Semana Santa y rebajas). Estos patrones deberán ser capturados por los modelos.

- La distribución de las afluencias es asimétrica positiva y no normal, lo que nos sugiere la conveniencia de transformaciones logarítmicas o el uso de modelos robustos.

- Existe una elevada correlación entre centros, lo que apunta a la influencia de factores exógenos comunes (clima, festivos, ciclo económico) que justifican la inclusión de variables externas.

- La recuperación pospandémica ha sido heterogénea entre centros, con trayectorias divergentes a partir de 2023.

- La autocorrelación temporal significativa confirma la idoneidad de modelos de series temporales (ARIMA/SARIMA), mientras que la presencia de relaciones no lineales y múltiples variables explicativas justifica la aplicación complementaria de modelos de Machine Learning.

4. Desarrollo del modelo predictivo

En este capítulo se desarrollan los modelos predictivos. Se implementan cuatro modelos con enfoques distintos. Se entrenan con el histórico de afluencias y se comparan entre sí. El objetivo es identificar el modelo más adecuado para generar las proyecciones del capítulo siguiente.

4.1. Modelos implementados

La elección de los cuatro modelos responde a una estrategia de triangulación. Cada uno representa un enfoque distinto del forecasting. Los modelos son: Naïve lag-7 (modelo de referencia), Random Forest (Machine Learning), SARIMA (serie temporal clásica) y Prophet (serie temporal moderna). Esto nos permite comparar técnicas con fundamentos diferentes.

Modelo 1: Naïve lag-7

El Naïve lag-7 es el modelo de referencia. Predice la afluencia de un día como la afluencia del mismo día de la semana anterior. Asume que los patrones semanales son estables. Sirve como umbral mínimo: cualquier modelo debe superarlo para ser útil (Hyndman & Athanasopoulos, 2021). Si un modelo más complejo no supera al Naïve, su complejidad añadida no aporta valor.

Modelo 2: Random Forest

El Random Forest es el modelo principal del trabajo. Es un algoritmo de Machine Learning propuesto por Breiman (2001). Combina múltiples árboles de decisión mediante bagging. Cada árbol se entrena con un subconjunto aleatorio de datos y variables. La predicción final es el promedio de las predicciones individuales. Esto reduce la varianza y mitiga el sobreajuste. Se implementa con scikit-learn (Pedregosa et al., 2011).

El modelo opera sobre 24 variables explicativas agrupadas en seis bloques:

• Temporales (5): día de la semana, mes, semana del año, fin de semana, día del año.

• Cíclicas (4): funciones seno y coseno del período anual y semanal. Capturan la naturaleza circular del tiempo.

• Contextuales (3): dos indicadores COVID y código del centro.

• Retardos (2): afluencia de hace 7 y 14 días. Son las variables más importantes del modelo.

• Calendario (7): festivos, vísperas, Navidad, rebajas y Black Friday.

• Meteorológicas AEMET (3): temperatura media, precipitación y horas de sol.

Los hiperparámetros se ajustan por validación cruzada temporal. Son: n_estimators=300, max_depth=20 y min_samples_leaf=5. La validación respeta el orden cronológico. Esto evita la fuga de información futura durante el entrenamiento.

Modelo 3: SARIMA

SARIMA(1,1,1)(1,1,1)[7] es un modelo estadístico clásico (Box, Jenkins, Reinsel & Ljung, 2015). Combina componentes autorregresivos, de medias móviles y diferenciaciones. El orden [7] indica estacionalidad semanal. Se ajusta por separado a cada centro. Trabaja solo con la serie univariante. No incorpora variables externas. Este contraste frente al Random Forest permite evaluar el valor añadido de las variables contextuales.

Modelo 4: Prophet

Prophet es un modelo desarrollado por Meta (Taylor & Letham, 2018). Descompone la serie en cuatro componentes: tendencia, estacionalidad anual, estacionalidad semanal y efectos de festivos. Cada componente se modela por separado y luego se suman. Prophet es robusto ante valores atípicos y datos faltantes. Esto es relevante dado el impacto del COVID-19 en la serie.

Se le incorporan los festivos del BOE y los regresores meteorológicos de AEMET.

Partición de datos

El entrenamiento abarca desde enero de 2020 hasta diciembre de 2024 (20.882 registros). El test va de enero a octubre de 2025 (3.617 registros). Esta partición asegura que los modelos se evalúan sobre un período futuro real. Es decir, datos no vistos durante el entrenamiento.

4.2. Evaluación y comparación de resultados

La evaluación se realiza mediante tres métricas: RMSE, MAE y R^2. El RMSE penaliza más los errores grandes porque los eleva al cuadrado. El MAE los trata por igual. El R^2 mide la varianza explicada por el modelo. Va de $-\infty$ a 1. Un R^2 de 1 indica predicción perfecta. Un R^2 de 0 indica que el modelo no mejora a la simple media histórica.

Resultados globales

La Tabla 6 presenta los resultados de los cuatro modelos sobre el conjunto de test.

Tabla 6. Comparación global de modelos (test: ene–oct 2025)

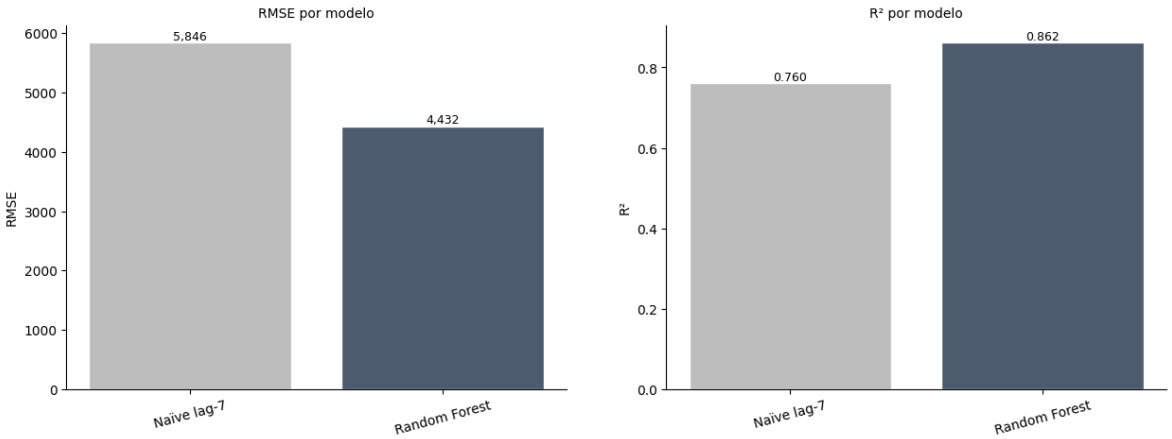
Modelo	RMSE	MAE	R^2	Mejora vs Naïve
Naïve (lag 7)	5.846	3.173	0,760	—
Random Forest	4.432	2.594	0,862	24,2%
Prophet + AEMET	4.845	3.233	0,835	17,1%
SARIMA(1,1,1)(1,1,1)[7]	22.215	16.769	-2,459	—

El Random Forest obtiene el mejor resultado. Alcanza un R^2 de 0,862 y un RMSE de 4.432. Supera al Naïve en un 24,2%. El enfoque de Machine Learning con variables externas captura mejor la dinámica de las afluencias. Las variables externas aportan información contextual relevante (Makridakis, Spiliotis & Assimakopoulos, 2020).

Prophet se sitúa segundo con un R^2 de 0,835. Supera al Naïve en un 17,1%. Aunque va por detrás del Random Forest, destaca por su facilidad de uso e interpretabilidad. Es una alternativa valiosa para equipos con menor especialización técnica.

SARIMA obtiene un R^2 negativo (-2,459), peor que la propia media histórica. El resultado no invalida el método: refleja las condiciones del experimento. Como señalan Hyndman y Athanasopoulos (2021), estos modelos requieren un ajuste de parámetros específico por serie (auto.arima o búsqueda en rejilla con AIC/BIC). Aquí se usó deliberadamente una parametrización fija y común a todos los centros —SARIMA(1,1,1)(1,1,1)[7]—, sin optimizar y sin regresores exógenos, para contrastarla con el Random Forest. La conclusión no es que SARIMA sea malo, sino que, mínimamente especificado y sin variables externas, resulta insuficiente para una serie de esta complejidad. Un SARIMAX optimizado por centro queda como línea futura.

Figura 15. Comparación de modelos (RMSE y R^2)



Resultados por centro (Random Forest)

La Tabla 7 muestra las métricas desagregadas por centro. Permite identificar dónde funciona mejor el modelo.

Tabla 7. Métricas por centro comercial

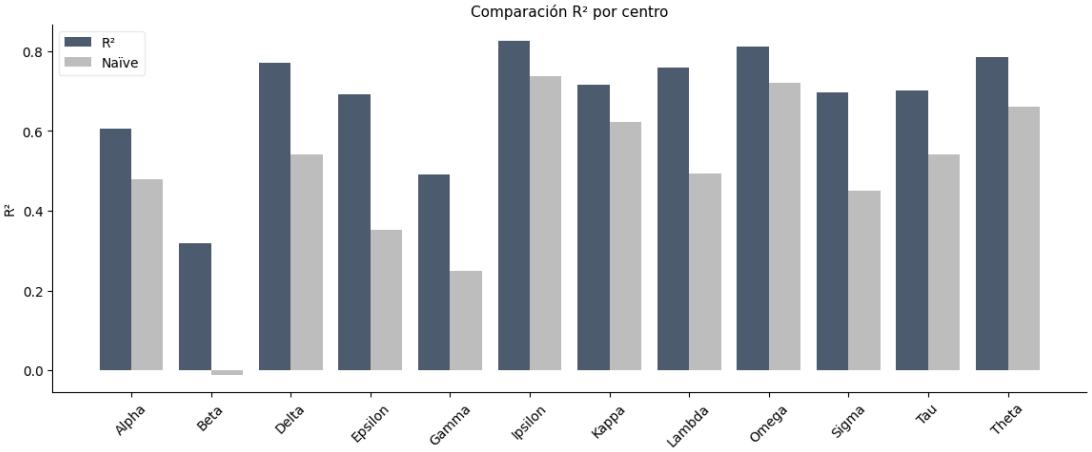
Centro	RMSE	MAE	R^2	R^2 Naïve	Mejora RMSE
Alpha	5.068	3.262	0,606	0,478	13,2%
Beta	4.110	2.761	0,319	-0,012	18,0%
Delta	4.402	3.019	0,770	0,542	29,1%
Epsilon	2.321	1.652	0,693	0,353	31,1%
Gamma	1.788	1.263	0,490	0,250	17,5%
Ipsilon	2.604	1.713	0,826	0,737	18,6%
Kappa	2.404	1.536	0,717	0,622	13,5%
Lambda	6.178	4.040	0,760	0,494	31,1%
Omega	3.124	2.046	0,812	0,720	18,1%

Sigma	8.888	5.138	0,698	0,451	25,8%
Tau	4.508	3.175	0,702	0,542	19,3%
Theta	2.318	1.519	0,785	0,661	20,3%

El Random Forest supera al Naïve en todos los centros sin excepción. Las mejoras oscilan entre el 13,2% (Alpha) y el 31,1% (Epsilon y Lambda). Los mejores ajustes están en Ipsilon (R^2=0,826), Omega (R^2=0,812) y Theta (R^2=0,785). Estos centros presentan patrones muy regulares.

El peor ajuste es Beta (R^2=0,319). Se explica por su alta proporción de valores atípicos. El Naïve en este centro tenía un R^2 negativo. El Random Forest mejora significativamente incluso en el caso más difícil.

Figura 16. Comparación R^2 por centro



Importancia de variables

El análisis de importancia revela qué variables son más relevantes para la predicción:

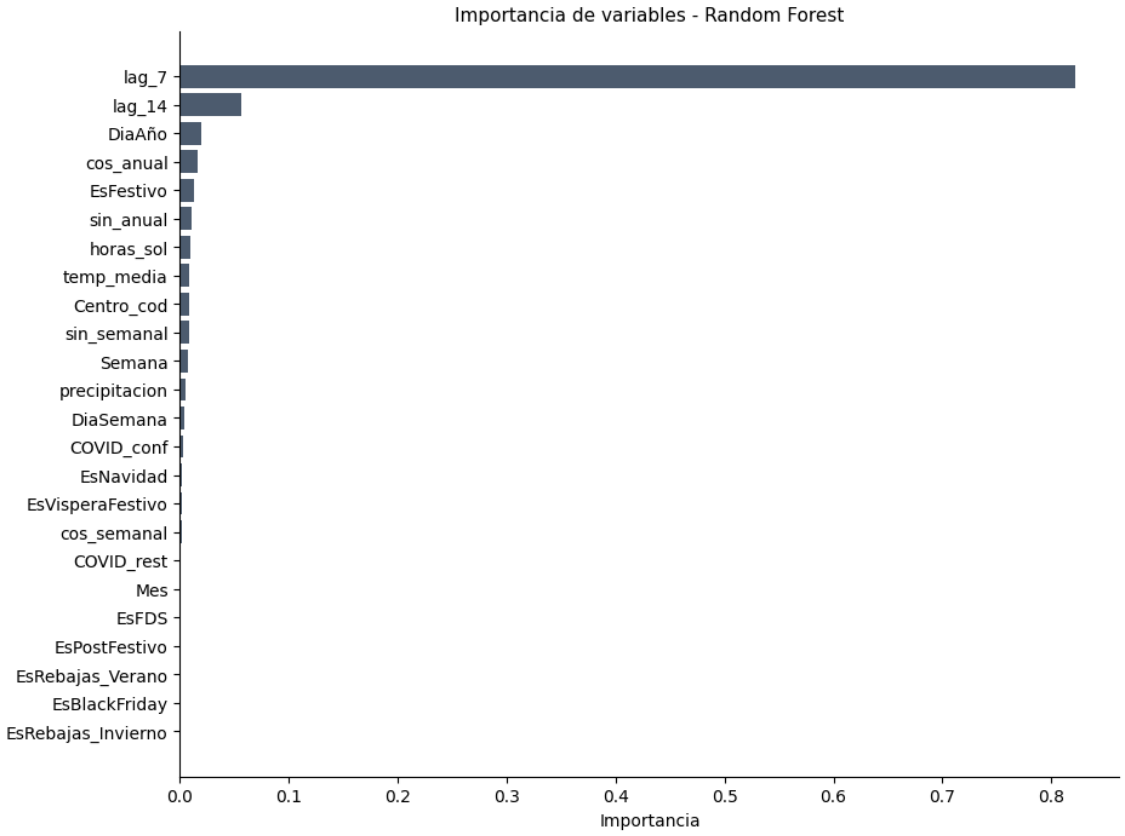
- lag_7 (82,2%) y lag_14 (5,6%): el patrón semanal es el factor dominante.
- horas_sol (0,9%): la variable meteorológica más importante.

- EsFestivo (1,3%): los festivos ocupan el quinto lugar.

- temp_media (0,9%) y precipitacion (0,5%): el resto de variables meteorológicas.

Las variables externas (festivos + AEMET) concentran un 4% de la importancia agregada, frente al 87,8% que acumulan los retardos lag_7 y lag_14. Esta distribución podría sugerir, a primera vista, que el peso explicativo de los factores exógenos es marginal. Conviene, sin embargo, evitar esa lectura simplista por dos motivos. Primero, en series con fuerte autocorrelación (como es el caso) los retardos absorben gran parte de la varianza explicable, dejando a las variables externas un papel de ajuste fino sobre los días en los que el patrón semanal se rompe (festivos, vísperas, rebajas, eventos climatológicos extremos); este es justamente el comportamiento que se observa en las Figuras 18, 19 y 20. Segundo, la mejora del RMSE de Random Forest frente al Naïve lag-7 (-24,2%) no puede atribuirse a los retardos (que el propio Naïve ya utiliza implícitamente), sino al conjunto de variables temporales cíclicas, calendario y meteorología que el Random Forest sí incorpora. En otras palabras: el 4% de importancia agregada de las variables externas se traduce en aproximadamente una cuarta parte de la mejora total del modelo sobre la referencia, lo que justifica plenamente el esfuerzo de integración de las fuentes externas (AEMET, BOE) en el pipeline.

Figura 17. Importancia de variables - Random Forest



Visualización de predicciones

Las siguientes figuras muestran las predicciones del modelo para tres centros representativos.

Las líneas verdes marcan los festivos nacionales.

Figura 18. Predicción vs Real – Alpha

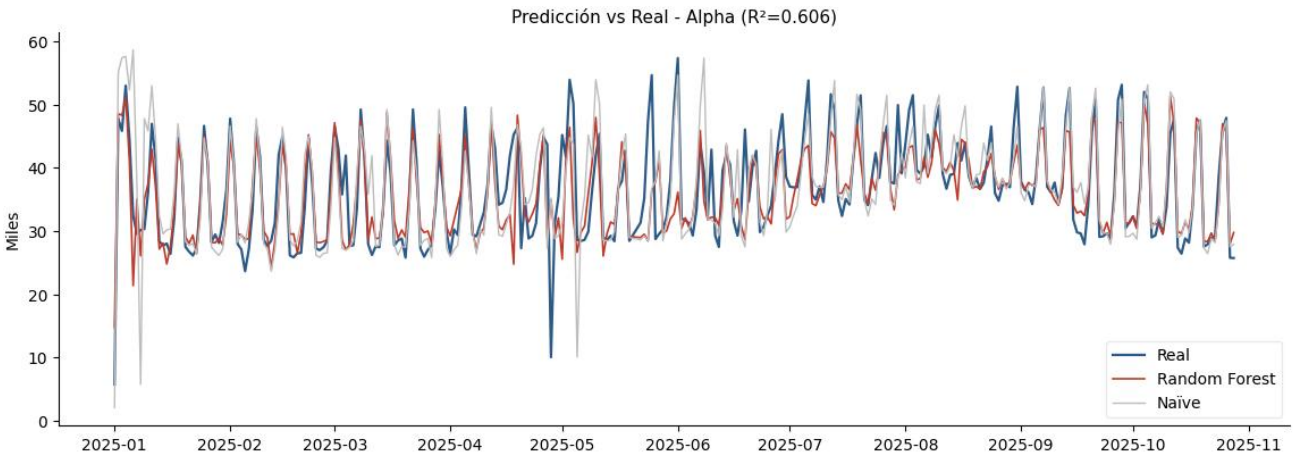


Figura 19. Predicción vs Real – Kappa

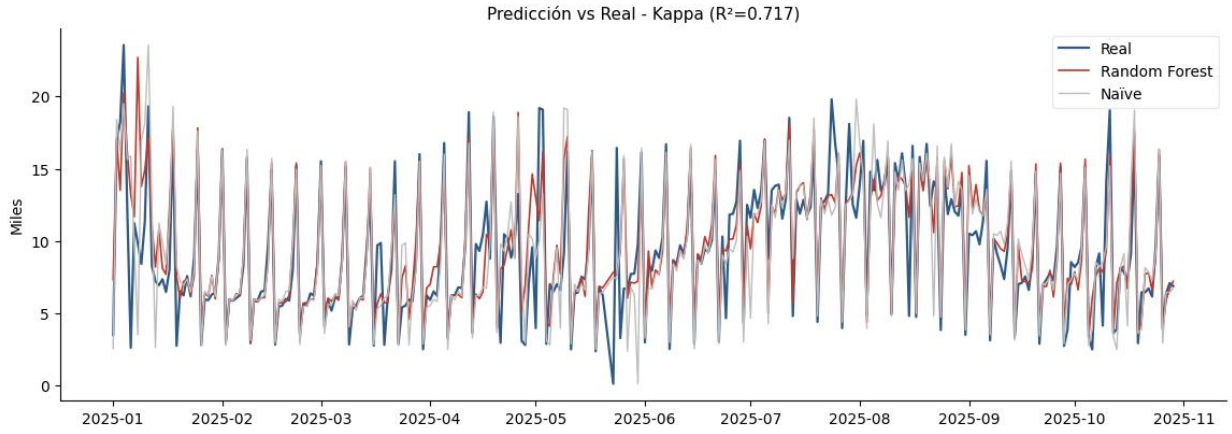
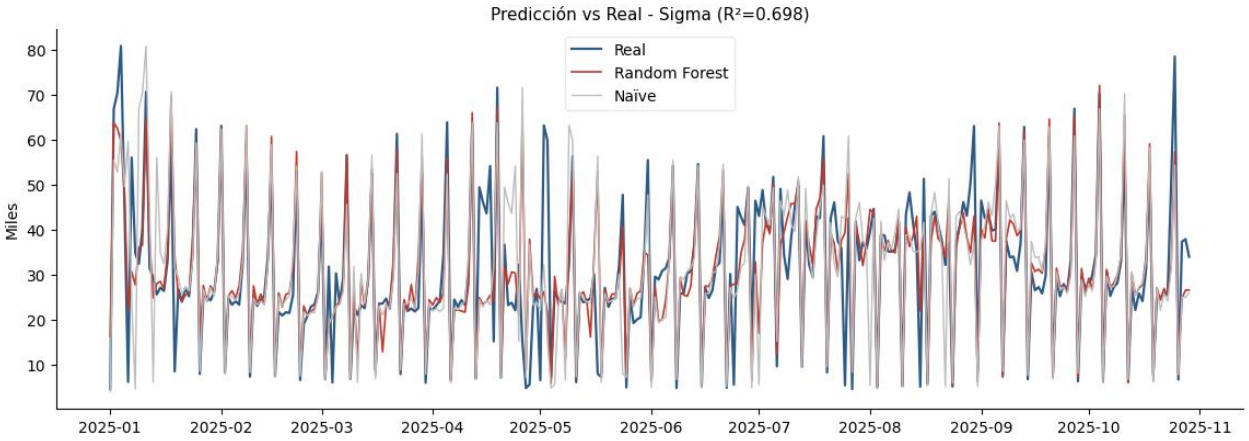


Figura 20. Predicción vs Real – Sigma



En los tres casos las predicciones siguen bien la forma de la serie real. El modelo captura los fines de semana (picos), los laborables (valles) y los patrones estacionales. Las mayores desviaciones ocurren en fechas atípicas o cercanas a festivos.

5. Predicciones y análisis de escenarios

Una vez validado el Random Forest, se generan proyecciones a medio y largo plazo. El período de proyección es 2026-2035 (10 años). Este horizonte es relevante para las decisiones estratégicas del sector inmobiliario comercial (CBRE, 2025). Los ciclos de inversión y las decisiones sobre reformas requieren visibilidad a largo plazo.

Las proyecciones a largo plazo tienen mucha incertidumbre. Cuanto más lejos se proyecta, más probable es que aparezcan factores imprevistos. Por eso no se presenta una única predicción. Se construyen tres escenarios: base, optimista y pesimista. Adicionalmente, se realiza un análisis de sensibilidad sobre las variables externas.

5.1. Metodología de proyección

El modelo Random Forest se aplica sobre un dataset futuro. Cubre desde el 30 de octubre de 2025 hasta el 31 de diciembre de 2035. Son 3.715 días × 12 centros = 44.580 observaciones. El dataset futuro contiene las mismas variables que el modelo original.

Las variables del dataset futuro se construyen así:

- Temporales y cíclicas: se calculan del calendario. No requieren supuestos adicionales.
- Festivos: se extrapolan los festivos fijos del BOE a los años futuros. Semana Santa se calcula según el algoritmo litúrgico.
- Meteorológicas: media histórica por centro, mes y día de la semana. Asume que el patrón climático es estable a medio plazo.
- Retardos: media estacional histórica del mismo centro y día. Este enfoque evita acumular errores de predicciones iterativas.

A partir de ahí se generan tres escenarios con un ajuste anual compuesto:

- Escenario base: sin ajustes. Refleja la continuación de las tendencias 2022-2025.
- Escenario optimista: +1,5% anual. Asume recuperación sostenida del sector.
- Escenario pesimista: -1,0% anual. Asume impacto creciente del e-commerce (Verhoef, Kannan & Inman, 2015).

Los porcentajes se basan en la tasa media observada en 2022-2024 (+2,2% medio anual) y en estimaciones sectoriales de CBRE (2025) y la AECC. La asimetría refleja que los escenarios pesimistas en el retail español han sido históricamente más moderados que los optimistas.

5.2. Proyección agregada a 10 años

La Tabla 8 muestra las proyecciones agregadas de afluencias anuales para los 12 centros. Se presentan los años clave del horizonte.

Tabla 8. Proyección de afluencias agregadas (millones) – 12 centros

Año	Pesimista	Base	Optimista	Rango
2024 (real)	97,4	97,4	97,4	—
2025 (parcial)	79,9	79,9	79,9	—
2026	90,6	91,6	92,9	±1,2%
2027	89,6	91,4	94,2	±2,5%
2028	89,0	91,7	95,9	±3,8%
2030	87,1	91,6	98,7	±6,3%

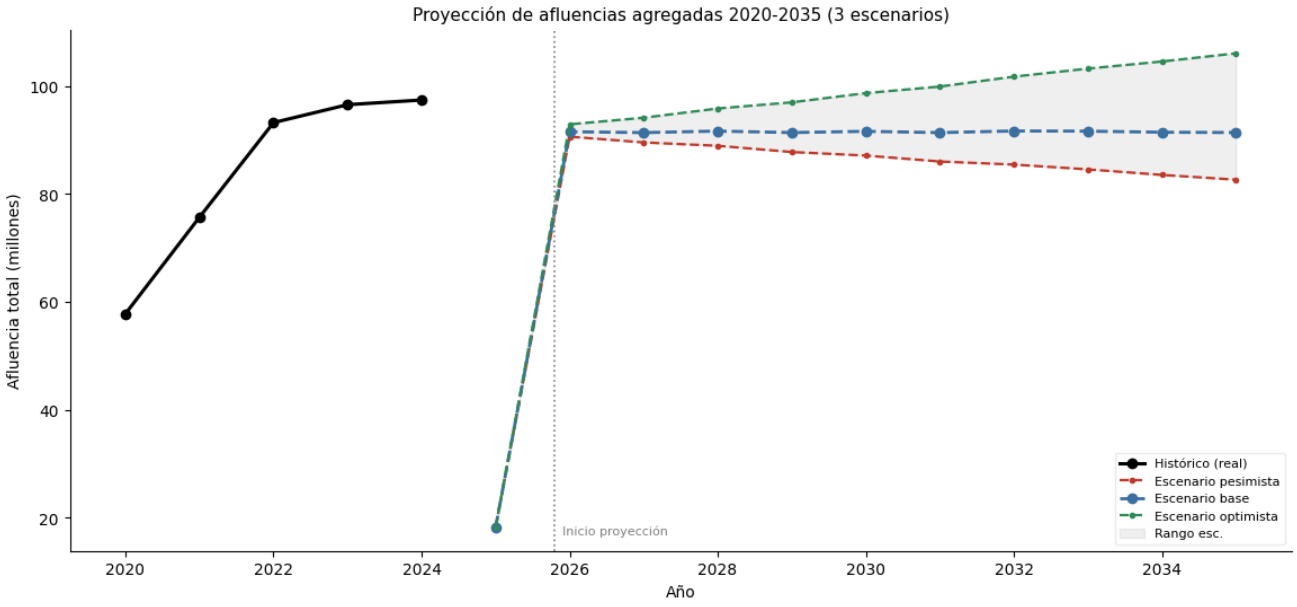
2033	84,6	91,7	103,3	±10,2%
2035	82,7	91,4	106,1	±12,8%

El escenario base se estabiliza en 91 millones anuales. Esto confirma que el ciclo de recuperación pospandémico ha finalizado. El sector entra en una fase de madurez. El nivel es coherente con los 96-97 millones observados en 2023-2024.

El escenario optimista llega a 106,1 millones en 2035 (+16% respecto al base). El pesimista cae a 82,7 millones (-9,5% respecto al base). La asimetría entre ambos responde al hecho de que las caídas del sector retail español han sido históricamente más moderadas que las fases expansivas.

El rango entre escenarios se ensancha con el tiempo. Pasa del ±1,2% en 2026 al ±12,8% en 2035. Esto refleja la acumulación de incertidumbre propia de las proyecciones a largo plazo. Es una propiedad inherente al forecasting extendido.

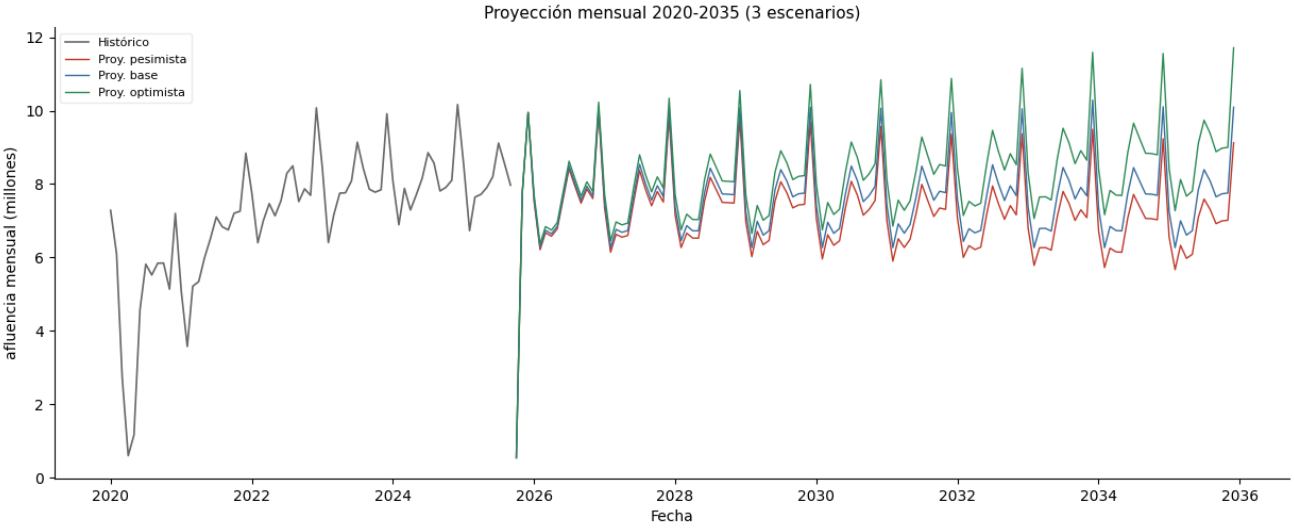
Figura 21. Proyección de afluencias agregadas 2020-2035



La Figura 22 muestra la proyección mensual. Se observa que las estacionalidades se

mantienen: picos navideños, caídas de agosto y ciclo semanal. Confirma que el modelo proyecta bien hacia el futuro. Es un indicador de calidad.

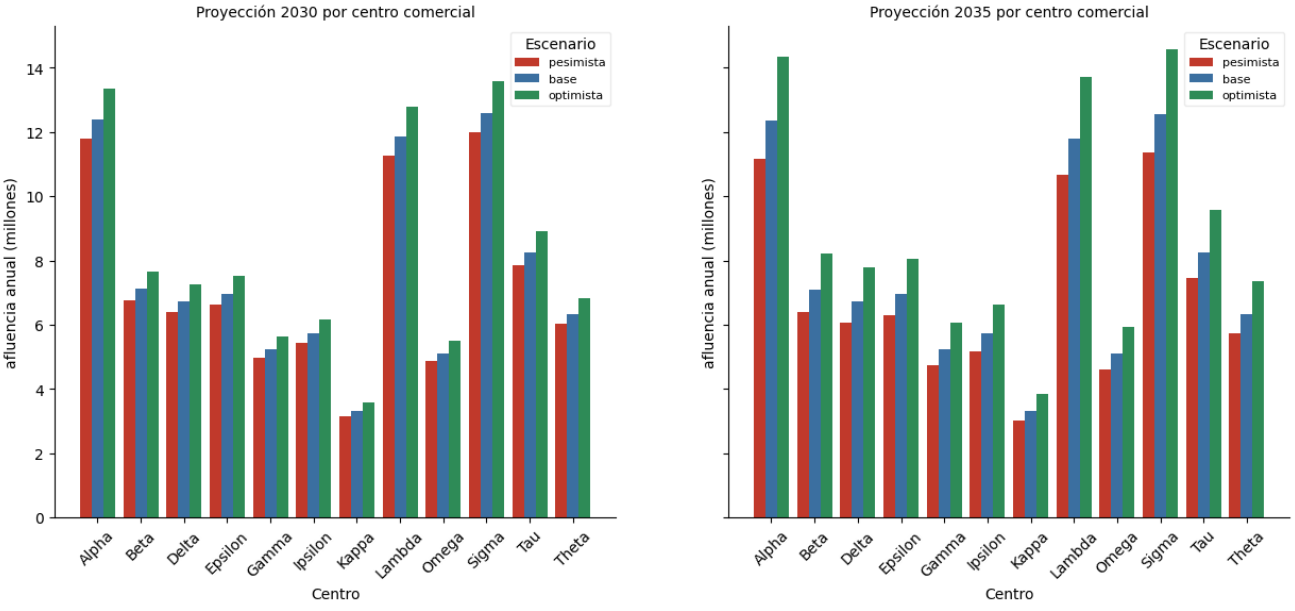
Figura 22. Proyección mensual 2020-2035 por escenario



5.3. Proyecciones por centro comercial

La Figura 23 presenta las proyecciones por centro para 2030 y 2035. Permite identificar qué centros tienen más potencial y cuáles más riesgo.

Figura 23. Proyección por centro comercial para 2030 y 2035



Los centros con mayor volumen proyectado son Sigma, Lambda y Alpha (>10 millones anuales).

Son los que generan más negocio y atraen más interés inversor. Los de menor volumen son Gamma y Kappa (<4 millones).

La diferencia entre escenarios es proporcionalmente mayor en los centros grandes. En términos absolutos, esto implica que están más expuestos al riesgo de caída. Pero también tienen mayor potencial de revalorización. Las estrategias de mitigación deben ser especialmente activas en los centros grandes.

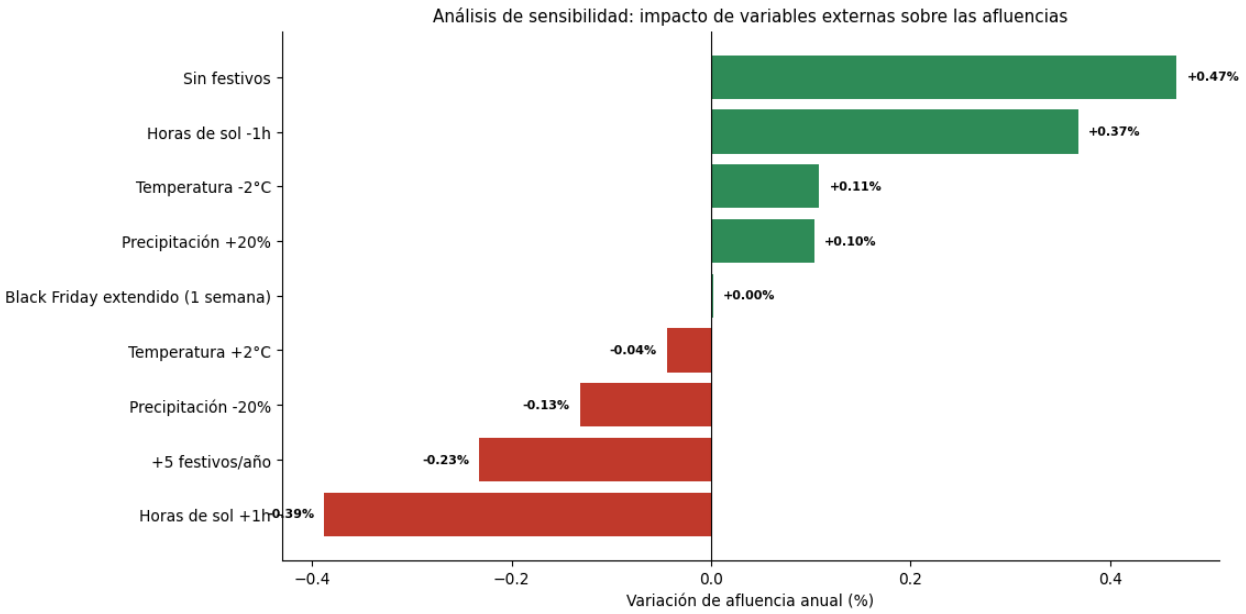
5.4. Análisis de sensibilidad

El análisis de sensibilidad mide cómo responden las afluencias ante cambios en las variables externas. Se modifican una a una. Los resultados están en la Tabla 9.

Tabla 9. Resultados del análisis de sensibilidad

Factor modificado	Variación afluencia anual
Temperatura +2°C	-0,04%
Temperatura -2°C	+0,11%
Precipitación +20%	+0,10%
Precipitación -20%	-0,13%
Horas de sol +1h/día	-0,39%
Horas de sol -1h/día	+0,37%
+5 festivos/año adicionales	-0,23%
Sin festivos (eliminar todos)	+0,47%
Black Friday extendido (1 semana)	0,00%

Figura 24. Análisis de sensibilidad



Los factores climáticos tienen un impacto limitado (inferior al 0,4%). La variable más sensible es las horas de sol. Con más horas de sol baja la afluencia un 0,39%. Tiene sentido económico: los consumidores prefieren actividades al aire libre. Los centros comerciales pierden atractivo relativo.

El calendario tiene el mayor impacto individual. Eliminar todos los festivos subiría la afluencia un 0,47%. Muchos festivos nacionales coinciden con viajes fuera del entorno urbano. Por eso reducen las visitas a centros. Añadir cinco festivos más tendría el efecto contrario (-0,23%).

El Black Friday extendido a una semana tiene un efecto prácticamente nulo sobre la afluencia anual agregada (0,00%), lo que sugiere que su impacto se concentra en una redistribución temporal de las visitas más que en un incremento neto. Es coherente con la tendencia del sector a prolongar estas campañas: primero a fin de semana, luego a semana completa y más recientemente a un mes.

Todos los factores tienen sensibilidad baja (<0,5%). El modelo es robusto ante variaciones individuales. Pero el efecto acumulado de varios factores en la misma dirección sí podría ser relevante. Un escenario con más temperatura, menos lluvia y más sol (compatible con el

cambio climático) podría situarse en torno al 0,6% anual.

5.5. Discusión estratégica y conclusiones del capítulo

Los resultados del modelo permiten extraer varias lecturas estratégicas para la gestión de centros comerciales, que van más allá de la mera proyección de afluencias.

1. Fin del ciclo de recuperación y transición a una fase de madurez. El escenario base se estabiliza en torno a 91 millones anuales, nivel coherente con los 96-97 millones observados en 2023-2024. Esto sugiere que la recuperación pospandémica ha concluido y que el sector entra en una fase de madurez en la que los crecimientos serán incrementales y dependerán cada vez más de la capacidad de cada activo para diferenciarse. Implicación para la gestión: los planes de negocio basados en hipótesis de recuperación continuada deben revisarse hacia escenarios de crecimiento plano o ligeramente positivo.

2. La incertidumbre crece con el horizonte y obliga a una gestión adaptativa. El rango entre escenarios pasa del ±1,2% en 2026 al ±12,8% en 2035 (amplitud total del 25,6%). Esto refleja una propiedad inherente al forecasting de largo plazo (Petropoulos, Makridakis & Assimakopoulos, 2022). Implicación para la gestión: las proyecciones a cinco años pueden utilizarse con razonable confianza para decisiones operativas y de marketing; las proyecciones a diez años deben tratarse como herramienta de planificación estratégica y *stress testing*, no como pronóstico puntual. Se recomienda reentrenar el modelo con cadencia anual incorporando los datos del último año cerrado.

3. El calendario domina sobre la meteorología como palanca externa. El análisis de sensibilidad sitúa al calendario (festivos, Black Friday) como el factor externo con mayor impacto (hasta ±0,47% sobre la afluencia anual), por delante de las variables meteorológicas (hasta ±0,39%,

correspondiente a las horas de sol). Implicación para la gestión: la planificación de campañas comerciales, dotación de personal y programación de eventos debe construirse sobre la columna del calendario, con especial atención a vísperas de festivos, periodos de rebajas y la creciente extensión del Black Friday, antes que sobre escenarios meteorológicos.

4. Heterogeneidad entre centros: el riesgo está concentrado en los grandes. Sigma, Lambda y Alpha aglutinan los mayores volúmenes proyectados (>10 millones anuales), pero también la mayor exposición absoluta a los escenarios pesimistas. Implicación para la gestión: las estrategias activas de mitigación (renovación de *tenant mix*, inversión en experiencia, eventos, restauración y ocio) deben priorizarse en los centros grandes, donde la asimetría de riesgo es mayor. En los centros pequeños (Gamma, Kappa) la prioridad debería ser la consolidación de la base de visitantes existente.

5. El comercio físico sigue siendo predecible, pero a costa de mayor sofisticación analítica. El hecho de que un Random Forest con variables externas mejore al Naïve en un 24,2% mientras que un SARIMA básico no lo consiga ilustra una tendencia más amplia: en un sector cada vez más sensible a factores contextuales, los modelos univariantes tradicionales pierden capacidad y los enfoques de Machine Learning con *feature engineering* enriquecido se vuelven el estándar (Giannopoulos et al., 2025). Implicación para la gestión: la inversión en capacidades analíticas internas (datos, talento, tooling) deja de ser un lujo y se convierte en un requisito competitivo.

En conjunto, las proyecciones describen un sector que ha completado su recuperación, ha entrado en madurez y necesita gestionarse con instrumentos analíticos más finos que en el pasado. La actualización periódica del modelo y la integración progresiva de nuevas fuentes de información (ventas, *tenant mix*, ticket medio, competencia local) serán claves para mantener la utilidad de este tipo de análisis a lo largo del tiempo.

6. Conclusiones y futuras líneas de investigación

6.1. Conclusiones generales

El presente trabajo ha analizado la dinámica de las afluencias en una cartera de doce centros comerciales españoles durante el período 2020-2025 y ha desarrollado un sistema predictivo capaz de proyectarlas a un horizonte de diez años. De la combinación del análisis exploratorio, la comparativa de modelos y el ejercicio de escenarios se desprenden las siguientes conclusiones:

1. Las afluencias presentan una estructura de estacionalidad múltiple (intradía, semanal y anual) y una elevada correlación entre centros, lo que confirma la influencia de factores exógenos compartidos (calendario, clima, ciclo económico).

2. El Random Forest con variables externas se consolida como el mejor modelo de los evaluados (R^2=0,862; RMSE=4.432), superando al Naïve lag-7 en un 24,2% y a Prophet en términos absolutos de error.

3. SARIMA, en su parametrización fija y sin regresores exógenos, resulta insuficiente para este tipo de serie. La incorporación de variables externas y la optimización individualizada por centro son condiciones necesarias para que los modelos clásicos sean competitivos.

4. El sector ha completado la recuperación pospandémica y entra en una fase de madurez con un nivel base aproximado de 91 millones de visitas anuales para la cartera analizada.
5. El calendario es la palanca externa con mayor impacto sobre las afluencias, por encima de las variables meteorológicas, lo que tiene implicaciones directas sobre la planificación comercial.

6. Los resultados son reproducibles: el código y los notebooks asociados están disponibles en GitHub (https://github.com/juanjpumar-collab/TFG-CC), lo que permite la verificación externa y la reutilización de la metodología en otros activos o sectores.

6.2. Limitaciones del estudio

El estudio presenta varias limitaciones que conviene reconocer explícitamente:

- Anonimización y validación externa: la pseudonimización de los centros impide contrastar los resultados con factores competitivos o de localización específicos, así como analizar el efecto del *tenant mix* concreto de cada activo.

- Supuestos de estabilidad: el Random Forest asume continuidad en los patrones históricos, lo que puede no cumplirse ante cambios estructurales (nuevas regulaciones, disrupciones tecnológicas, crisis sectoriales). Las variables meteorológicas se han proyectado mediante medias históricas, ignorando posibles efectos del cambio climático.

- SARIMA no optimizado: la parametrización fija de SARIMA limita su capacidad predictiva. Un SARIMAX optimizado por centro arrojaría resultados muy distintos.

- Ausencia de variables clave: el modelo no incorpora información sobre competencia local, ventas, ticket medio, campañas de marketing específicas o cambios en el *tenant mix*, todos ellos factores potencialmente relevantes.

- Horizonte de diez años: la incertidumbre acumulada hace que las proyecciones a diez años deban interpretarse como herramienta de planificación estratégica y no como pronóstico puntual; las proyecciones a cinco años son sustancialmente más fiables.

- Generalización limitada: el análisis se basa en una única cartera de doce centros gestionados por una misma empresa, lo que limita la extrapolación al conjunto del

sector.

6.3. Futuras líneas de investigación

A partir de las limitaciones identificadas y de los hallazgos del trabajo, se proponen las siguientes líneas de investigación:

1. Optimización individualizada de SARIMA/SARIMAX mediante procedimientos automáticos (*auto.arima*, búsqueda en rejilla con criterios AIC/BIC) y comparación rigurosa frente al Random Forest una vez ambos modelos parten de condiciones equivalentes.

2. Modelos de Gradient Boosting (XGBoost, LightGBM, CatBoost) que, según la literatura reciente, suelen superar a Random Forest en problemas de forecasting de retail con *feature engineering* enriquecido (Giannopoulos et al., 2025).

3. Modelos de Deep Learning para series temporales (LSTM, GRU, *Temporal Fusion Transformers*), que podrían capturar dependencias de largo plazo y estacionalidades múltiples de forma más nativa, especialmente sobre la serie horaria original.

4. Modelos jerárquicos y de reconciliación, que permitan integrar de forma coherente las predicciones a nivel horario, diario, mensual y anual, y a nivel centro y cartera (Hyndman & Athanasopoulos, 2021).

5. Integración de nuevas fuentes de datos**:** ventas, ticket medio, indicadores macroeconómicos del INE, datos de movilidad (Google Mobility, telcos), datos de competencia local e información sobre el *tenant mix* específico de cada activo.

6. Modelos causales y de inferencia, que permitan ir más allá de la predicción y estimar el impacto de palancas concretas de gestión (apertura de un nuevo operador ancla, reformas, campañas de marketing) sobre la afluencia.

7. Escenarios climáticos basados en proyecciones IPCC, que sustituyan las medias históricas por trayectorias coherentes con los escenarios de cambio climático, especialmente relevantes en el horizonte 2030-2035.

8. Extensión del estudio a una muestra multi-empresa que permita generalizar las conclusiones al conjunto del sector y comparar el comportamiento entre carteras con diferentes posicionamientos.

9. Desarrollo de un dashboard interactivo (Streamlit, Dash o similar) que ponga los resultados del modelo al alcance de equipos de gestión no técnicos y permita la simulación de escenarios en tiempo real.

7. Bibliografía

Asociación Española de Centros y Parques Comerciales. (2024, julio 10). Las ventas y la afluencia a los centros comerciales crecen hasta un 6 % en el primer semestre. *Forbes España*. https://forbes.es/economia/802406/las-ventas-y-la-afluencia-a-los-centros-comerciales-crecen-hasta-un-6-en-el-primer-semestre/

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: Forecasting and control* (5th ed.). Wiley.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Carbonneau, R., Laframboise, K., & Vahidov, R. (2008). Application of machine learning techniques for supply chain demand forecasting. *European Journal of Operational Research*, *184*(3), 1140–1154. https://doi.org/10.1016/j.ejor.2006.12.004

CBRE. (2025). *CBRE retail index – Abril 2025*. https://www.cbre.es/insights/reports/cbre-retail-index-abril-2025

Cleveland, R. B., Cleveland, W. S., McRae, J. E., & Terpenning, I. (1990). STL: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics, 6*(1), 3–73.

DataCamp. (s. f.). *Cómo calcular el RMSE en Python*. https://www.datacamp.com/es/tutorial/rmse

DataCamp. (s. f.). *Modelo ARIMA en Python: cómo hacer predicciones con series temporales*. https://www.datacamp.com/es/tutorial/arima

Fildes, R., Goodwin, P., & Nikolopoulos, K. (2019). Forecasting retail demand: Research and practice.
International Journal of Forecasting, *35*(3), 877–891.
https://doi.org/10.1016/j.ijforecast.2019.01.010

Fildes, R., Goodwin, P., Lawrence, M., & Nikolopoulos, K. (2009). Effective forecasting and judgmental adjustments: An empirical evaluation and strategies for improvement in supply-chain planning.
International Journal of Forecasting, *25*(1), 3–23. https://doi.org/10.1016/j.ijforecast.2008.11.010

Giannopoulos, P. G., Dasaklis, T. K., Tsantilis, I., & Patsakis, C. (2025). Machine learning algorithms in intermittent demand forecasting: A review. *International Journal of Production Research*. Advance online publication. https://doi.org/10.1080/00207543.2025.2578701

Healy, K. (2018). *Data visualization: A practical introduction*. Princeton University Press. https://socviz.co/

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer. https://hastie.su.domains/ElemStatLearn/

Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and practice* (3rd ed.). OTexts. https://otexts.com/fpp3/

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). The M4 Competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, *36*(1), 54–74. https://doi.org/10.1016/j.ijforecast.2019.04.014

Ministerio de Agricultura, Alimentación y Medio Ambiente. (2013). Centros comerciales en España: Situación, evolución e interpretación empírica. *Distribución y Consumo*, *127*, 5–21. https://www.mapa.gob.es/ministerio/pags/biblioteca/revistas/pdf_dyc/dyc_2013_127_5_21.pdf

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

Petropoulos, F., Makridakis, S., & Assimakopoulos, V. (2022). Forecasting in practice: The role of machine learning methods. *International Journal of Forecasting*, *38*(3), 1232–1248. https://doi.org/10.1016/j.ijforecast.2021.03.008

Probabilidad y Estadística. (s. f.). *Modelo ARIMA*. https://www.probabilidadyestadistica.net/modelo-arima/

Retlife. (2024, marzo 12). Footfall en centros comerciales: por qué esta métrica es clave

para atraer inversión rentable. https://retlife.es/footfall-en-centros-comerciales-por-que-esta-metrica-es-clave-para-atraer-inversion-rentable/

Shmueli, G., & Koppius, O. (2011). Predictive analytics in information systems research. *MIS Quarterly*, *35*(3), 553–572. https://doi.org/10.2307/23042796

Taylor, S. J., & Letham, B. (2018). Forecasting at scale. *The American Statistician*, *72*(1), 37–45. https://doi.org/10.1080/00031305.2017.1380080

Verhoef, P. C., Kannan, P. K., & Inman, J. J. (2015). From multi-channel retailing to omni-channel retailing: Introduction to the special issue on multi-channel retailing. *Journal of Retailing*, *91*(2), 174–
181. https://doi.org/10.1016/j.jretai.2015.02.005

Wilke, C. O. (2019). *Fundamentals of data visualization: A primer on making informative and compelling figures*. O'Reilly Media. https://clauswilke.com/dataviz/time-series.html