Trabajo Fin de Grado
Grado en Ingeniería Electrónica, Robótica y Mecatrónica

Desarrollo y Aplicación de Metodología de Modelado Predictivo de Series Temporales

Autor: Tomás Vázquez Ruiz

Tutores: Amparo Núñez Reyes

Fernando Pavón Pérez

Dpto. Ingeniería de Sistemas y Automática
Escuela Técnica Superior de Ingeniería
Universidad de Sevilla

Sevilla, 2023

Trabajo Fin de Grado

Grado en Ingeniería Electrónica, Robótica y Mecatrónica

Desarrollo y Aplicación de Metodología de Modelado Predictivo de Series Temporales

Autor:

Tomás Vázquez Ruiz

Tutores:

Amparo Núñez Reyes

Profesora Titular

Fernando Pavón Pérez

CEO Gamco S.L

Dpto. Ingeniería de Sistemas y Automática

Escuela Técnica Superior de Ingeniería

Universidad de Sevilla

Sevilla, 2023

Trabajo Fin de Grado: Desarrollo y Aplicación de Metodología de Modelado Predictivo de Series Temporales

Autor: Tomás Vázquez Ruiz

Tutores: Amparo Núñez Reyes
Fernando Pavón Pérez

El tribunal nombrado para juzgar el trabajo arriba indicado, compuesto por los siguientes profesores:

Presidente:

Vocal/es:

Secretario:

acuerdan otorgarle la calificación de:

El Secretario del Tribunal

Fecha:

Agradecimientos

Este trabajo ha sido posible gracias al proyecto PREDDICCO (Predicciones y decisiones dinámicas con control de contingencias y optimización de la toma de decisiones en base a modelos predictivos y simulación) EXP 00136057 / TIC-20200448, financiado por el CDTI (Centro para el Desarrollo Tecnológico Industrial). Durante la fase del proyecto: Simulación de datos reales, generación de conjuntos de datos y simulación de la respuesta del sistema. Donde se ha disfrutado de una beca de investigación en AICIA (Asociación de Investigación y Cooperación Industrial de Andalucía), dentro del proyecto "Generation of datasets created by the Generative Adversial Network (GAN) to train predictive models" con referencia: PI-2201/24/2022.

También me gustaría agradecer a mis padres y a mi hermana, quienes han estado a mi lado brindándome su cariño y apoyo inquebrantable desde siempre. A Javier Gil, uno de mis mejores amigos de la infancia. Coincidir con él en la carrera es lo mejor que me podría haber pasado y le agradezco de todo corazón toda la ayuda y el respaldo que me ha proporcionado a lo largo de estos últimos cuatro años. A mi grupo de amigos del Europa, a quienes voy a echar mucho de menos el año que viene cuando esté viviendo en Madrid.

Gracias a Fernando Pavón y a todo el equipo de GAMCO por su generosidad al compartir su valioso conocimiento y experiencia conmigo a lo largo de este proyecto, y en especial, a Álvaro Muñoz, quien no solo ha estado siempre dispuesto a solucionarme cualquier tipo de duda que tuviese, sino que lo ha hecho de la manera más amable y paciente que existe. Gracias por todo Álvaro.

Por último, pero no menos importante, quiero agradecer a mi tutora, Amparo Núñez Reyes. Gracias por brindarme la oportunidad de realizar la beca, y confiar en mi para formar parte de este proyecto.

Tomás Vázquez Ruiz

Sevilla, 2023

Resumen

La capacidad de anticiparse a eventos futuros nunca había estado tan a nuestro alcance como en el momento actual. El crecimiento exponencial experimentado por la Inteligencia Artificial en los últimos años ha impulsado la aplicación generalizada de algoritmos basados en Aprendizaje Automático, convirtiendo esta posibilidad en una realidad tangible y consolidándola como una herramienta poderosa que ha sido ampliamente adoptada por las empresas a nivel global. Los beneficios de anticipar el futuro son, como uno puede imaginar, incalculables, y las organizaciones están tomando cada vez más conciencia de esta ventaja.

Es por ello que en este Trabajo de Fin de Grado (TFG) se presenta una metodología de predicción de series temporales, desarrollada en colabaración con GAMCO S.L. que tiene como objetivo aportar una solución innovadora, fácil de implementar y que pueda adaptarse a una amplia variedad de contextos.

Esta metodología, basada en técnicas de minería de datos, utiliza como entradas los valores pasados de la serie temporal para predecir los valores futuros. Estas entradas, conocidas como entradas autorregresivas, se implementan en dos tipos de modelos predictivos: los modelos lineales de tipo ARX y los modelos no lineales de tipo RBF.

Este método se aplica a dos conjuntos de datos reales bastante diferentes entre si. En primer lugar, a una serie temporal de la demanda de energía eléctrica en los hogares de tres ciudades españolas, y en segundo lugar, se utilizará para tratar de predecir la serie temporal que conforman las transacciones de los dispensadores de un banco.

De esta manera, el objetivo principal de este trabajo consiste en evaluar la adaptación de nuestra metodología a dichas series, analizando a su vez, qué tipo de modelos se ajustan mejor a cada serie, para qué entradas, y cómo de buenas son las predicciones al comparar los resultados obtenidos en las diversas pruebas realizadas.

Abstract

The ability to anticipate future events has never been as within our reach as it is at the present moment. The exponential growth experienced by Artificial Intelligence in recent years has driven the widespread application of Machine Learning-based algorithms, turning this possibility into a tangible reality and establishing it as a powerful tool that has been widely adopted by companies globally. The benefits of anticipating the future are, as expected, incalculable, and organizations are increasingly aware of this advantage.

This is why in this Bachelor's Thesis (TFG), a methodology for time series prediction is presented, developed in collaboration with GAMCO S.L. The goal is to provide an innovative, easy-to-implement solution that can adapt to a wide variety of contexts.

This methodology, based on data mining techniques, uses past values of the time series as inputs to predict future values. These inputs, known as autoregressive inputs, are implemented in two types of predictive models: linear ARX models and non-linear RBF models.

This method is applied to two quite different real datasets. First, to a time series of household electricity demand in three Spanish cities, and second, it will be used to try to predict the time series formed by the transactions of a bank's ATMs.

In this way, the main objective of this work is to evaluate the adaptation of our methodology to these series, while also analyzing what types of models are better suited to each series, for which inputs, and how good the predictions are when comparing the results obtained in the various tests conducted.

Índice Abreviado

Índice

1 Introducción

La inteligencia artificial (IA) ha experimentado un crecimiento exponencial en las últimas décadas, y su aplicación en el ámbito empresarial se ha erigido como uno de los impulsores clave de esta revolución tecnológica. La capacidad de las máquinas para aprender, razonar y tomar decisiones de manera autónoma ha abierto un amplio abanico de posibilidades para mejorar la eficiencia y la productividad en las empresas de todo el mundo.

En sus primeras etapas, la IA en el ámbito empresarial se centraba principalmente en el análisis de datos y la generación de informes. Desde la década de los 90, con el avance de la tecnología informática y la acumulación masiva de datos, las empresas comenzarón a percatarse del potencial de esta tecnología para la automatización de tareas rutinarias, la optimización de procesos y la mejora de la toma de decisiones.

En el siglo XXI la IA se ha vuelto más sofisticada y ha ampliado su ámbito de aplicación en el entorno empresarial. En la actual era de la información, las empresas se encuentran ante un creciente volumen de datos generado a diario, los cuales albergan información de gran valor susceptible de ser aprovechada para la adopción de decisiones estratégicas y la mejora de la eficiencia en diversos sectores.

Actualmente, las organizaciones emplean soluciones basadas en IA para una amplia variedad de tareas, que engloban desde la atención al cliente, la gestión de inventario, el análisis de riesgos, el marketing digital, la ciberseguridad, la optimización de la cadena de suministro, hasta las predicciones de comportamiento del consumidor, entre otras.

Un ámbito en el cual la IA ha evidenciado un potencial significativo es la predicción de eventos futuros, particularmente **la predicción de series temporales**, temática que será abordada en este Trabajo de Fin de Grado (TFG).

Las series temporales se refieren a conjuntos de datos que evolucionan en función del tiempo, tales como las ventas mensuales, los precios de las acciones o los datos climáticos, entre otros ejemplos.

La habilidad para identificar patrones y tendencias en vastos conjuntos de datos que previamente resultaban arduos de discernir mediante métodos convencionales, ha resultado fundamental para las empresas en el proceso de toma de decisiones estratégicas y la planificación de recursos.

Mediante la implementación de algoritmos de aprendizaje automático, las empresas tienen la capacidad de aprovechar los datos históricos para entrenar modelos capaces de identificar patrones ocultos y efectuar predicciones precisas sobre el comportamiento futuro de una serie temporal. Estos modelos pueden considerar factores como la estacionalidad, tendencias, ciclos y eventos especiales, lo que conlleva a la generación de pronósticos más acertados.

A medida que la tecnología siga evolucionando, se espera que la IA desempeñe un papel aún más importante en la mejora de la eficiencia y la competitividad de las empresas en el futuro.

1.1 Motivación y objetivos del proyecto

El presente Trabajo de Fin de Grado (TFG) tiene como objetivo el desarrollo de una metodología de creación de dos tipos de modelos predictivos de aprendizaje automático: los modelos de tipo ARX (Auto-Regresión con Variables Exógenas) y los RBF (Redes Neuronales de Base Radial), así como su aplicación a dos series temporales, ambas con datos reales y sustancialmente dispares en cuanto a su naturaleza y características.

En una primera instancia, la metodología se emplea con la finalidad de pronosticar los valores correspondientes a la demanda de energía eléctrica en tres localidades de la región de Andalucía. Para llevar a cabo esta tarea, se utiliza una serie temporal histórica que comprende la potencia consumida en cada una de las 24 horas del día en dichas localidades. A lo largo del proceso, se concede una atención meticulosa a las particularidades inherentes a esta serie temporal, incluyendo las variables exógenas que puedan ejercer influencia sobre la misma, adaptando en consecuencia el enfoque y las técnicas aplicadas.

Una vez implementada y validada la metodología en la predicción de la demanda de energía eléctrica, se procede a su utilización en un contexto real de mayor complejidad. En esta fase, el propósito es anticipar los saldos relacionados con una institución bancaria, específicamente, los saldos de transacciones efectuadas en los dispositivos automáticos de una sucursal bancaria, haciendo uso de la serie temporal correspondiente.

Este entorno financiero representa un desafío mucho más complejo en comparación con el escenario previo, debido a la naturaleza no lineal de los datos y la influencia significativa de numerosas variables exógenas, las cuales serán cuidadosamente expuestas durante el proceso de análisis. Este enfoque nos brinda la oportunidad de evaluar minuciosamente la efectividad y capacidad de adaptación de la metodología propuesta.

Las utilidades y aplicaciones prácticas derivadas de esta investigación son de alcance considerable para la entidad financiera en cuestión. A continuación, se presentan dos de las aplicaciones más destacadas:

1. **Gestión de efectivo:** La predicción de saldos en los dispositivos permite a los bancos planificar y administrar de manera más efectiva el suministro de efectivo en sus sucursales. Al anticipar las necesidades futuras, el banco puede evitar situaciones problemáticas como la escasez o el exceso de efectivo, lo que garantiza que los clientes siempre tengan acceso a efectivo en los dispositivos automáticos.

2. **Optimización de la logística:** Conociendo con antelación la demanda esperada en los dispositivos automáticos, el banco puede optimizar las rutas y frecuencias de recarga y descarga de efectivo. Esto puede reducir costos logísticos y tiempos de espera en el reabastecimiento, mejorando la eficiencia operativa. Además, la optimización logística contribuye a la mejora de la calidad del servicio, ya que los tiempos de espera en el reabastecimiento se reducen, lo que beneficia tanto a los clientes como a la propia entidad financiera.

Al emplear técnicas avanzadas de aprendizaje automático y análisis de datos, se pueden desarrollar modelos predictivos que permitan hacer pronósticos precisos sobre cómo evolucionarán dichas series en el futuro.

La metodología de creación de modelos predictivos ofrece un enfoque sistemático y basado en el análisis de series temporales. Esta metodología combina técnicas y herramientas avanzadas de minería de datos, como el procesamiento de grandes volúmenes de información, la selección de variables relevantes y la implementación de algoritmos de aprendizaje automático. Además, en la etapa final se lleva a cabo una evaluación rigurosa de los modelos predictivos obtenidos, comparando su rendimiento mediante distintas métricas y analizando la precisión de las predicciones realizadas.

Todo esto ha sido llevado a cabo en colaboración con la empresa GAMCO S.L. gracias a una beca de investigación en AICIA (Asociación de Investigación y Cooperación Industrial de Andalucía), dentro del proyecto “Generation of datasets created by the Generative Adversial Network (GAN) to train predictive models", con Referencia: PI-2201/24/2022.

En conclusión, con este Trabajo de Fin de Grado (TFG), se pretende realizar una contribución significativa al campo de las predicciones en series temporales, y ofrecer una solución altamente efectiva y adaptable para tratar de predecir el comportamiento futuro de un fenómeno específico, así como explorar y comprender el comportamiento pasado y presente de dicho fenómeno.

1.2 Estado del Arte

La Inteligencia Artificial (IA) es una rama de la ciencia de la computación que se ocupa del desarrollo de algoritmos y sistemas que buscan emular la inteligencia humana y realizar tareas que, de otra manera, requerirían la intervención humana [1]. El objetivo principal de la IA es crear programas y máquinas capaces de aprender, razonar, planificar, percibir el entorno y tomar decisiones inteligentes [2].

Dentro de la IA, existen diversos enfoques y técnicas que permiten a las máquinas llevar a cabo tareas complejas. En este capítulo, nuestro enfoque se dirige principalmente hacia un campo particular de la IA que ha experimentado un crecimiento exponencial en aplicaciones y en popularidad en la última década: el **aprendizaje automático** (Machine Learning).

El aprendizaje automático es una rama de la Inteligencia Artificial que se enfoca en el desarrollo de algoritmos y técnicas que permitan a las máquinas aprender patrones y realizar predicciones o tomar decisiones basadas en datos sin una programación explícita [3]. En lugar de programar instrucciones específicas, el objetivo es que el sistema aprenda automáticamente de los datos y mejore su rendimiento con la experiencia [4].

El término "aprendizaje automático" se acuñó por primera vez en 1959 [5]. Se podría decir que

este campo es el resultado de la convergencia entre la informática y la estadística. La informática proporciona las técnicas de programación, arquitectura de software y las herramientas necesarias para el desarrollo de algoritmos complejos, fundamentales para implementar modelos de machine learning [6]. Mientras que la estadística juega también un papel crucial al proporcionar los fundamentos matemáticos que respaldan los algoritmos y las técnicas utilizadas. Conceptos como la inferencia estadística, la regresión, la probabilidad y la validación cruzada son esenciales para evaluar y mejorar el rendimiento de los modelos de machine learning.

Mientras que las ciencias de la computación por sí solas buscan resolver cómo programar manualmente sistemas informáticos, el aprendizaje automático se enfoca en permitir que las computadoras aprendan de manera autónoma. Y a diferencia de la estadística, cuyo objetivo es el análisis de datos, el aprendizaje automático implica implementar algoritmos que procesen automáticamente los datos para ejecutar acciones específicas.

En esencia, se trata de una simbiosis única entre la informática y la estadística, que aprovecha el potencial de las máquinas para adquirir conocimiento y tomar decisiones basadas en los datos recopilados.

Como se ha mencionado en las líneas anteriores, la relevancia y popularidad del campo del aprendizaje automático ha aumentado significativamente en los últimos años, y esto se debe a dos factores fundamentales: el aumento en la capacidad de computación y el crecimiento exponencial de los datos digitales (Big Data).

Big Data, en español "grandes volúmenes de datos", es un término utilizado para describir conjuntos de datos extremadamente grandes y complejos que superan la capacidad de las herramientas tradicionales de procesamiento y análisis de datos [7]. Estos conjuntos de datos se caracterizan por su volumen, variedad y rápido crecimiento, lo que representa un desafío significativo para su almacenamiento, gestión, procesamiento y análisis.

Este auge masivo de los datos que se ha producido en el siglo XXI ha proporcionado a las empresas una enorme cantidad de información que antes no estaba disponible. Los datos se han convertido en una valiosa materia prima para el aprendizaje automático, ya que estos modelos necesitan grandes cantidades de datos para aprender patrones y tomar decisiones precisas.

Por otro lado, el avance tecnológico en la capacidad de procesamiento ha permitido realizar cálculos complejos y entrenar modelos de aprendizaje automático en grandes conjuntos de datos de manera más eficiente y rápida [8]. Esto ha permitido desarrollar algoritmos más sofisticados y precisos que pueden abordar problemas cada vez más complejos y brindar soluciones más precisas. Como resultado, en los últimos años han surgido nuevas técnicas de aprendizaje automático que se han convertido en una parte fundamental del ecosistema del Big Data [9].

Una de las más relevantes es el **aprendizaje profundo** (Deep Learning). Se trata de una subárea del Machine Learning que se basa principalmente en **redes neuronales artificiales** para realizar tareas complejas de procesamiento de datos. Estas redes, inspiradas en la estructura del cerebro humano, son modelos matemáticos que permiten, entre otras cosas, realizar un aprendizaje automático eficiente y un reconocimiento de patrones en conjuntos de datos masivos [10].

La estructura básica de una red neuronal artificial se compone de nodos llamados neuronas, que se organizan en capas (Figura 1.1). Cada neurona recibe entradas, realiza cálculos matemáticos con estas entradas y produce una salida [11]. Las conexiones entre las neuronas se representan mediante "pesos", que determinan la importancia relativa de cada entrada en el cálculo realizado por la neurona [12].

Entre las capas principales de una red neuronal artificial se incluyen:

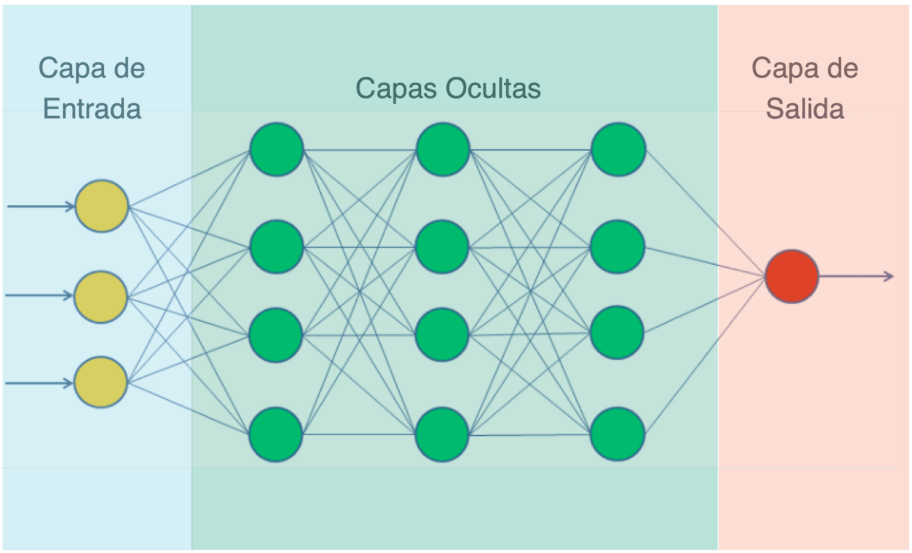


Figura 1.1 Arquitectura de una red neuronal artificial.

1. **Capa de entrada:** Es la primera capa de la red y se encarga de recibir los datos de entrada. Cada neurona de esta capa representa una característica del dato.
2. **Capas ocultas:** Estas capas intermedias realizan cálculos complejos utilizando los pesos y las entradas de la capa anterior y permiten que la red aprenda representaciones jerárquicas y características complejas de los datos. El término "oculto" se refiere a que estas capas no interactúan directamente con el mundo exterior y suelen ser invisibles para el usuario. La cantidad de capas ocultas y el número de neuronas en cada capa pueden variar según la arquitectura de la red neuronal.
3. **Capa de salida:** Es la última capa de la red y su función es producir las salidas finales de la red neuronal en función de los cálculos realizados en las capas ocultas..

La capacidad de una red neuronal para aprender y adaptarse a los datos se basa en el proceso de **entrenamiento**, en el cual los pesos se ajustan iterativamente para minimizar una función de error, lo que permite a la red aprender patrones y realizar predicciones precisas. El objetivo es que la red pueda generalizar y realizar predicciones precisas en datos nunca antes vistos.

El aprendizaje profundo ha demostrado ser especialmente eficaz en una amplia variedad de aplicaciones, como el procesamiento de imágenes y videos, el procesamiento del lenguaje natural, la visión por computadora, la traducción automática y la clasificación de datos entre otras. Sin embargo, en este trabajo, como ya se ha mencionado anteriormente, el enfoque principal se centrará en la **predicción de series temporales**.

En términos simples, la predicción de series temporales es una disciplina dentro del aprendizaje automático que se enfoca en prever valores futuros de una variable en función de su comportamiento pasado en el tiempo.

En la literatura se pueden encontrar multitud de metodologías y técnicas propuestas para la predicción de series temporales. Entre las más populares se incluyen los modelos ARIMA (AutoRegressive Integrated Moving Average) (Box & Jenkins, 1970), los métodos de suavizado exponencial, las redes neuronales recurrentes (RNN) (Hochreiter & Schmidhuber, 1997), y los métodos de Aprendizaje Profundo, como las redes neuronales LSTM (Long Short-Term Memory) o las redes neuronales feed-forward (FNN).

A continuación, se presenta una breve descripción de los principios en los que se sustentan los métodos más destacados, desde los más históricos hasta los más contemporáneos y avanzados:

- **Modelos ARIMA**

  Desde la década de 1960, el campo de la predicción ha estado dominado por enfoques estadísticos lineales, como los modelos ARIMA (AutoRegressive Integrated Moving Average). Estos modelos se componen de tres partes: autorregresión (AR), media móvil (MA) e integración (I). El componente AR implica que el valor actual se relaciona con valores previos a través de una regresión lineal. El componente MA se basa en errores pasados para predecir el valor actual. La integración se utiliza para hacer que la serie sea estacionaria, eliminando tendencias y estacionalidad.

- **Promedio Ponderado Exponencialmente (Exponentially Weighted Moving Average):**

  Esta técnica asigna pesos exponenciales a los valores pasados de una serie temporal, enfocándose más en los valores recientes [13]. Esto la hace efectiva para identificar tendencias o patrones emergentes. Es eficiente computacionalmente y se adapta a series temporales de diferentes longitudes y frecuencias.

- **Suavizado Holt-Winters:**

  Se trata de una extensión del Promedio Ponderado Exponencial. Se utiliza especialmente para abordar datos que exhiben patrones cíclicos o estacionales. Se compone de tres elementos esenciales: el nivel, la tendencia y la estacionalidad [14]. Estos componentes se combinan de manera aditiva o multiplicativa, dependiendo de la naturaleza de la serie temporal, y se ajustan de manera iterativa para mejorar la precisión de las predicciones.

- **Suavizado Exponencial (Exponential Smoothing):**

  Similar al Promedio Ponderado Exponencial, utiliza ponderación exponencial para valores pasados, dando más importancia a los recientes [15]. Se aplica iterativamente para predecir valores futuros, siendo efectivo en la predicción de tendencias.

Sin embargo, en los últimos tiempos, los modelos de aprendizaje automático han ganado relevancia y se han convertido en los referentes principales en el campo de predicción de series. Numerosos estudios han demostrado la superioridad clara de estos algoritmos frente a los métodos estadísticos tradicionales [16]. Algunos de los más relevantes son:

- **Redes Neuronales Recurrentes**

  Se trata de un tipo de modelo de aprendizaje profundo diseñado específicamente para trabajar con datos secuenciales, como series temporales [17]. A diferencia de las redes neuronales tradicionales, las RNN tienen conexiones recurrentes que les permiten mantener y utilizar información de elementos anteriores de una secuencia [18].

  La arquitectura básica de una RNN consiste en unidades recurrentes (celdas) que se activan de manera secuencial a lo largo de la serie temporal (Figura 1.2). Cada celda recibe de la iteración anterior una entrada y un estado oculto (hidden state) donde se almacena la información importante de iteraciones anteriores. Con esta información se produce una salida y un nuevo estado oculto que se utilizará en la siguiente iteración. Esto es especialmente útil en la predicción de series temporales, donde las observaciones en el pasado pueden tener un impacto significativo en los valores futuros.

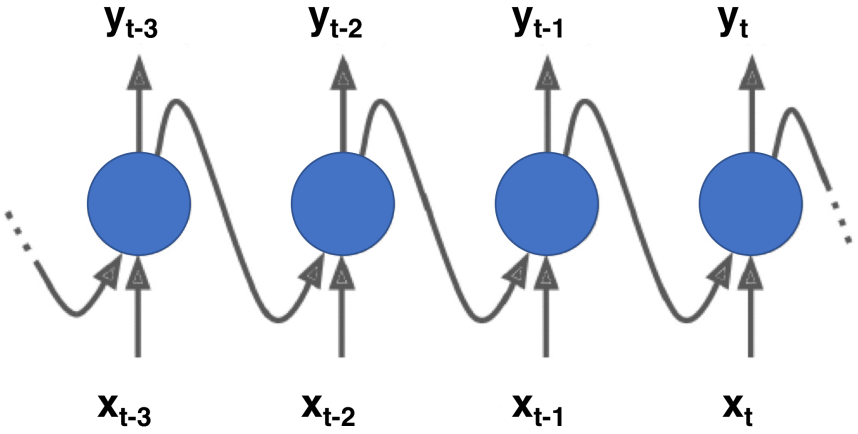


Figura 1.2 Arquitectura de una red neuronal recurrente.

Sin embargo, en las RNN tradicionales pueden aparecer problemas al utilizarse en secuencias de datos que son muy largas. Esto se conoce como el "problema de la dependencia a largo plazo" el cual se divide en dos vertientes: el problema del “desvanecimiento del gradiente” y el de la “explosión del gradiente" [19].

El gradiente, en este contexto, es una medida que indica cómo cambia la función de error en relación con los pesos de la red neuronal. Esta función de error es una medida que evalúa qué tan bien está realizando una red neuronal sus predicciones, es decir, representa la diferencia entre las predicciones del modelo y los valores reales del conjunto de entrenamiento [20].

El gradiente muestra por tanto la dirección y la magnitud del cambio en la función de pérdida cuando los pesos de la red se ajustan ligeramente.

Sin embargo, durante el entrenamiento, los gradientes se multiplican repetidamente a medida que retroceden en el tiempo para ajustar los pesos, por lo que si este es menor que 1 puede disminuir exponencialmente (desvanecimiento del gradiente) hasta que eventualmente se vuelven tan pequeños que la red tiene dificultades para actualizar los pesos de las capas

anteriores y aprender relaciones a largo plazo.

O puede darse el caso contrario, en el que el gradiente al ser mayor que 1, aumente exponencialmente al multiplicarse repetidamente (explosión del gradiente) y provoque que los pesos de la red crezcan desmesuradamente, afectando la estabilidad del modelo y dificultando el aprendizaje.

Esta divergencia o convergencia rápida de los gradientes puede hacer que la RNN tenga serias dificultades para capturar patrones a largo plazo en series temporales. Para abordar estas limitaciones, se han desarrollado variantes más avanzadas de RNN, como las redes neuronales LSTM (Long Short-Term Memory) o las redes neuronales de tipo feed-forward (FNN).

- **LSTM**

Las LSTM (Long Short-Term Memory) son un tipo de red neuronal recurrente (RNN) diseñada para abordar el problema de la memoria a largo plazo en el procesamiento de secuencias de datos. Fueron introducidas por Hochreiter y Schmidhuber en 1997 como una mejora significativa sobre las RNN tradicionales [21].

Las LSTMs abordan este problema incorporando una estructura más compleja conocida como "celda de memoria" o "celda LSTM". Estas actúan como una especie de memoria interna que se puede leer, escribir y borrar de manera controlada. Para ello tienen una estructura de puertas que regulan el flujo de información, lo que les permite recordar u olvidar información relevante de la secuencia.

Los 3 tipos de puertas en una celda LSTM son:

1. Puerta de entrada (input gate): Decide qué nueva información se va a agregar a la memoria de la celda.
2. Puerta de olvido (forget gate): Decide qué información existente en la memoria de la celda se va a descartar.
3. Puerta de salida (output gate): Decide qué información almacenada en la celda se va a usar como salida de la LSTM.

Al aprender a mantener y actualizar la información relevante en la celda de memoria, las LSTM pueden capturar patrones a largo plazo en secuencias, lo que las hace muy efectivas en tareas de predicción y modelado de series temporales largas.

Además, las redes LSTM han establecido récords en los campos de reconocimiento del habla con vocabulario extenso, traducción automática, modelado de lenguaje y procesamiento de lenguaje multilingüe [22].

En la actualidad, se han llevado a cabo diversos estudios en el ámbito de la predicción de series temporales empleando algoritmos basados en LSTM. Un ejemplo reciente es el trabajo llevado a cabo por Chandra, R., Jain, A., y Chauhan, D. S. (2022), titulado "Aprendizaje profundo a través de modelos LSTM para la predicción de infecciones por COVID-19 en India" [23], publicado el 28 de enero de 2022. Este estudio evaluó una variedad de modelos LSTM univariados y multivariados junto con distintos enfoques para la elección de conjuntos de datos de entrenamiento y prueba.

Las variantes del modelo LSTM evidenciaron fortalezas y debilidades específicas en diversas situaciones, lo que complicó la elección de un modelo único. En términos generales, se determinó que el modelo univariado ED-LSTM con división aleatoria exhibió el rendimiento más sobresaliente en las pruebas, en comparación con el resto de los modelos evaluados. No obstante, los resultados subrayan las dificultades inherentes a la predicción con datos limitados, particularmente en un contexto marcado por sesgos significativos, debido a los dos picos destacados en el transcurso de la pandemia en India.

- **Perceptrón Multicapa o Red Neuronal Feedforward (FNN):**

  Se trata de un algoritmo de aprendizaje profundo que ha manifestado su eficacia de manera destacada en el ámbito de la predicción de series temporales [24]. Este enfoque ha ganado notoriedad en el contexto contemporáneo gracias a su capacidad para modelar relaciones de alta complejidad en datos secuenciales, consolidándose como una herramienta de gran valía en la predicción de series temporales.

  En su formulación elemental, el Perceptrón Multicapa se compone de una capa de entrada, una o múltiples capas ocultas, y una capa de salida. Cada capa alberga nodos o neuronas interconectadas. En el contexto de la predicción de series temporales, la capa de entrada se emplea comúnmente para representar observaciones pasadas, mientras que la capa de salida se encarga de proporcionar las predicciones para el futuro. Las capas ocultas revisten una importancia esencial, al permitir el aprendizaje y modelado de patrones intrincados en los datos secuenciales [25].

  El rasgo distintivo y beneficioso de los Perceptrones Multicapa radica en su capacidad de capturar y representar las dependencias a largo plazo existentes en la serie temporal mediante el mecanismo de propagación hacia atrás de gradientes (backpropagation) durante el proceso de entrenamiento [26]. Este procedimiento viabiliza la adaptación de los pesos de las conexiones neuronales con el objetivo de minimizar el error de predicción.

  Numerosos estudios han corroborado el éxito de los Perceptrones Multicapa en la predicción de series temporales. A modo de ejemplo, en un estudio llevado a cabo por Saied S. Sharif1 y James H. Taylor (2000), se aplicó un conjunto de 24 redes neuronales feedforward (FNN) para reducir el error en la predicción de la carga eléctrica por hora del día siguiente [27]. Es decir, se utilizaron 24 redes independientes para predecir la carga de las siguientes 24 horas. Las variables de entrada para las FNN se se seleccionaron a partir de tres categorías de relevancia: 1) variables de carácter calendario, 2) variables de índole meteorológica, y 3) variables concernientes a datos de carga eléctrica.

  Los resultados derivados de las simulaciones exhibieron que la aproximación basada en FNN logró reducir el error absoluto medio porcentual (MAPE) en la predicción de la carga por hora en una proporción superior al 30 % en comparación con el mejor modelo hasta la fecha.

  Otro ejemplo ilustrativo se halla en un estudio conducido por Sina E. Charandabi y Kamyar Kamyar en el año 2021, en el cual se empleó una Red Neuronal Feedforward para anticipar los precios de diversas criptomonedas [28]. Las conclusiones de este estudio apuntaron a que el algoritmo FNN se presentó como adecuado en la mayoría de los casos para predecir los valores de las criptomonedas, incluso en un contexto caracterizado por la inusitada volatilidad que experimentó el mercado de la mayoría de las criptomonedas contempladas en el marco temporal de la investigación.

La elección de la mejor técnica depende de numerosos factores, como la cantidad y naturaleza de los datos, la presencia de tendencias y estacionalidades, la complejidad del problema y los recursos computacionales disponibles. A menudo, se requiere experimentación y ajuste para determinar la técnica óptima para una tarea de predicción de series temporales específica. Es importante también tener en cuenta que el campo de la inteligencia artificial está en constante evolución, lo que conlleva la necesidad de mantenerse actualizado y estar dispuesto a adoptar nuevas técnicas y enfoques a medida que estos se desarrollen.

1.3 Estructura del Documento

Este documento se estructura en cinco capítulos, cada uno de los cuales aborda aspectos específicos relacionados con la metodología y las aplicaciones de la misma a las series temporales de estudio. A continuación, se presenta un resumen de cada capítulo:

- **Capítulo 1: Introducción**

  Este capítulo establece el contexto general del trabajo y presenta una visión general de los objetivos, la relevancia y el alcance de la investigación. Se proporciona una introducción a los conceptos clave y se repasa brevemente la literatura.

- **Capítulo 2: Metodología Empleada**

  En este capítulo, se detalla de manera exhaustiva la metodología utilizada en el estudio. Se describen las técnicas de análisis y los enfoques de aprendizaje automático empleados para el análisis de series temporales. También se explican los procedimientos de recopilación, preprocesamiento y definición de entradas, así como la configuración de los distintos tipos de modelos predictivos.

- **Capítulo 3: Aplicación a Serie Temporal de la Demanda del Consumo Eléctrico**

  Este capítulo se centra en la aplicación de la metodología desarrollada a una serie temporal específica: la demanda de consumo eléctrico en hogares. Se detallan los resultados obtenidos, se presentan gráficos y se analizan las relaciones entre las variables de entrada y la salida. Se discuten las implicaciones y los hallazgos relacionados con esta serie temporal.

- **Capítulo 4: Aplicación a Serie Temporal de las Transacciones Bancarias**

  Similar al capítulo anterior, en este se aborda la aplicación de la metodología a otra serie temporal, en este caso, las transacciones bancarias. Se analiza la serie, se definen las entradas a utilizar en los modelos, se analizan las relaciones entre las variables exógenas y la serie principal, se presentan los resultados, y se evalúa el rendimiento de los modelos predictivos.

- **Capítulo 5: Conclusiones**

  El último capítulo resume las conclusiones clave del estudio. Se destacan los hallazgos más relevantes, las implicaciones prácticas y las limitaciones del enfoque utilizado.

2 Metodología empleada

La metodología en la que se ha trabajado en colaboración con GAMCO S.L, con el objetivo de lograr la predicción de series temporales relacionadas con entidades bancarias con la máxima precisión posible, se fundamenta en gran medida en la disciplina de la minería de datos.

La Minería de Datos se trata de una disciplina cuya finalidad radica en la identificación de patrones, tendencias y relaciones significativas en extensos conjuntos de datos. Mediante la aplicación de una variedad de técnicas analíticas, se logra la extracción de conocimientos de valor intrínseco y con potencial utilidad, los cuales pueden ser empleados para la toma de decisiones fundamentadas o la mejora de procesos en diversas áreas.

La Minería de Datos implica varias etapas, que incluyen:

1. **Recopilación de Datos:** es la primera fase donde se obtienen y se recopilan los datos necesarios para el análisis. Estos datos pueden provenir de diversas fuentes, como bases de datos, registros transaccionales, sensores, redes sociales, entre otros.

2. **Preprocesamiento de Datos:** antes del análisis, los datos suelen requerir limpieza y transformación para eliminar valores atípicos, datos faltantes o ruidos que puedan afectar los resultados.

3. **Análisis y Selección de Variables:** esta etapa implica examinar en detalle los datos recopilados, identificar patrones, relaciones y tendencias. A través de diversas técnicas estadísticas y herramientas de visualización, se busca comprender la información subyacente en los datos y extraer conocimientos valiosos. En esta etapa, se eligen las variables más relevantes o significativas del conjunto de datos que se utilizarán en el modelado, lo que puede ayudar a reducir la complejidad y mejorar el rendimiento del modelo.

4. **Modelado:** se aplican técnicas de inteligencia artificial, como algoritmos de aprendizaje automático, para realizar modelos predictivos.

5. **Evaluación e Interpretación de Resultados:** se evalúa el rendimiento del modelo y se interpretan los resultados obtenidos para obtener conocimientos útiles y aplicables.

La combinación de técnicas de aprendizaje automático con la minería de datos constituye el cimiento de nuestro enfoque metodológico. En las siguientes secciones, se brindará una explicación

detallada de cómo se han abordado las distintas etapas mencionadas en las anteriores líneas durante el desarrollo de nuestro proyecto.

Todas las fases de nuestra metodología se han implementado mediante el uso de scripts en el entorno de programación MATLAB® desarrollados en colaboración con GAMCO. No obstante, es importante señalar que, debido a que estos códigos pertenecen a GAMCO, no se incluirán sus contenidos en un anexo final en este Trabajo de Fin de Grado. En su lugar, se proporcionan resúmenes descriptivos de las principales funcionalidades y enfoques utilizados en dichos scripts, asegurando así la comprensión de la metodología empleada.

2.1 Recopilación de datos

Los datos relacionados con la serie temporal del consumo eléctrico y los datos asociados a la entidad bancaria provienen de proyectos previamente llevados a cabo por GAMCO, y, en consecuencia, son las propias empresas que contrataron los servicios de GAMCO las que suministraron estos datos.

La primera serie temporal de la demanda eléctrica data de hace más de 20 años y consta de un fichero principal que contiene los datos correspondientes a la potencia consumida en cada una de las 24 horas de un día en 3 ciudades andaluzas.

Los datos bancarios corresponden a un proyecto actual en desarrollo por GAMCO, y por lo tanto, no se revelarán nombres ni detalles específicos. Además, todos los datos utilizados en las pruebas presentadas en este TFG estarán normalizados.

Con el objetivo de mantener un alcance manejable, el foco principal de este trabajo han sido los archivos de datos que conforman las series temporales de las transacciones de los diferentes dispositivos automáticos del banco, concretamente de los dispensadores de una de las sucursales bancarias.

Un **dispensador** es un dispositivo automático que permite a las personas retirar dinero sin la necesidad de acudir a una ventanilla de atención al cliente. Cada retiro de dinero se corresponde con una transacción. Por consiguiente, los hábitos y horarios de las personas para realizar transacciones en los dispensadores tendrán un impacto significativo en la serie temporal. Las tendencias de retiros pueden variar según diferentes factores, como los días de la semana, los períodos del mes, las festividades, las horas del día y más.

Precisamente, este es el desafío principal que se abordará mediante el modelaje predictivo en esta serie. El objetivo principal será comprender y capturar los patrones inherentes en los hábitos y horarios de retiro de dinero en los dispensadores.

2.2 Preprocesamiento de los datos

En lo que respecta a la preparación inicial de los datos en bruto, se ha seguido un procedimiento que involucra la conversión de los archivos originales en formato CSV al formato MAT. Esta conversión se lleva a cabo con el propósito de facilitar su manipulación y utilización en el entorno de Matlab de manera más eficaz.

La mayoría de los ficheros originales se encuentran en formato CSV, un formato de archivo que se utiliza comúnmente para almacenar datos tabulares en forma de texto plano. En un archivo CSV, los datos se estructuran en filas y columnas, donde cada fila corresponde a una unidad de la serie temporal, y los campos de las columnas contienen información adicional que variará según la serie temporal.

A modo de ejemplo, en el caso de la serie temporal de transacciones bancarias que exploraremos, cada fila corresponderá a una transacción efectuada, y sus respectivas columnas contendrán detalles tales como la fecha de la transacción, la cantidad de dinero involucrada, junto con otros datos relevantes propios del banco. A través de esta estructura, cada operación se encuentra exhaustivamente definida a partir de la información consignada en las columnas de la misma fila.

Como es evidente, la información presente en cada fila experimentará variaciones según el tipo de serie temporal, no obstante, todas deben mantener una característica común esencial: el **campo temporal**. Asimismo, es de suma importancia que los datos se encuentren organizados en orden cronológico, desde los eventos más antiguos hasta los más recientes. Una de las tareas inherentes al proceso de preprocesamiento es, precisamente, garantizar esta disposición ordenada de los datos.

Una vez que los datos han sido dispuestos en orden cronológico, se procederá a llevar a cabo un proceso de filtrado y limpieza de los mismos, con el objetivo de garantizar su calidad y coherencia.

Finalmente, cada archivo CSV original será convertido en un archivo MAT, que contendrá las columnas consideradas relevantes del archivo original. Esta conversión de formatos se realizará mediante un script de Matlab diseñado para aplicar un proceso de mapeo a los campos del archivo original que se encuentren en formato de texto hacia un formato numérico. La elección y documentación meticulosa de este mapeo reviste gran importancia, ya que en el futuro será crucial comprender el significado detrás de cada número asignado.

Todo este proceso de conversión es esencial, dado que las etapas posteriores de nuestra metodología se desarrollarán también mediante scripts de Matlab. La conversión a formato MAT se erige como un paso crucial para garantizar la compatibilidad y la eficacia de las operaciones posteriores en nuestro enfoque metodológico.

2.3 Análisis de la serie

En lo que respecta a la fase de análisis, es crucial subrayar la aplicación de dos enfoques diferenciados y complementarios. El primero de estos enfoques se orienta hacia el aspecto estadístico de los datos, buscando entender las características fundamentales de la serie temporal, como tendencias, estacionalidades y ciclos. Este enfoque estadístico proporciona una visión inicial de la estructura de los datos y ayuda a identificar patrones que pueden ser relevantes para la modelización posterior.

El segundo enfoque, por otro lado, se enfoca en la identificación de correlaciones lineales entre la salida del modelo y sus posibles entradas. Este análisis busca establecer relaciones directas entre las variables de entrada y la variable objetivo que se pretende predecir. Este proceso brinda una primera idea de qué variables de entrada tienen una influencia significativa en el comportamiento de la serie temporal y, por lo tanto, son candidatas a ser utilizadas en nuestros modelos de predicción.

La combinación de estos dos enfoques en la fase de análisis permite una comprensión profunda y completa de los datos, ayudando a identificar tanto patrones estadísticos como relaciones causales que pueden ser cruciales para la construcción de modelos predictivos precisos y efectivos. Estos análisis preliminares sirven como cimientos sólidos sobre los cuales se basará el diseño y la implementación de nuestros modelos de predicción.

2.3.1 Análisis gráfico y estadístico

En primer término, se realiza un análisis estadístico exhaustivo de la serie temporal con el propósito de identificar información crucial, como la detección de períodos en los que no existen registros de datos, la distribución de operaciones en función del período de tiempo y otros aspectos pertinentes que dependerán de las características propias de la serie temporal bajo estudio.

Junto a este enfoque estadístico, se llevará a cabo un análisis gráfico adicional en esta etapa. Esto implica la representación visual de la serie temporal y la creación de otras representaciones gráficas que se consideren pertinentes para una comprensión más profunda de la serie. A través de la visualización de datos, se podrán identificar con eficacia patrones, comportamientos atípicos y relaciones entre variables que podrían no ser evidentes mediante el análisis estadístico por sí solo.

Este análisis gráfico complementa y enriquece el enfoque estadístico al proporcionar una comprensión más holística de la serie temporal. Juntos, ambos enfoques nos brindan una visión completa de los datos.

Una vez finalizada esta fase de análisis de la serie en cuestión, se da paso al estudio de las correlaciones como parte del proceso de selección de las variables de entrada para los modelos de predicción.

2.3.2 Análisis de correlaciones

Se ha desarrollado un script en Matlab encargado de calcular la correlación de Pearson entre la variable objetivo a predecir y los datos correspondientes a los instantes de tiempo previos. En este análisis, emplearemos el coeficiente de correlación de Pearson, a menudo referido simplemente como "correlación de Pearson". Esta medida estadística cuantifica la relación lineal entre dos conjuntos de datos numéricos. Es una métrica ampliamente utilizada para evaluar la fuerza y la dirección de la relación entre dos variables continuas. El coeficiente de correlación de Pearson varía entre -1 y 1, donde:

- 1 indica una correlación positiva perfecta: cuando una variable aumenta, la otra también lo hace en una proporción constante.
- -1 indica una correlación negativa perfecta: cuando una variable aumenta, la otra disminuye
- 0 indica que no hay correlación lineal entre las variables.

La fórmula que hemos implementado en nuestro script para calcular el coeficiente de correlación de Pearson entre dos variables X e Y es:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2.1}$$

Donde:

- n es el número de instantes.
- x_i es el valor de la variable X en el instante i.
- $\bar{x}$ es la media de los valores de la variable X.
- y_i es el valor de la variable Y en el instante i.
- $\bar{y}$ es la media de los valores de la variable Y.

Es importante tener en cuenta que la correlación de Pearson solo mide la relación lineal, por lo que puede no capturar relaciones no lineales entre las variables.

Una relación no lineal en el contexto de series temporales se refiere a patrones y comportamientos que no se pueden modelar o entender adecuadamente utilizando una línea recta.

Por ejemplo, en términos financieros, una relación no lineal podría manifestarse de varias maneras:

1. **Curvas y patrones no lineales:** Los datos podrían seguir patrones no lineales, como curvas, ondas o formas complejas, que no se ajustan a una relación directamente proporcional o inversamente proporcional.
2. **Efectos umbral:** Podrían existir umbrales en los datos, donde la relación entre las variables cambia abruptamente en ciertos puntos. Estos efectos no se capturan adecuadamente mediante la correlación de Pearson.
3. **Relaciones asimétricas:** Las relaciones financieras pueden ser asimétricas, lo que significa que los cambios en una variable pueden afectar de manera diferente a la otra variable según su magnitud o dirección.
4. **Interacciones complejas:** Las interacciones entre diferentes factores financieros pueden ser intrincadas y no lineales, lo que puede dificultar la captura completa de estas relaciones con la

correlación de Pearson.

No obstante, el enfoque de nuestro análisis en esta etapa se centra en comprender cómo los datos están relacionados entre sí en distintos momentos temporales. A través de este proceso, podemos representar gráficamente el valor de la correlación lineal entre la variable a predecir y los valores de tiempos pasados, en función del desfase temporal.

Las correlaciones fuertes sugieren una relación lineal entre las variables y, por lo tanto, indican que estas variables son candidatas prometedoras para ser incluidas en los modelos de predicción.

Este análisis de correlaciones desempeña un papel esencial en la selección de las variables de entrada, ya que ayuda a enfocar el proceso de modelización en las variables que tienen un mayor potencial predictivo. Así, se garantiza que los modelos estén construidos de manera eficiente y sean capaces de capturar las relaciones más significativas dentro de la serie temporal.

2.4 Definición de entradas

Se identifican múltiples tipos de variables de entrada que serán integradas en nuestros modelos predictivos, las cuales pueden ser definidas de manera fundamentada gracias a la fase anterior de análisis.

En una primera instancia, en el caso de cualquier serie temporal, nuestro enfoque se dirige hacia la implementación de "**Entradas Autorregresivas**". Como sugiere su nombre, estas entradas utilizan los valores previos de la serie temporal para anticipar valores futuros. La autorregresión implica que los valores previos de la serie tienen importancia para predecir sus valores subsiguientes. Es decir, modelamos la serie temporal en relación con su propio historial.

Las entradas de tipo autorregresivo utilizan comúnmente la notación "*k+d*", donde "*k*" denota el instante actual y "*d*" el desfase temporal. Como es evidente en este contexto, las entradas autorregresivas presentarán un desfase negativo, ya que se refieren a instantes pasados, y las salidas un desfase positivo, pues siempre se referirán a instantes futuros.

Por ejemplo, si se considera una entrada con un desfase de k-14 y se desea predecir la salida k+7, se estaría empleando el valor de la serie temporal que ocurrió hace 14 pasos en el pasado para anticipar el valor que se presentará 7 pasos en el futuro.

A lo largo de este TFG, se usará de manera intercambiable los términos "desfase" y "entrada", ya que estos desfases temporales se convierten en las entradas autorregresivas de nuestros modelos.

Como se ha explicado previamente en la sección anterior, la elección de los desfases temporales que serán utilizados como entradas en los modelos se realizará mediante el análisis de correlaciones lineales. Se seleccionarán aquellos desfases temporales que muestren una correlación más alta con la salida a predecir, y se llevará a cabo una inclusión gradual de estos desfases a través de un proceso iterativo. El objetivo es encontrar la combinación óptima de desfases de entradas que resulte en los mejores niveles de precisión en las predicciones.

Los pasos de este método de inclusión progresiva de entradas se detallarán en la última sección de este capítulo, una vez estén explicados los procesos de modelado y evaluación.

Además de considerar los instantes temporales previos, es fundamental tener en cuenta la posibilidad de incorporar otras series temporales relacionadas con la serie que se está tratando de predecir, o considerar características relevantes derivadas de la propia serie principal que puedan contribuir a anticipar su comportamiento futuro. A estas variables se les denomina "**Entradas Exógenas**". Estas entradas pueden surgir de series temporales externas a la serie que se está prediciendo, o bien, pueden derivar de cálculos realizados sobre la propia serie original. Por ejemplo, calcular la media de los datos de los 5 momentos de tiempo previos a "k+d" sería un ejemplo de cálculo basado en la serie original.

Otra forma en que las entradas exógenas se pueden presentar al modelo consiste en asignar diferentes identificadores a los diversos tipos de días identificados en nuestros análisis. Por ejemplo, en el caso de las transacciones bancarias, si se sabe que en ciertos días de la semana o meses existe una mayor demanda de retiros de efectivo debido a pagos de salarios, eventos o vencimientos de facturas, esta información puede ser utilizada para ajustar las predicciones de manera adecuada.

Se profundizará en el concepto de variables exógenas en los capítulos siguientes, ya que cada tipo de serie tendrá sus propias variables exógenas específicas que requerirán una explicación detallada.

2.4.1 Notación empleada

Para este TFG, se ha utilizado una notación sencilla y accesible a la hora de definir las entradas y salidas con el propósito de facilitar su comprensión. Todas las entradas utilizadas en este trabajo se encuentran detalladamente recopiladas en el glosario ubicado al final del documento.

Para identificar una entrada, se utilizará una letra que identifique la serie temporal que estemos tratando en cuestión, seguida de un subíndice que caracterice el tipo de entrada, seguido a su vez, de @ y el desfase seleccionado.

Letra Subíndice @ Desfase

Se ha asignado la letra "**A**" a la serie temporal de la **demanda eléctrica**, y la letra "**B**" a las **transacciones bancarias de los dispensadores**. La notación de las entradas exógenas específicas de cada serie se abordará en sus respectivos capítulos, sin embargo, habrá dos tipos de entradas exógenas comunes a ambas series: las entradas de tipo **calendario**, a las que se les ha asociado la letra "**C**", y las entradas **derivadas**, que utilizan la letra "**D**". Más adelante se explicará en que consisten estos dos tipos de entradas.

Por lo general, las entradas autorregresivas llevarán el subíndice "**1**", aunque más adelante se verán algunas variaciones. Por ejemplo, si se desea representar la entrada autorregresiva de desfase -14 de la serie de la demanda eléctrica para predecir la salida k+7, se escribiría de la siguiente manera:

Entrada: A1@-14

Salida: A1@7

Como se mencionó anteriormente, es importante recordar que en el caso de las variables autorregresivas, se utilizará un desfase positivo o negativo, según se trate de una salida a predecir o una entrada. Sin embargo, en el caso de las entradas exógenas de tipo calendario y las derivadas, se utilizará un desfase positivo, coincidente con el de la salida correspondiente, ya que su función será la de proporcionar información relevante sobre la salida que se busca predecir, Se abordará con mayor profundidad estos aspectos en los capítulos correspondientes a la aplicación de la metodología.

2.5 Modelado

Una vez finalizadas las etapas de recopilación y preprocesamiento de datos, así como el análisis y la selección de entradas, se puede dar paso al proceso de modelado. En esta fase, se hará uso de técnicas de inteligencia artificial, en particular, algoritmos de aprendizaje automático, para desarrollar modelos predictivos fundamentados en los datos que se han recopilado y preparado previamente.

Los pilares fundamentales de nuestra metodología se centran en dos algoritmos principales, los cuales servirán como piedra angular para la construcción de nuestros modelos predictivos: el enfoque lineal de **Auto-Regresión con Variables Exógenas (ARX)** y las **Redes Neuronales de Base Radial (RBF)**. Sin embargo, antes de adentrarnos en los aspectos específicos de cada tipo de modelo, es esencial comprender cómo se proporciona la información necesaria de entrada a ambos tipos de modelos.

En primer lugar, resulta esencial especificar la salida del modelo que se desea predecir, así como las entradas que serán utilizadas con dicho propósito, y además llevar a cabo la partición del conjunto de datos completo en tres subconjuntos fundamentales: el conjunto de entrenamiento (CE), el conjunto de prueba (CP) y el conjunto de datos nuevos (CN) destinados al proceso de modelado. Con este propósito, se ha implementado un script en Matlab que automatiza la creación de un **Fichero de Configuración**. Este fichero incluye la especificación de todas las entradas que se desean incorporar, así como la salida del propio modelo, además de los porcentajes correspondientes a cada uno de los subconjuntos mencionados.

La separación de los datos en subconjuntos de entrenamiento, prueba y nuevos es fundamental para garantizar que el modelo sea capaz de hacer predicciones precisas en situaciones del mundo real y que no esté simplemente memorizando los datos de entrenamiento. A continuación se explica la función de cada uno:

1. **Conjunto de Entrenamiento (CE):** se utiliza para entrenar o construir el modelo. En el aprendizaje automático, los datos de este subconjunto se utilizan para ajustar los parámetros internos del modelo a medida que aprende a capturar los patrones. El objetivo es que el modelo generalice a partir de estos datos de entrenamiento y pueda hacer predicciones precisas en situaciones similares. Es importante que el conjunto de entrenamiento sea representativo y diverso para que el modelo pueda generalizar correctamente en datos no vistos.

2. **Conjunto de Prueba (CP):** se utiliza para evaluar el rendimiento del modelo entrenado. Estos datos no se utilizan durante el proceso de entrenamiento, por lo que son nuevos para el modelo. La evaluación se realiza al alimentar los datos de prueba al modelo y comparar las predicciones del modelo con los valores reales de la variable objetivo. El conjunto de prueba proporciona una idea de cómo se comporta el modelo en datos no vistos y ayuda a medir su calidad.

3. **Conjunto Nuevo (CN):** se refiere a un conjunto de datos completamente nuevo que el modelo no ha visto durante el proceso de entrenamiento ni de prueba. Estos datos representan situaciones futuras o no observadas y se utilizan para probar la capacidad de generalización a largo plazo del modelo. Evaluar el rendimiento en datos nuevos es crucial para verificar si el modelo puede hacer predicciones precisas y útiles en situaciones que no ha encontrado anteriormente.

En nuestro enfoque, siempre dividiremos los datos en al menos dos subconjuntos. La mayor porción (80 %) constituye el conjunto de entrenamiento (CE) y el resto (20 %) se destina al conjunto de prueba (CP), que se utiliza para evaluar el rendimiento del modelo. Como se mencionó antes, los porcentajes de estos subconjuntos se especifican en el fichero de configuración junto con las entradas y la salida a predecir, utilizando la notación especificada en la sección anterior.

Una vez generado este fichero de configuración, contenedor de toda la información necesaria para la construcción de los modelos, se procede a la generación del archivo que será el que finalmente utilicen los algoritmos de modelado.

Este último archivo se ha denominado "**CX**" debido a que en su interior contendrá los vectores de entrada correspondientes a los conjuntos de entrenamiento (CE) y prueba (CP) que utilizarán los algoritmos de modelaje predictivo.

La creación del archivo CX implica varios pasos fundamentales. En primer lugar, se aplican los porcentajes de los subconjuntos de entrenamiento y prueba al conjunto de datos global. Luego, se lleva a cabo la normalización de los valores de los datos correspondientes. Finalmente, se calculan los **vectores de entrada** correspondientes a las entradas definidas en el fichero de configuración. Estos vectores de entrada se calculan mediante un código en Matlab que ha sido desarrollado para incluir todas las entradas que se han definido para cada serie temporal.

Un vector de entrada, en este contexto, hace referencia al conjunto de valores numéricos que se obtiene al aplicar la definición de una entrada, previamente definida en el fichero de configuración, a la serie temporal correspondiente. Estos vectores de entrada son, en esencia, las variables que constituyen las auténticas entradas que son proporcionadas a los algoritmos de modelado. No obstante, con el propósito de simplificar la comunicación y hacerla más accesible, a lo largo de este TFG se ha optado por emplear el término "entradas" para hacer referencia a estos vectores de entrada.

En el caso de los vectores de entrada correspondientes a las entradas autorregresivas, por lo general, no requieren de cálculos adicionales ya que consisten en tomar simplemente el dato con un desfase específico. Sin embargo, como se verá más adelante, existen algunas excepciones; y para el caso de las entradas exógenas, la mayoría de ellas sí requerirán de cálculos adicionales para obtener sus vectores de entrada.

En resumen, el archivo CX no es más que el conjunto de vectores de entrada correspondiente a las entradas definidas en el fichero de configuración. Cada fichero de configuración se traducirá siempre en un archivo CX. Es por ello que a lo largo de este TFG, con el propósito de simplificar la comunicación y hacerla más accesible, se referirá con frecuencia a ambos ficheros como sinónimos, y se empleará el término "entradas" para hacer referencia a los "vectores de entrada".

Por el contrario, para un mismo CX podrán existir infinitud de modelos predictivos. Esto se ilustra esquemáticamente en la Figura 2.1, y se debe a que, como veremos en la siguiente sección, los modelos de tipo ARX y RBF, cuentan a su vez con una serie de parámetros de entrenamiento cuyos valores habrá que variar a la hora de encontrar buenos resultados en la predicción. De ahí que existan miles de combinaciones posibles, y miles de modelos posibles para un único conjunto de entradas.

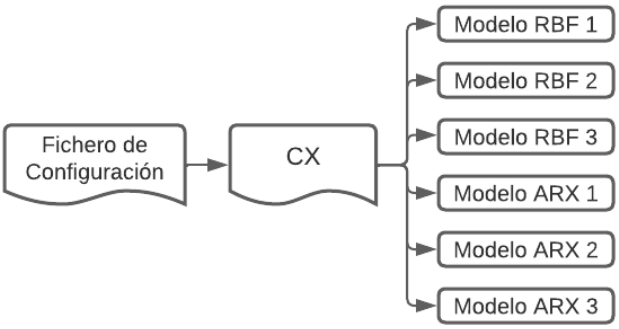


Figura 2.1 Esquema de relación entre Ficheros de Configuración, CXs y Modelos.

Una vez aclarada la forma en la que se le pasa la información de entrada a los algoritmos de Aprendizaje Automático, se pasa a describir estos algoritmos en detalle.

2.5.1 Auto-Regresión con Variables Exógenas (ARX)

El algoritmo de Auto-Regresión con Variables Exógenas (ARX) es un método utilizado en el análisis de series temporales para predecir valores futuros de una variable objetivo, considerando tanto su historia pasada como variables externas que pueden influir en su comportamiento [29].

Es importante mencionar que nuestros modelos también incorporan componentes de medias móviles. Por lo tanto, estaríamos trabajando con el Modelo ARMAX (Auto-Regresión con Medias Móviles y Variables Exógenas), que es una extensión del Modelo ARX [30]. Sin embargo, con el fin de mantener la simplicidad a lo largo de este TFG, se refiere a ambos enfoques como ARX.

El Modelo ARX (Auto-Regresión con Variables Exógenas) es lineal, es decir, las relaciones entre las variables de entrada y la variable objetivo se expresan como combinaciones lineales ponderadas por coeficientes. Cada coeficiente representa la contribución lineal de una variable.

Utilizaremos una función de entrenamiento en Matlab la cual se encargará de crear modelos lineales de tipo ARMAX utilizando el algoritmo de Mínimos Cuadrados Recursivos (MCR) sobre el conjunto de entrenamiento. A continuación, se explica de manera sintetizada las fases del mismo:

Algoritmo de Mínimos Cuadrados Recursivos (MCR):

1. **Inicialización**: Los parámetros iniciales del modelo ARMAX, representados por el vector $\boldsymbol{\theta}$, se establecen con valores iniciales por defecto. Además, la matriz $P(0)$ se inicializa multiplicando el factor $FactP$ por la matriz identidad. Esta matriz desempeña un papel fundamental en la actualización de los parámetros en cada iteración del algoritmo.

2. **Iteración sobre los datos**: El algoritmo recorre los datos de entrenamiento, uno a uno, en orden secuencial. Para cada dato de entrada en el tiempo actual, se realiza una predicción utilizando los parámetros actuales del modelo y los datos anteriores. La predicción se compara con el valor real de salida en el mismo tiempo, y se calcula el error entre la predicción y el valor real.

3. **Actualización de parámetros**: Los parámetros del modelo, $\theta(t)$, se actualizan de manera recursiva utilizando la siguiente fórmula:

$$\theta(t) = \theta(t-1) + \frac{P(t-1)\mathbf{x}(t)}{\gamma + \mathbf{x}^T(t)P(t-1)\mathbf{x}(t)}(y(t) - \mathbf{x}^T(t)\theta(t-1)) \tag{2.2}$$

Donde:

- $\theta(t)$ es el vector de parámetros en el tiempo t.
- $P(t)$ es la matriz de covarianza en el tiempo t.
- $\mathbf{x}(t)$ es el vector de entrada en el tiempo t.
- $y(t)$ es el valor real de salida en el tiempo t.
- γ es el parámetro ‘gamma‘.

La matriz de covarianza $P(t)$ se actualiza de manera recursiva utilizando la siguiente fórmula:

$$P(t) = \frac{1}{\gamma}\left(P(t-1) - \frac{P(t-1)\mathbf{x}(t)\mathbf{x}^T(t)P(t-1)}{\gamma + \mathbf{x}^T(t)P(t-1)\mathbf{x}(t)}\right) \tag{2.3}$$

4. **Repetición**: El proceso se repite para cada dato de entrenamiento en orden secuencial hasta que se hayan procesado todos los datos.

5. **Resultados**: Al final de la iteración sobre los datos, el algoritmo devuelve los parámetros del modelo ARMAX, que son los valores finales de $\theta(t)$.

Además del CX, la función de entrenamiento tomará como entradas los parámetros **Gamma** y **FactP**, cuyos valores tendremos que elegir:

- **Gamma**: Controla el factor de olvido en el algoritmo. Un valor cercano a 1 da más peso a los datos más recientes, mientras que un valor cercano a 0 da un peso similar a todos los datos pasados. En general, un valor de 1 suele ser apropiado cuando se espera que el modelo sea más sensible a cambios recientes en los datos.
- **FactP**: Se utiliza para inicializar la matriz de covarianza $P(0)$. Un valor elevado como 10 suele ser apropiado, pero puede requerir ajustes según el problema.

Es importante destacar que, debido a su naturaleza lineal, el Modelo ARX es más adecuado para situaciones en las que las relaciones entre las variables son relativamente simples y pueden ser aproximadas de manera efectiva con combinaciones lineales. Si se espera que las relaciones sean más complejas y no lineales, podría ser necesario explorar otros enfoques de modelado, como modelos no lineales o redes neuronales.

En los siguientes capítulos, se evaluará la eficacia de los modelos ARX, tanto en una serie sencilla, como es el caso de la demanda eléctrica, como en una mucho más compleja, como las transacciones de una entidad financiera.

2.5.2 Redes Neuronales de Base Radial (RBF)

Una red neuronal de base radial (RBF, por sus siglas en inglés) es un tipo de red neuronal artificial que utiliza una función de base radial como función de activación en sus neuronas ocultas [31]. A diferencia de las redes neuronales tradicionales con funciones de activación como la sigmoide, las RBF utilizan funciones radiales, como la función gaussiana, para modelar las transformaciones no lineales en los datos de entrada. Esta **no linealidad** es crucial para que la red pueda modelar y capturar relaciones complejas en los datos de entrada.

La RBF, al igual que la mayoría de redes neuronales artificiales, se caracteriza por su arquitectura compuesta de tres capas principales: la capa de entrada, la capa oculta o intermedia y la capa de salida.

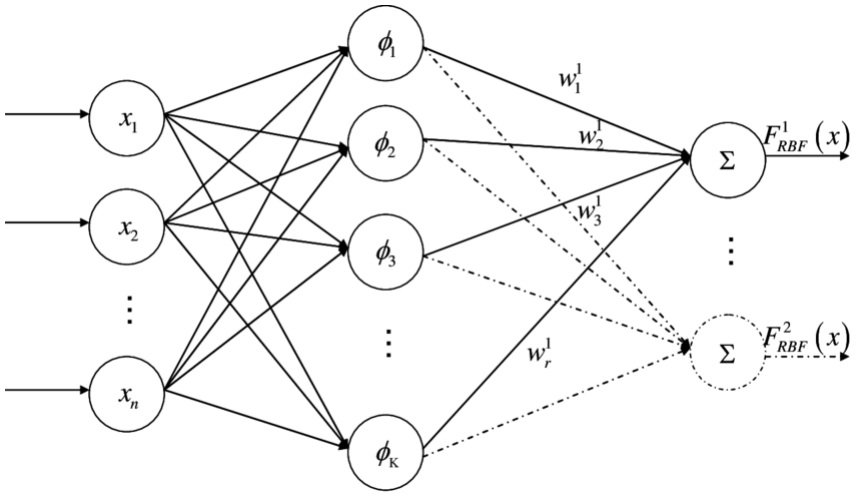


Figura 2.2 Arquitectura de una red neuronal de base radial (RBF).

- **Capa de Entrada:** su función principal es la de recibir los datos de entrada y transferirlos directamente a la siguiente capa, que es la capa oculta.

- **Capa Oculta:** es el corazón de la red neuronal RBF, donde se lleva a cabo la transformación no lineal de los datos. Está compuesta por neuronas RBF, que utilizan funciones de base radial (ϕ) como su función de activación.

- **Capa de Salida:** es la última etapa de la red neuronal RBF y genera las salidas finales del modelo. después de que los datos hayan sido procesados y transformados a través de las capas anteriores.

Función de Base Radial ($\phi(r)$):

En el contexto de las redes neuronales, una función de activación es una función matemática que se aplica a la salida de una neurona para introducir no linealidad en el proceso de cálculo. En las redes RBF, las funciones de activación son funciones de base radial, como la función gaussiana.

$$\phi(r) = e^{-\frac{r^2}{2\kappa^2}} \tag{2.4}$$

Donde:

- $\phi(r)$ es el valor de la función de base radial.
- r es la distancia entre el punto de evaluación y el centro de la función.
- κ es un parámetro llamado "kappa" que controla la amplitud de la función.

Cada neurona RBF se caracteriza por tres elementos principales: un centro, el valor kappa que controla la amplitud de la función y, en ocasiones, un peso asociado. La salida de la neurona se calcula mediante la aplicación de una función de base radial utilizando la distancia entre el centro de la neurona y el vector de entrada como parámetros.

La característica clave de las funciones de base radial, como la gaussiana, es que a medida que aumenta la distancia entre el vector de entrada actual y el centro de la neurona, menor será el valor de salida de la función de base radial. Esto significa que las funciones de base radial asignan valores más altos a puntos cercanos al centro y valores más bajos a puntos lejanos. Esta propiedad es muy útil porque permite que las neuronas en la capa oculta respondan más fuertemente a entradas que están cerca de su centro y más débilmente a aquellos que están más lejos, capturando así patrones no lineales en los datos de entrada.

Por lo que cada vez que se presenta un dato de entrada se calcula la distancia entre el centro de cada neurona RBF y el dato de entrada. Esta distancia se pasa a través de la función de base radial para obtener la activación de la neurona oculta. Las activaciones de todas las neuronas ocultas se pasan como entradas a la capa de salida. Finalmente, la capa de salida produce las salidas finales del modelo.

El entrenamiento de la red RBF involucra ajustar los pesos en la capa de salida para que las salidas se asemejen a los valores deseados. Este proceso de entrenamiento implica dos pasos principales: la colocación de los centros de las neuronas y el ajuste de los pesos. La ubicación de los centros implica determinar sus posiciones basándose en los datos de entrenamiento.

Se utiliza una función de Matlab que se encarga del entrenamiento de una red neuronal RBF utilizando el algoritmo LMS (Least Mean Squares en inglés).

Algoritmo LMS (Least Mean Squares):

El algoritmo LMS en este contexto se utiliza para ajustar los pesos de salida de la red RBF (Radial Basis Function) de manera que minimice el error cuadrático medio (ECM) entre las salidas predichas por la red y las salidas reales del conjunto de entrenamiento. A continuación, se describen los pasos:

1. **Inicialización de pesos:** En el inicio, los pesos de salida se inicializan con una serie de valores iniciales por defecto.

2. **Paso de entrenamiento:** El algoritmo LMS se ejecuta durante un número especificado de pasadas a través del conjunto de entrenamiento. En cada pasada, se toma un valor del conjunto de entrenamiento y se realiza el siguiente proceso:

   - Se aplica el ejemplo de entrenamiento a la red RBF para calcular su salida predicha.
   - Se calcula el error entre la salida predicha y la salida deseada (diferencia entre la predicción y el valor real).
   - Se ajustan los pesos de salida de la red RBF en función de este error y los parámetros de aprendizaje, en este caso, controlado por el parámetro γ.

3. **Actualización de pesos:** La actualización de los pesos de salida se realiza utilizando la regla LMS, que es una forma de descenso de gradiente. Los pesos se ajustan de acuerdo con la siguiente fórmula:

$$W_i^{(n+1)} = W_i^{(n)} + \gamma \cdot e \cdot \phi_i(x) \tag{2.5}$$

   Donde:

   - $W_i^{(n+1)}$ es el peso i en la iteración $n+1$.
   - $W_i^{(n)}$ es el peso i en la iteración n.
   - γ es el parámetro que controla la velocidad de ajuste de los pesos.
   - e es el error entre la salida predicha y la salida deseada.
   - $\phi_i(x)$ es la función de base radial correspondiente al centro i.

4. **Convergencia:** El algoritmo LMS se repite durante un número especificado de pasadas o hasta que el error cuadrático medio (ECM) alcance un valor aceptable. El ECM se calcula en cada iteración y se utiliza como criterio de convergencia.

En resumen, el algoritmo LMS se utiliza para ajustar los pesos de salida de la red RBF de manera iterativa, de modo que la red pueda aprender a predecir con precisión las salidas deseadas a partir de los ejemplos de entrenamiento. El objetivo es minimizar el error cuadrático medio entre las salidas predichas y las salidas reales.

Nuestra función de entrenamiento de modelos RBF toma como argumentos de entrada el archivo CX, así como los siguientes parámetros cuyos valores habrá que definir para cada prueba:

- **Alpha**: Paso de entrenamiento para la atracción de los centros RBF. Controla la velocidad de convergencia de los centros hacia las entradas de entrenamiento.
- **Kappa**: Es el parámetro que controla la apertura y solapamiento de las funciones de base radial, que aparece en la Ecuación 2.4. La elección de Kappa es fundamental para adaptar las funciones RBF al comportamiento de los datos y puede influir mucho en la precisión de las predicciones.
- **Número de neuronas**: Este parámetro determina la cantidad de neuronas. Debe seleccionarse con cuidado, ya que afectará la complejidad del modelo. Un número insuficiente de neuronas podría resultar en un modelo subajustado, mientras que un número excesivo podría llevar a un sobreajuste.
- **Número de pasadas**: Número de iteraciones para la actualización de los centros RBF durante el entrenamiento. Un número apropiado de pasadas es importante para lograr la convergencia del modelo.
- **Gamma**: Es el parámetro de adaptación LMS, que aparece en la Ecuación 2.5 Controla la velocidad de ajuste de los pesos de salida de la red. La elección de Gamma puede afectar la rapidez con la que el modelo se ajusta a los datos de entrenamiento. En la práctica usaremos siempre un valor de 1.

Posteriormente en las pruebas, se verá que los dos parámetros que más influencia tienen en los resultados son kappa y el número de neuronas. El objetivo será crear numerosas baterías de modelos, barriendo los valores de todos los parámetros de entrenamiento con el fin de encontrar la combinación óptima que se ajuste adecuadamente a la serie que se esté tratando.

2.6 Evaluación

La etapa final de nuestra metodología se centra en la evaluación de los modelos que hemos creado. Una vez que el modelo ha sido entrenado con el conjunto de entrenamiento y los pesos han sido ajustados, se procede a la evaluación del rendimiento del modelo mediante la utilización del conjunto de prueba. Con este propósito, se han diseñado funciones específicas, las cuales, tomando como entrada un modelo previamente definido, se encargan de aplicar dicho modelo a los datos pertenecientes al conjunto de prueba. Como resultado de esta evaluación, estas funciones generan dos vectores columnas para ambos conjuntos: uno que contiene las salidas predichas por el modelo y otro que contiene las salidas reales.

Con esta información en mano, se puede realizar una evaluación exhaustiva de la calidad y eficacia de los modelos que se han desarrollado. A continuación, se expondrán las diferentes métricas de evaluación que se han utilizado en las pruebas.

Notación previa común a todas las métricas:

- N: Número de observaciones o registros del conjunto evaluado.
- k: Índice de observación / Instante de tiempo.
- y_k: Salida real/esperada en el instante k.
- $\hat{y}_k$: Predicción realizada por el modelo en el instante k.

- **Error Cuadrático Medio (ECM):** calcula la raíz cuadrada del promedio de los errores al cuadrado entre los valores reales y las predicciones. El elevar al cuadrado da más peso a los errores grandes, lo que puede hacer que esta métrica sea sensible a valores atípicos [32].

$$ECM = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(y_k - \hat{y}_k)^2} \tag{2.6}$$

- **Error Absoluto Medio (EAM):** es una métrica que calcula el promedio de las diferencias absolutas entre los valores reales de la serie temporal y_k y las predicciones $\hat{y}_k$ [33].

$$EAM = \frac{1}{N}\sum_{k=1}^{N}|y_k - \hat{y}_k| \tag{2.7}$$

- **Error Porcentual Medio Absoluto Arcotangente (MAAPE):** calcula el promedio de los arcos tangentes de las diferencias porcentuales absolutas entre los valores reales y las predicciones. Tiene la particularidad de manejar los errores porcentuales de manera simétrica, evitando la tendencia de las métricas tradicionales a penalizar más los errores de sobreestimación que los de subestimación [34].

$$MAAPE = \frac{1}{N}\sum_{k=1}^{N}\arctan\left(\frac{|y_k - \hat{y}_k|}{y_k}\right) \tag{2.8}$$

- **Error Porcentual Medio Absoluto Simétrico (SMAPE):**calcula el promedio de las diferencias porcentuales absolutas, teniendo en cuenta tanto los valores reales como las predicciones en el denominador [35].

$$SMAPE = \frac{1}{N}\sum_{k=1}^{N}\frac{|y_k - \hat{y}_k|}{|y_k| + |\hat{y}_k|} \tag{2.9}$$

Estas métricas ofrecen diferentes enfoques para evaluar la calidad de las predicciones de series temporales, y se utilizarán tanto para evaluar cada modelo individualmente, así como para comparar los resultados obtenidos por los distintos tipos de modelos.

Esta comparación nos permitirá identificar cuál de los enfoques, ya sea ARX o RBF, se ajusta mejor a las características específicas de la serie temporal que se esté tratando, y dentro de cada enfoque, cuáles han sido los modelos que mejores resultados han dado, para de esta forma, tener en cuenta cuáles han sido las entradas autorregresivas y los parámetros de entrenamiento que mejor han funcionado.

Concretamente, a lo largo de este TFG, para exponer los resultados de las pruebas se utiliza principalmente el Error Cuadrático Medio (ECM) y el Error Porcentual Absoluto Medio (SMAPE).

Aunque las cuatro métricas expuestas anteriormente son importantes para evaluar la calidad de los modelos y es fundamental considerar todas ellas al llegar a conclusiones sobre qué pruebas han dado los mejores resultados, a la hora de representar los errores de los modelos en gráficas, se priorizará principalmente la utilización del Error Cuadrático Medio (ECM). Esto se debe a que los valores del ECM suelen ser más altos en comparación a las otras métricas, lo que facilita la visualización de los incrementos o descensos en los errores entre distintos modelos y pruebas.

Una vez conocidos los algoritmos de modelado, y las métricas de evaluación de los modelos, para finalizar este capítulo de metodología, se explica la estrategia mencionada en la Sección 2.4. Esta estrategia es el enfoque principal para la mayoría de las pruebas realizadas en este TFG, ya que permite encontrar la combinación óptima de entradas autorregresivas a utilizar.

Este enfoque es conocido como el **Método de Inclusión Progresiva de Entradas**, y se trata de un enfoque iterativo que nos permitirá determinar qué combinación de entradas autorregresivas y exógenas produce los mejores resultados en términos de predicción.

2.6.1 Método de Inclusión Progresiva de Entradas

El Método de Inclusión Progresiva de Entradas se desarrolla en varias fases:

1. En la fase de análisis se realiza un estudio de las correlaciones de Pearson para cada entrada potencial única definida, ya sea autorregresiva o exógena. Esto implica calcular la correlación lineal entre la salida y la entrada desfasada en el rango de desfases predefinido.

2. Para cada entrada, se identifican los 15 desfases de mayor correlación, que serán las entradas autorregresivas que usemos en los modelos de la primera iteración. De esta forma se reduce en gran medida el número de entradas iniciales a probar, agilizando el proceso de selección de entradas y, por tanto, de creación de modelos.

3. Se crean 15 ficheros de configuración incluyendo cada uno de esos 15 desfases de entrada, lo cual se trasladará a la creación de 15 CXs, y posteriormente a 15 modelos predictivos (ARX o RBF según la prueba que estemos haciendo).

4. Se evalúan esos 15 modelos mediante las 4 métricas expuestas en la Sección 2.5, enfocándonos especialmente en el ECM. A partir de esta evaluación, se identificará el modelo que haya brindado los mejores resultados, y la entrada presente en el fichero de configuración utilizado para crear dicho modelo, con el fin de utilizarla en la siguiente iteración.

5. Se vuelven a crear nuevos ficheros de configuración, los cuales incluiran el desfase elegido en la iteración anterior y cada uno de los 14 desfases restantes con las correlaciónes más altas que no fueron elegidos en la anterior iteración. Se vuelve a generar un CX, por cada uno de estos ficheros y un modelo por cada uno de los CXs. Los modelos se vuelven a evaluar, y finalmente se elige para la siguiente iteración la combinación de desfases de entrada presentes en el modelo que haya demostrado mejor rendimiento.

6. Se repite este proceso iterativamente hasta dejar de observar mejora en la evaluación de modelos, lo cual se ha comprobado que tiende a ocurrir antes de incluir los 15 desfases en el proceso.

A continuación, en la Tabla 2.1 se ilustra el procedimiento seguido. Cada fila representa los desfases de entrada de cada fichero de configuración, y cada columna representa una iteración del proceso. En **negrita** se indica el conjunto de entradas elegido en la iteración correspondiente.

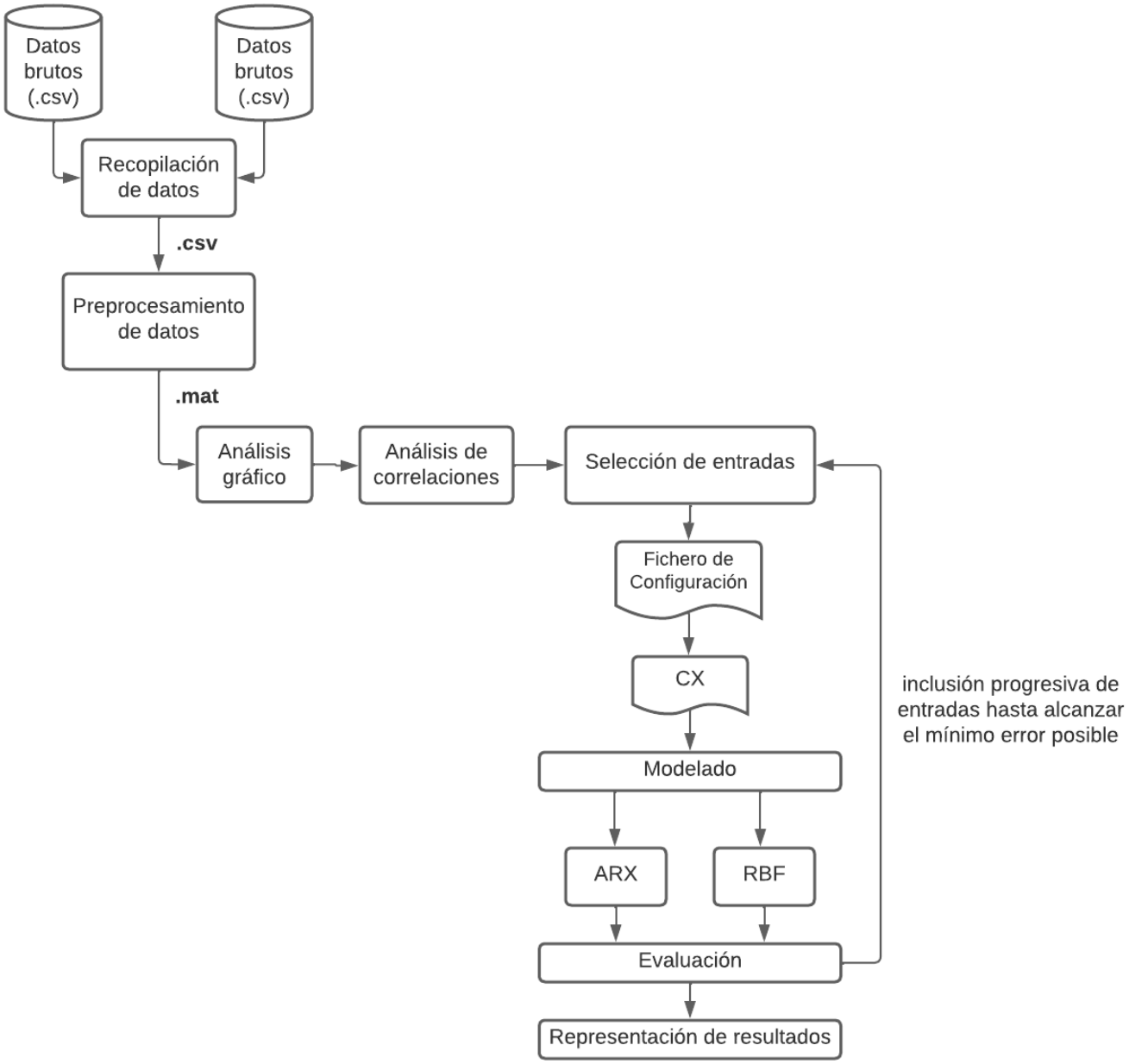
Tabla 2.1 Esquema del Método de Inclusión Progresivo de Entradas.

Iteración 1	Iteración 2	Iteración 3	...	Iteración 10
-21	**-17,-21**	-17,-21,-56	...	**-17, -21, -23, -25, -24, -18, -53, -49, -52, -56**
-56	-17,-56	-17,-21,-24	...	-
-24	-17,-24	-17,-21,-52	...	-
-52	-17,-52	-17,-21,-49	...	-
-49	-17,-49	-17,-21,-53	...	-
-17	-17,-53	-17,-21,-25	...	-
-53	-17,-25	-17,-21,-18	...	-
-25	-17,-18	**-17,-21,-23**	...	-
-18	-17,-23	-	...	-
-23	-	-	...	-

Una vez completado este proceso, se habrá obtenido el conjunto de desfases de entradas que mejor se ajuste a la serie temporal que se esté tratando de predecir. El siguiente paso a seguir con este conjunto de entradas dependerá del tipo de prueba que se esté realizando. En el caso de una prueba con modelos RBF, generalmente se lleva a cabo un barrido exhaustivo de los valores de los parámetros de entrenamiento del modelo, como se detalla en la Subsección 2.5.2. El objetivo de este barrido es encontrar la combinación óptima de estos parámetros para el conjunto de entradas seleccionado, lo que permitirá mejorar la capacidad predictiva del modelo.

2.7 Diagrama de las fases de la metodología

Finalmente, concluimos el capítulo con un diagrama que proporciona una síntesis visual de la metodología detallada en las secciones previas.



3 Aplicación a serie temporal de la demanda de la energía eléctrica

En este capítulo, nos sumergiremos en la aplicación práctica de la metodología que se ha detallado cuidadosamente en el capítulo anterior. Abordaremos una serie temporal con datos reales que corresponden al consumo de energía eléctrica en los hogares de tres ciudades ubicadas en España.

El objetivo central es desarrollar modelos predictivos capaces de anticipar el consumo eléctrico. Para lograrlo, se llevará a cabo una serie de pruebas exhaustivas y experimentos que involucrarán diferentes enfoques y horizontes temporales de predicción. Nuestro propósito es evaluar la capacidad de estos modelos para hacer pronósticos precisos y útiles en el contexto del consumo de electricidad en estas ciudades.

Pero antes de adentrarnos en las pruebas, se procederá a explicar las etapas de recopilación, preprocesamiento y análisis de los datos con el fin de prepararlos y tener una visión clara de la serie temporal que estamos tratando.

3.1 Recopilación de datos

Originalmente los datos provienen de la antigua Sevillana de Electricidad. Contamos con un archivo principal que recopila información sobre el consumo promedio de energía eléctrica en tres ciudades: Sevilla, Córdoba y Málaga. Este archivo abarca un período que se extiende desde enero de 1994 hasta enero de 1998. Los datos se expresan en megavatios (MW) y están registrados con una frecuencia horaria, lo que implica que se tiene la cantidad de electricidad consumida en cada una de las 24 horas del día.

Adicionalmente, se dispone de un archivo secundario que contiene información sobre las temperaturas medias diarias. Estas temperaturas medias diarias se han calculado como el promedio de las temperaturas registradas en las mismas tres ciudades mencionadas anteriormente.

Es importante destacar que este segundo archivo corresponde a una serie temporal externa a la serie del consumo eléctrico. Sin embargo, debido a la estrecha relación entre la temperatura ambiente y el consumo de energía eléctrica en los hogares, se explorará cómo esta serie temporal externa puede ser aprovechada para generar una entrada exógena que contribuya a mejorar las

predicciones.

3.2 Preprocesamiento de datos

En primer lugar, se verifica que los datos estén organizados en orden cronológico y que no haya valores faltantes ni valores atípicos que deban ser eliminados. Como ya se explicó en el Capítulo 2, se debe llevar a cabo la tarea de mapear los campos de texto a valores numéricos con el fin de facilitar su manipulación en Matlab. En el caso de ambos ficheros, dado que todos los datos son numéricos, este mapeo será innecesario.

Los campos del fichero del **consumo de la potencia eléctrica** son:

- La variable autoincremental del día (va de 0 hasta el número de filas totales).
- Consumo en la hora 1.
- Consumo en la hora 2.
- ...
- Consumo en la hora 24.

Mientras que en el de la **temperatura media** los dos únicos campos serán:

- La variable autoincremental del día (va de 0 hasta el número de filas totales).
- La temperatura ambiente en Sevilla, Córdoba y Málaga calculada como la media de la temperatura en dichas ciudades.

3.3 Análisis de los datos

A la hora de analizar esta serie temporal, se plantean dos enfoques posibles: el primero implica analizar el consumo a **nivel diario**, empleando los días como puntos de referencia temporales. El segundo enfoque, en cambio, consiste en examinar el consumo a **nivel horario**, utilizando las horas como unidades temporales.

Estas dos perspectivas en el análisis también se reflejan en dos enfoques distintos a la hora de predecir el consumo energético. Como se observará más adelante, esto se traducirá en la creación de dos tipos de entradas autorregresivas distintas: una diseñada para predecir el consumo total diario y otra para predecir el consumo horario.

Predicción del Consumo Total Diario:

- Los datos de consumo se agrupan por días, lo que resulta en un menor número de puntos de datos en comparación con el enfoque horario.
- Los patrones de consumo a lo largo de los días pueden ser más visibles, especialmente si hay patrones semanales o estacionales.
- Las fluctuaciones intradía desaparecen, por lo que si hay variaciones significativas dentro de los días, como picos de consumo en ciertas horas, estos detalles se perderán debido a la agregación diaria.
- La resolución diaria puede no ser suficiente para capturar patrones de corto plazo o eventos excepcionales.

Predicción del Consumo Horario:

- Se trabajan con datos de consumo en intervalos horarios, lo que nos proporciona una mayor cantidad de puntos de datos en comparación con el enfoque diario.
- La mayor resolución temporal puede permitir que los modelos encuentren patrones más sutiles y relaciones complejas en los datos, como picos de consumo durante horas pico o cambios en el comportamiento en momentos específicos del día debido a las diferentes rutinas y hábitos de las personas, lo que puede llevar a predicciones más precisas.
- Los detalles intradía se conservan, lo que puede ser útil si estamos interesados en comprender y predecir cambios en el consumo en intervalos más pequeños.
- Al disponer de más puntos de datos, los modelos pueden volverse más complejos y requerir más capacidad computacional para su entrenamiento.

Ambos enfoques son válidos y presentan sus propias ventajas y desafíos, sin embargo, es importante mencionar que la elección entre predecir a nivel diario o a nivel horario dependerá principalmente del propósito de la predicción. La elección siempre debe estar basada en la ventana temporal que tengamos interés en predecir, es decir, en el horizonte de tiempo para el cual queremos obtener predicciones precisas a través de nuestros modelos.

En este TFG, se llevarán a cabo pruebas con ambos enfoques y distintas ventanas temporales para determinar que combinación se ajusta mejor a los datos.

3.3.1 Análisis gráfico

En primer lugar, se verifica que no haya datos faltantes y se procede a representar gráficamente las series temporales contenidas en los archivos ya preprocesados.

Primero se representa el consumo total diario de energía eléctrica. Para lograrlo, se suma la consumida en las 24 horas de cada día y se representan estos valores en función del tiempo.

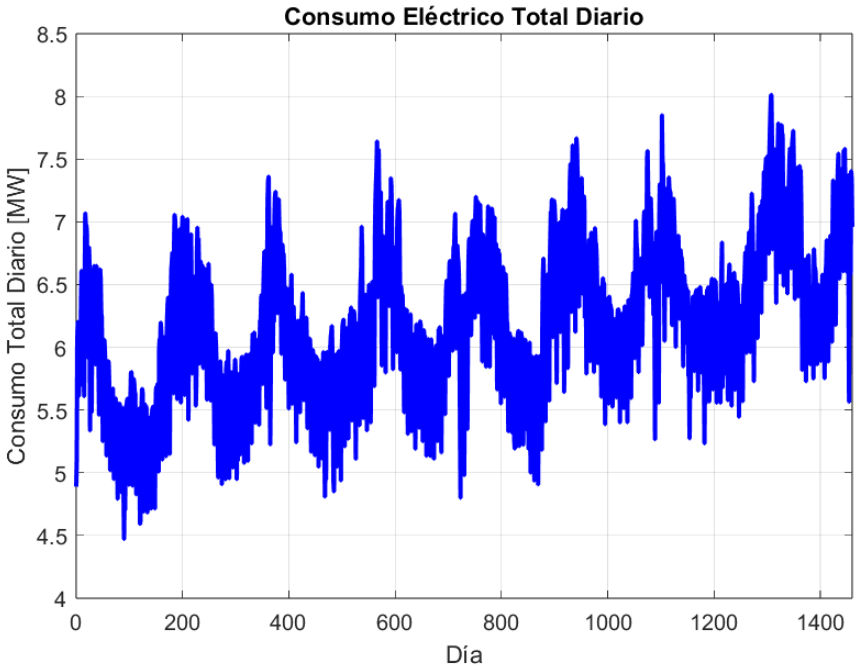


Figura 3.1 Consumo Eléctrico Total Diario entre Enero de 1994 y Enero de 1998.

Se observa un patrón muy interesante en la Figura 3.1 del que podemos deducir varias cosas. En primer lugar, es evidente que el consumo sigue una tendencia periódica. Los picos de consumo muestran una disposición regular en el tiempo.

El patrón observado en los datos revela ciclos recurrentes de aproximadamente 365 días, marcados por fases de alto consumo, indicativas de las estaciones de verano o invierno, cuando es común el uso de calefacción o aire acondicionado. Posteriormente, se registra una disminución en el consumo, que probablemente corresponde a la primavera u otoño.

Como resultado, cada pico de consumo se repite en ciclos de alrededor de 180 días, lo que representa medio año y coincide con el cambio entre las estaciones de verano e invierno. Dado el claro carácter periódico de este patrón, es de esperar una fuerte correlación lineal entre los momentos temporales con diferentes desfases en esta serie temporal.

Además, resulta interesante ver una tendencia general creciente a los largo de los 4 años, la cual podría estar influenciada por las variaciones en las temperaturas ambientales.

En referencia a esto último, también es altamente probable una correlación significativa con la

temperatura ambiente. Se anticipa que existirá una correlación positiva entre la temperatura y el consumo eléctrico durante las estaciones de verano, cuando ambos alcanzan sus niveles máximos. Por el contrario, durante las estaciones de invierno, caracterizadas por temperaturas más bajas, se prevé una correlación negativa entre la temperatura y el consumo eléctrico, ya que la demanda de calefacción, como se observó previamente, aumenta en climas más fríos.

A continuación, se representan los datos de la temperatura ambiente promedio para confirmar esta tendencia.

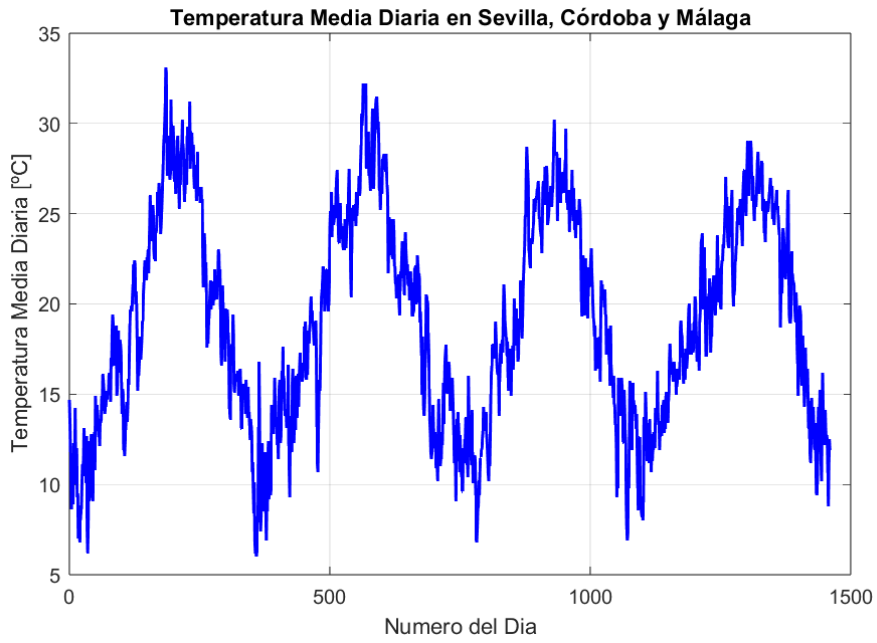


Figura 3.2 Temperatura Media Diaria en las ciudades de Sevilla, Córdoba y Málaga entre Enero de 1994 y Enero de 1998.

Como era de esperar, al observar la Figura 3.2 se puede confirmar que tanto los picos positivos como los negativos de la temperatura se repiten en intervalos de aproximadamente 365 días, coincidiendo con las estaciones de verano e invierno, respectivamente.

Con seguridad, se puede afirmar que la temperatura ambiente será una variable exógena que se incluirá en nuestro modelo para predecir el consumo diario. Esto se debe a la evidente relación que hemos observado en las representaciones gráficas anteriores entre la temperatura y el consumo eléctrico.

Por último, se representa el consumo eléctrico a nivel horario durante los 4 años.

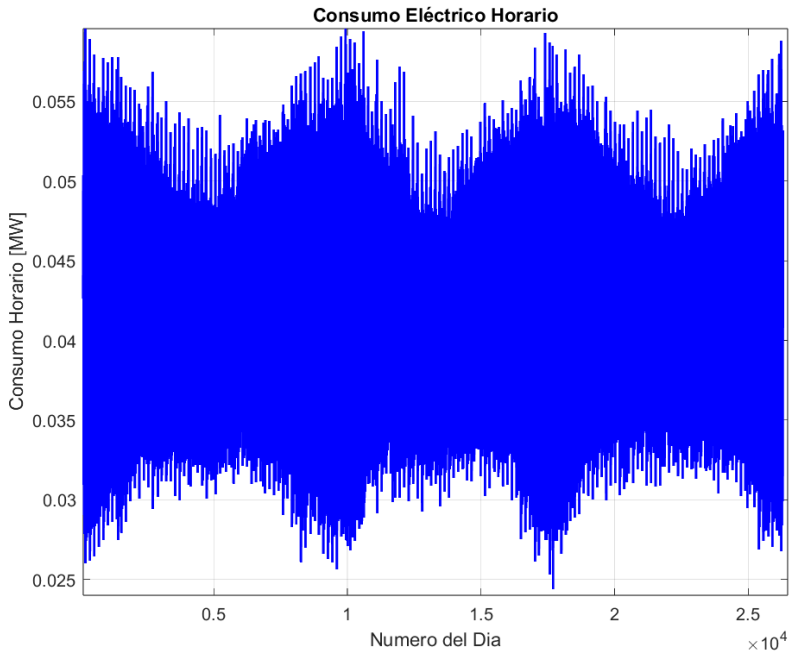


Figura 3.3 Consumo Eléctrico Horario entre Enero de 1994 y Enero de 1998.

Se realiza un acercamiento en la gráfica para una mejor visualización y para extraer posibles conclusiones de la representación.

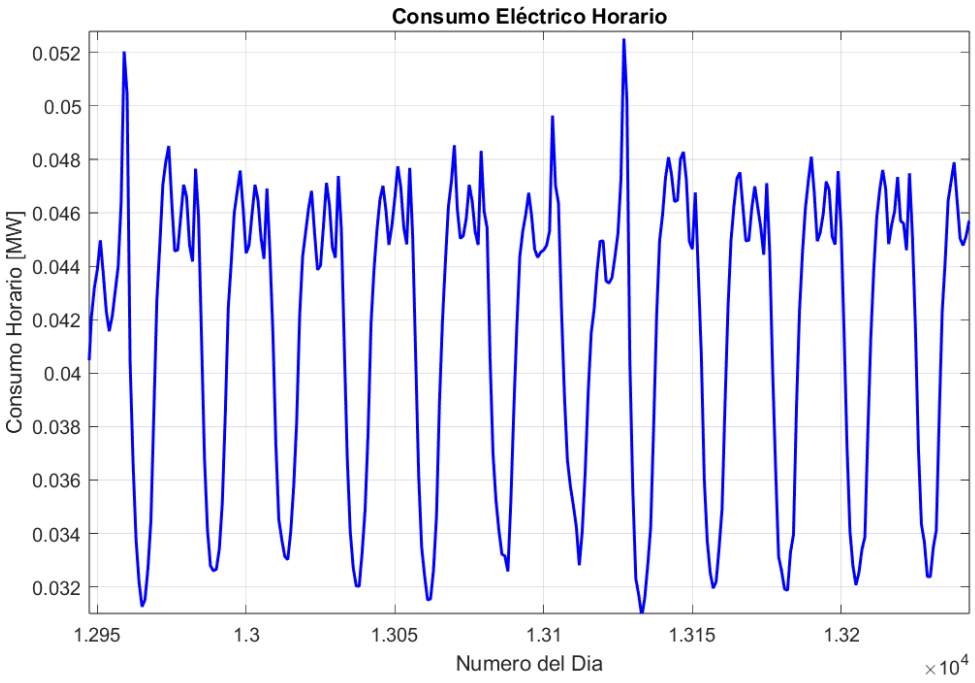


Figura 3.4 Consumo Eléctrico Horario entre Enero de 1994 y Enero de 1998.

En la Figura 3.4 se pueden apreciar con mayor claridad los picos de consumo durante el día y los descensos durante la noche. Además, se observa una cierta periodicidad semanal, ya que los picos de consumo máximo parecen repetirse con un intervalo de aproximadamente 7 días.

3.4 Definición de entradas

Tal como se explica detalladamente en la Sección 3.3, el problema de la predicción puede ser abordado desde dos enfoques diferentes según usemos las horas o los días como unidades temporales. Para adaptar nuestro método a ambas perspectivas, hemos creado dos tipos de entradas autorregresivas, cada una de ellas diseñada para ajustarse a cada uno de los enfoques mencionados anteriormente. Estas entradas son:

1. **A1@d**: se encargará de tomar los datos del consumo eléctrico total de cada día, obtenido como la suma de los consumos de cada hora. Cuando su desfase sea positivo, se utilizará como salida para predecir el consumo diario.

2. **A2@d**: se encargará de tomar los datos del consumo eléctrico de cada hora. Cuando su desfase sea positivo, se utilizará como salida para predecir el consumo diario.

Como se detallará en las secciones dedicadas al modelado, se llevarán a cabo una serie de pruebas empleando cada uno de estos dos tipos de entradas como punto de partida, y se considerarán varios horizontes temporales para la predicción. De manera específica, se utilizará las salidas k+1 y k+7 en las pruebas que involucran la entrada A1@d, con el fin de anticipar el consumo total diario del día siguiente y el consumo total de una semana en el futuro, respectivamente. Asimismo, se utilizará la salida k+1 en las pruebas relacionadas con la entrada A2@d, con el objetivo de prever el consumo eléctrico horario de la siguiente hora.

Además, con el propósito de mejorar aún más las predicciones se han creado los siguientes tipos de entradas exógenas.

3.4.1 Definición de entradas exógenas

Para el consumo diario:

1. **Entrada de tipo calendario (C)**

   - **C1@d:** Identificador asignado a cada día de la semana.
   - **C2@d:** Identificador asignado a cada día del mes.
   - **C3@d:** Identificador asignado a cada mes.
     Estas entradas permitirán al modelo tener en cuenta a que día de la semana, del mes y a que mes corresponde la salida a predecir. El objetivo es por tanto, caracterizar la salida a predecir. Por ello el desfase que se usa en estas entradas es positivo e igual a la salida.

2. **Entradas derivadas (D)**

   - **D1@d y D2@d:** seno y coseno del día de la semana respectivamente.
   - **D1@d y D2@d:** seno y coseno del día de la semana respectivamente.
   - **D3@d y D4@d:** seno y coseno del día del mes respectivamente.
   - **D5@d y D6@d:** seno y coseno del mes respectivamente.
   - **D7@d:** semana del mes.
   - **D8@d y D9@d:** seno y coseno de la semana del mes respectivamente.

   Como su propio nombre indica, estas entradas se derivan de la entrada de tipo calendario. Las funciones trigonométricas de seno y coseno permitirán capturar periodicidades o fluctuaciones regulares a lo largo del tiempo, como picos de consumo en ciertos días de la semana. El desfase utilizado también es positivo e igual a la salida con el objetivo de caracterizar el día a predecir.

3. **Entrada proveniente de la serie temporal externa de temperaturas (T)**

   - **T1@d:** Temperatura media del día anterior.

   Al introducir esta información como una entrada exógena en el modelo, se permite que este considere cómo las variaciones en la temperatura pueden afectar el comportamiento del consumo energético, lo que puede conducir a predicciones más precisas y útiles.

Para el consumo horario:

1. **Entradas que aportan información adicional extraída de la propia serie temporal del consumo horario (A)**

   - **A3@d:** Consumo horario del mismo día de la semana anterior.
   - **A4@d:** Consumo horario del mismo día del año anterior.

   Estas entradas aportan contexto histórico y permiten que el modelo considere cómo se ha comportado el consumo en situaciones similares en el pasado. La entrada A3@d permite al modelo capturar patrones recurrentes que se repiten semanalmente, mientras que la A4@d es útil para capturar patrones estacionales que pueden repetirse anualmente. Por ejemplo, si en un día específico del año suele haber un aumento de consumo debido a una festividad o evento recurrente, esta entrada permitirá que el modelo considere cómo se comportó el consumo en las mismas horas del día del año anterior.

3.5 Análisis de correlaciones de Pearson

En esta sección, se centrará en calcular y visualizar las correlaciones que existen entre los diferentes instantes de tiempo desfasados, tanto para la serie temporal del consumo diario como para la del consumo horario.

A continuación, se representan en la Figura 3.5 los valores de correlación de Pearson para el consumo eléctrico diario. Se utiliza un rango de desfases desde -1 hasta -180, lo que abarca un período de aproximadamente medio año.

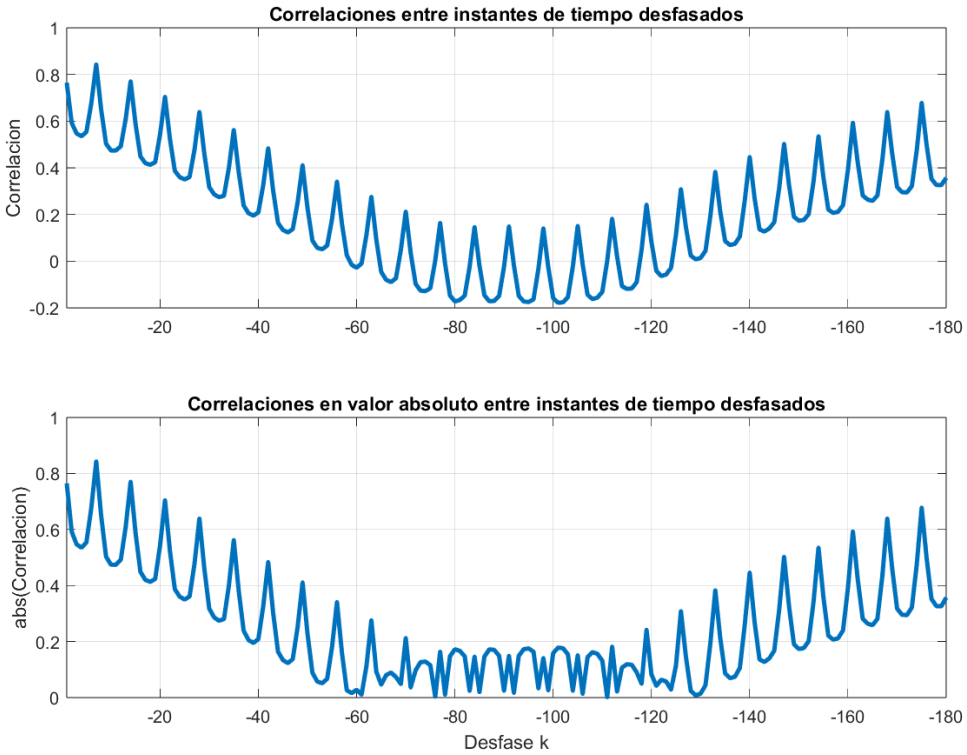


Figura 3.5 Correlaciones entre los datos de (k+d) desfasados con d=[-1,-180].

Se puede observar que los puntos más altos de correlación están presentes precisamente al principio y al final del rango, y su valor es alrededor de 0.8. Esto indica una fuerte correlación lineal entre los datos con un período de 180 días, lo cual es significativo ya que coincide con la periodicidad de los picos de consumo eléctrico que también ocurren cada 180 días, como habíamos observado previamente en la Subsección 3.3.1.

En la Tabla 3.1 se recogen los 15 desfases temporales con mayor correlación y sus valores.

Tabla 3.1 Tabla de Correlaciones de A1@d.

A1@d	
d	**abs(Corr. k+7)**
-7	0.8429
-14	0.7703
-1	0.7647
-21	0.7043
-175	0.6780
-6	0.6768
-8	0.6500
-28	0.6392
-168	0.6391
-13	0.6093
-161	0.5931
-2	0.5927
-15	0.5884
-35	0.5622
-5	0.5535

Ahora, se hace lo mismo para la serie del consumo horario utilizando el mismo rango de desfases. Sin embargo, en este caso, hay que considerar que al abarcar desde -1 hasta -180, se están desfasando los datos de los últimos 7 días y medio, dado que la unidad temporal es la hora.

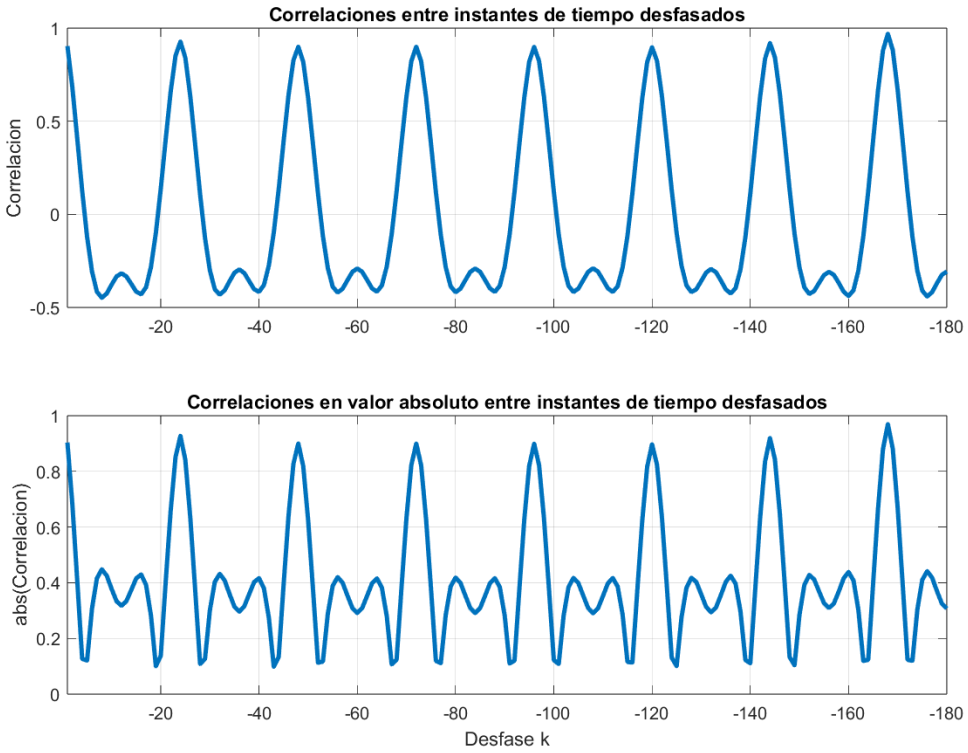


Figura 3.6 Correlaciones entre los datos de (k+d) desfasados con d=[-1,-180].

En la Figura 3.6 se observa una marcada tendencia periódica, con picos de casi el máximo de correlación lineal situados cada 24 horas exactas. Esto se trata de un hallazgo muy relevante ya que sugiere que los consumos horarios tienden a ser muy similares con una diferencia de 24 horas. Se recogen los 15 desfases temporales con mayor correlación y sus valores en la Tabla 3.2

Tabla 3.2 Tabla de Correlaciones de A2@d.

A2@d	
d	**abs(Corr. k+7)**
168	0.9696
24	0.9273
144	0.9197
1	0.9036
48	0.8999
72	0.8998
96	0.8991
120	0.8969
169	0.8811
167	0.8792
23	0.8510
145	0.8441
25	0.8407
143	0.8350
47	0.8272

Se puede apreciar que todos los desfases de la Tabla 3.2 son múltiplos de 24 o están cerca de uno. Estos serán los desfases que se emplearán como entradas autorregresivas de tipo A2@d en nuestros modelos.

3.6 Modelado ARX

Se comienza realizando pruebas de creación de modelos ARX. Como se mencionó previamente, estas pruebas se dividen en dos enfoques principales: el enfoque diario y el enfoque horario, y, a su vez, se subdividen según la variable de salida que se busque predecir. Para comenzar, se centra en las pruebas destinadas a la predicción del consumo diario de energía eléctrica.

3.6.1 Predicción del consumo diario (A1@d)

Se realizan dos pruebas distintas. En la primera prueba, nuestro objetivo será anticipar el consumo total diario del día siguiente, para lo cual utilizaremos una variable de salida con un desfase positivo de 1. En la segunda prueba, nos proponemos predecir los valores del consumo diario de una semana en el futuro, empleando una variable de salida con un desfase positivo de 7.

En todos los ficheros de configuración especificaremos la siguiente partición del conjunto de datos: un subconjunto de entrenamiento, que comprenderá el 80 % de los datos, y un conjunto de prueba con el 20 % restante.

Los parámetros FactP y Gamma de los modelos ARX se mantendrán constantes en 10 y 1, respectivamente. Estos valores se han establecido como óptimos tras llevar a cabo numerosas pruebas. Por lo tanto, para cada archivo CX, se generará un único modelo ARX, ya que no se realizará ninguna variación en los parámetros de entrenamiento.

Los desfases temporales utilizados como entradas autorregresivas serán los presentes en las tablas de las mayores correlaciones lineales del consumo diario y horario de la Sección 3.5. Se seguirá el método iterativo de inclusión progresiva de entradas explicado paso por paso al final del Capítulo 2 con el fin de determinar que combinación de entradas da lugar a las mejores predicciones.

Predicción del consumo diario del día siguiente (Salida A1@1)

Se recogen los modelos resultantes de cada iteración de la prueba en la Tabla 3.3. Este es el formato que se utilizará para presentar los resultados de todas las pruebas a lo largo de este TFG. Se incluyen tanto el SMAPE como el ECM de los subconjuntos de prueba y entrenamiento de los modelos, así como el valor promedio de los mismos, y se marca en negrita el modelo con los errores más bajos. Los errores de EAM y MAAPE serán omitidos, ya que resultan redundantes.

Tabla 3.3 Inclusión progresiva de entradas autorregresivas de tipo A1@d (Salida A1@1).

ID ARX	ID CX	Entradas	$SMAPE_{\text{Medio}}$	$SMAPE_{\text{CE}}$	$SMAPE_{\text{CP}}$	ECM_{Medio}	ECM_{CE}	ECM_{CP}
1005	6	-6	0.0383	0.0396	0.0370	0.5327	0.5375	0.5279
1017	18	-6, -1	0.0365	0.0379	0.0351	0.5094	0.5169	0.5019
1033	34	-6, -1, -8	0.0334	0.0345	0.0323	0.4799	0.4846	0.4752
1048	49	-6, -1, -8, -13	0.0319	0.0333	0.0305	0.4613	0.4675	0.4551
1062	63	-6, -1, -8, -13, -15	0.0314	0.0328	0.0301	0.4558	0.4633	0.4482
1069	70	-6, -1, -8, -13, -15, -28	0.0313	0.0324	0.0302	0.4529	0.4565	0.4493
1077	**78**	**-6, -1, -8, -13, -15, -28, -21**	**0.0313**	**0.0324**	**0.0303**	**0.4529**	**0.4564**	**0.4494**
1087	88	-6, -1, -8, -13, -15, -28, -21, -175	0.0311	0.0311	0.0310	0.4735	0.4667	0.4804
1092	93	-6, -1, -8, -13, -15, -28, -21, -175, -7	0.0316	0.0315	0.0317	0.4838	0.4774	0.4902

En la Tabla 3.3 se puede ver que el mejor modelo es el **1077**, creado con el **CX 78** el cual utiliza como entradas autorregresivas de consumo diario los desfases: -6, -1, -8, -13, -15, -28, -21.

Se representa el ECM promedio en negro, así como de los subconjuntos de prueba (rojo) y entrenamiento (azul) de cada modelo en la Figura 3.7. La representación visual de los errores hace que sea aún más evidente que el modelo superior es aquel creado mediante el CX 78, el cual presenta un **ECM Medio de 0.4529**.

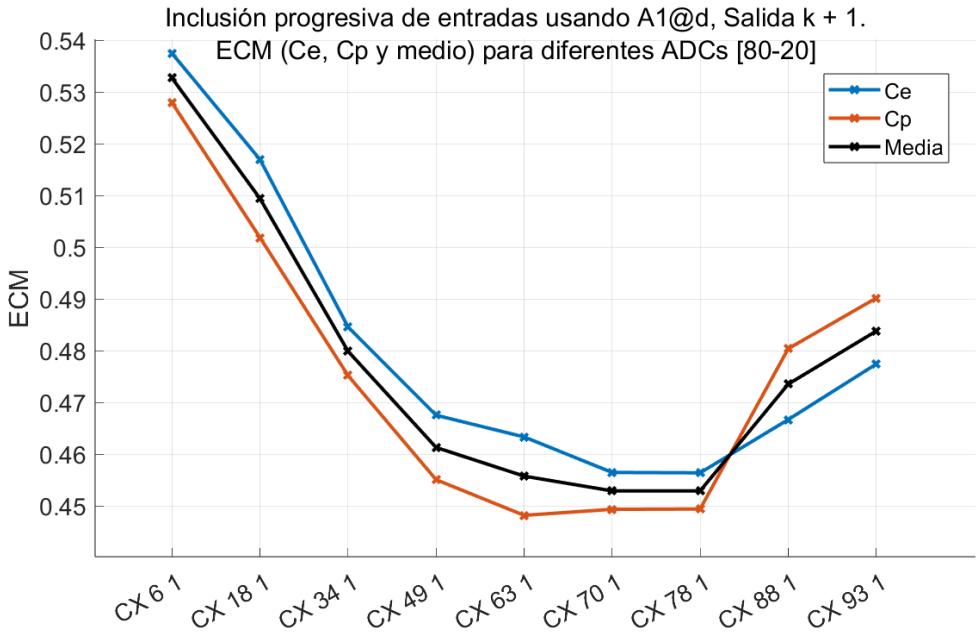


Figura 3.7 Inclusión progresiva de entradas de tipo A1@d (Salida A1@1).

Se procede a incluir al modelo 1077 obtenido en esta pueba las entradas exógenas para el consumo diario definidas en Subsección 3.4.1. La tabla Tabla 3.4 muestra los resultados de esta prueba. Se parte del conjunto de entradas autorregresivas del CX 78, y en cada iteración (fila de la tabla) se agrega una entrada exógena adicional. Estas entradas se acumulan, lo que significa que en cada fila se está utilizando la entrada exógena correspondiente y todas las entradas exógenas de las filas anteriores.

Tabla 3.4 Inclusión progresiva de entradas exógenas (Salida A1@1).

ID ARX	ID CX	Entradas	$SMAPE_{\text{Medio}}$	$SMAPE_{\text{CE}}$	$SMAPE_{\text{CP}}$	ECM_{Medio}	ECM_{CE}	ECM_{CP}
1077	78	A1@d	0.0313	0.0324	0.0303	0.4529	0.4564	0.4494
1206	207	D2@d	0.0312	0.0322	0.0301	0.4516	0.4552	0.4481
1207	**208**	**T1@d**	**0.0311**	**0.0322**	**0.0301**	**0.4515**	**0.4552**	**0.4477**
1211	211	D1@d	0.0313	0.0322	0.0303	0.4524	0.4545	0.4503

Se detiene a partir de la tercera entrada exógena incluida ya que se observa que los errores apenas muestran una disminución en comparación con el modelo inicial 1077, lo que sugiere que las entradas exógenas no tienen un impacto significativo en este caso, ya que las entradas autorregresivas por sí solas han proporcionado resultados decentes.

Se determina como modelo final resultante el **1207**, con un **0.4507 de ECM Medio**, y cuyas entradas son:

- **A1@d** con d = -6, -1, -8, -13, -15, -28, -21.
- **D2@d**: Coseno del día de la semana.
- **T1@d**: Temperatura media del día anterior.

En la Figura 3.8 se representan las salidas reales frente a las predichas por este modelo para hacernos una idea mejor de la calidad de las predicciones.

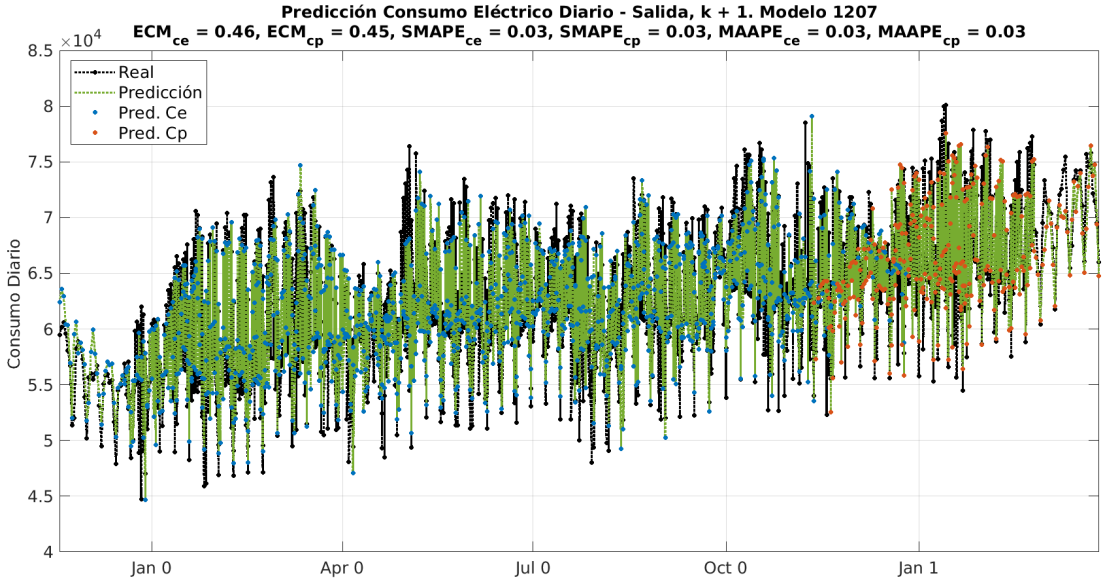


Figura 3.8 Salidas reales frente a las predichas por el modelo ARX 1207.

En verde, se muestran las predicciones del modelo, mientras que en rojo y azul se presentan los puntos discretos que representan el consumo eléctrico para los conjuntos de entrenamiento y prueba, respectivamente. Los valores reales se muestran en negro. A primera vista, parece que las predicciones en verde se superponen en gran medida con los valores reales en negro.

Predicción del consumo diario de dentro de 7 días (Salida A1@7)

Se realiza la misma prueba pero esta vez utilizando como salida el desfase k+7 con el objetivo de predecir los valores del consumo diario de una semana en el futuro.

Tabla 3.5 Inclusión progresiva de entradas autorregresivas de tipo A1@d (Salida A1@7).

ID ARX	ID CX	Entradas	$SMAPE_{\text{Medio}}$	$SMAPE_{\text{CE}}$	$SMAPE_{\text{CP}}$	ECM_{Medio}	ECM_{CE}	ECM_{CP}
1099	100	-7	0.0461	0.0486	0.0435	0.6271	0.6380	0.6163
1115	116	-7, -1	0.0445	0.0468	0.0423	0.6024	0.6150	0.5899
1132	133	-7, -1, -8	0.0427	0.0446	0.0408	0.5803	0.5924	0.5681
1141	142	-7, -1, -8, -14	0.0426	0.0440	0.0411	0.5740	0.5821	0.5660
1162	163	-7, -1, -8, -14, -15	0.0417	0.0432	0.0402	0.5666	0.5753	0.5578
1166	167	-7, -1, -8, -14, -15, -21	0.0416	0.0431	0.0401	0.5644	0.5731	0.5557
1184	**185**	**-7, -1, -8, -14, -15, -21, -35**	**0.0416**	**0.0430**	**0.0402**	**0.5619**	**0.5701**	**0.5536**
1192	193	-7, -1, -8, -14, -15, -21, -35, -161	0.0401	0.0380	0.0422	0.5838	0.5544	0.6131
1195	196	-7, -1, -8, -14, -15, -21, -35, -161, -14	0.0401	0.0380	0.0422	0.5838	0.5544	0.6131

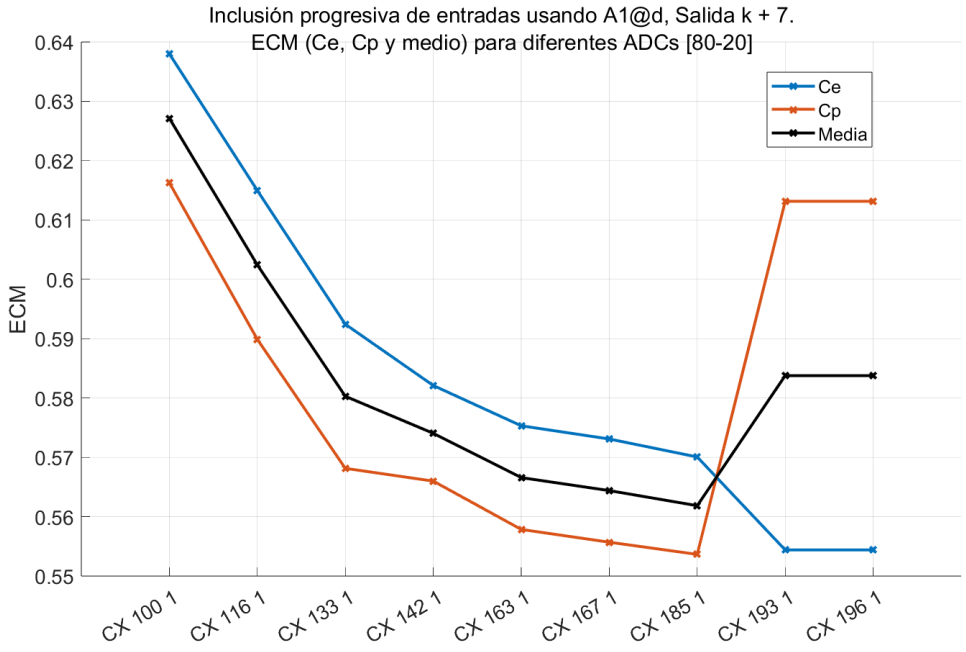


Figura 3.9 Inclusión progresiva de entradas de tipo A1@d (Salida A1@7).

Se observa tanto en la Tabla 3.5 como en la Figura 3.9 que el mejor modelo es el **1184** correspondiente al **CX 185**. Como se aprecia en la gráfica, a partir de este modelo, al incluir más entradas en las siguientes dos iteraciones, el ECM del conjunto de prueba se dispara.

El modelo **1184** cuenta con un ECM Medio de **0.5619**, unas 10 décimas por encima del mejor modelo de la prueba anterior. Esto sugiere una conclusión bastante evidente: al ampliar la ventana temporal hacia el futuro, la calidad de las predicciones tiende a disminuir.

Se prueba a incluir de nuevo las entradas exógenas diseñadas para el consumo diario, esta vez partiendo del modelo 1184.

Tabla 3.6 Inclusión progresiva de entradas exogenas (Salida A1@7).

ID ARX	ID CX	Entradas	$SMAPE_{\text{Medio}}$	$SMAPE_{\text{CE}}$	$SMAPE_{\text{CP}}$	ECM_{Medio}	ECM_{CE}	ECM_{CP}
1184	185	A1@d	0.0416	0.0430 0.0402	0.5619	0.5701	0.5536	
1213	213	C1@d	0.0412	0.0420	0.0405	0.5581	0.5588	0.5574
1218	218	D2@d	0.0406	0.0414	0.0397	0.5519	0.5540	0.5498
1219	**219**	**T1@d**	**0.0406**	**0.0414**	**0.0398**	**0.5506**	**0.5534**	**0.5479**
1221	221	D1@d	0.0414	0.0411	0.0418	0.5624	0.5470	0.5779

Se comprueba que para esta prueba las entradas exógenas han funcionado ligeramente mejor que para la anterior prueba, logrando bajar hasta una dos décimas el ECM Medio del modelo inicial.

El modelo resultante es el **1219**. Exhibe un ECM Medio de **0.5506** e incluye las entradas del CX 219, las cuales son las siguientes:

- **A1@d** con d = -7, -1, -8, -14, -15, -21, -35.
- **C1@d**: Identificador para cada día de la semana.
- **D2@d**: Coseno del día de la semana.
- **T1@d**: Temperatura media del día anterior.

Comparativa de resultados en la predicción del consumo diario en función de la salida

A continuación, en la Figura 3.10, se representa en un plano los errores de los de los modelos ARX más destacados de las dos pruebas anteriores. El eje vertical representa el error para los conjuntos de prueba, mientras que el eje horizontal refleja los errores correspondientes al conjunto de entrenamiento. Esto nos proporciona una visión más precisa de la calidad de las predicciones en función de la longitud de la ventana temporal que estamos tratando de predecir.

La calidad de la predicción de cada modelo será mayor cuanto más cerca estén sus errores correspondientes de la esquina inferior izquierda de cada subgráfica.

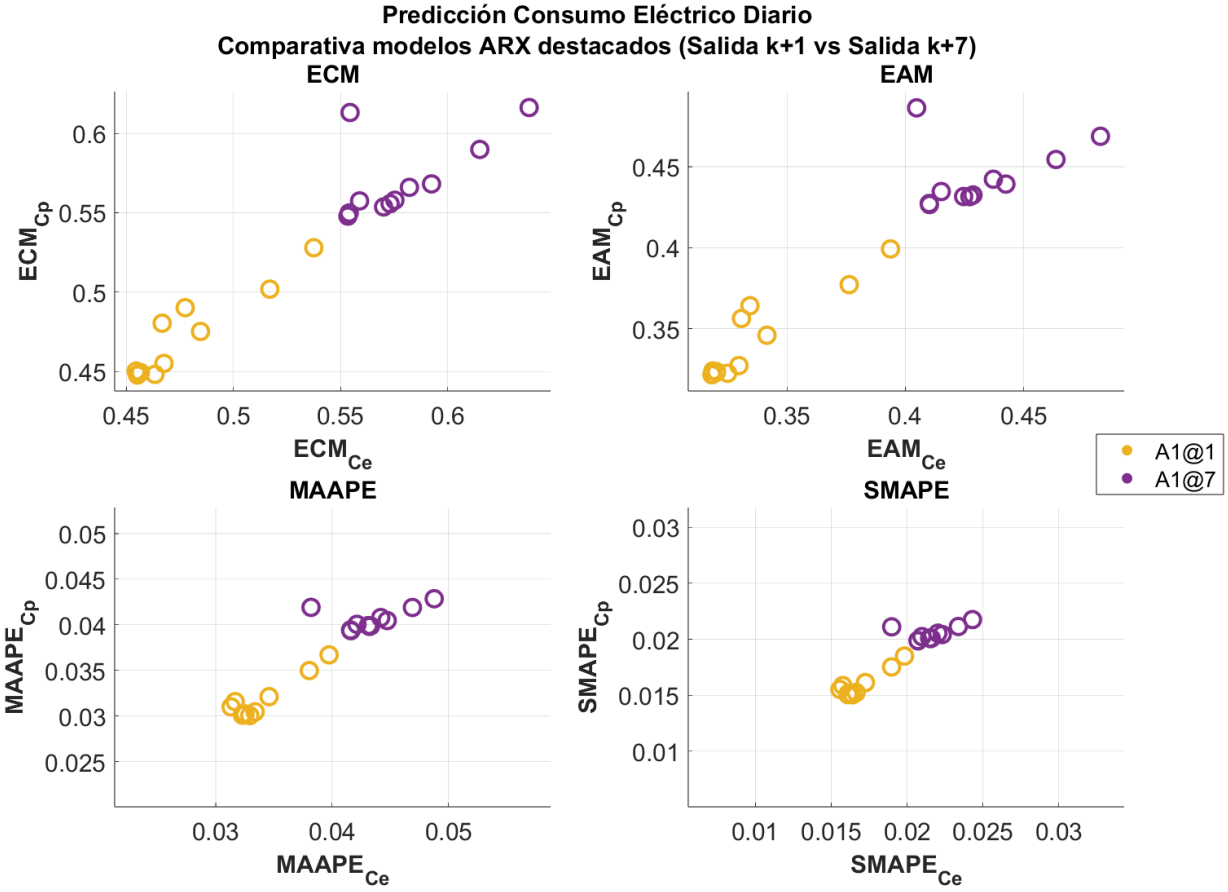


Figura 3.10 Comparativa Modelos ARX (A1@1 vs A1@7).

Como se puede observar, los modelos que predicen para una salida k+1, representados en amarillo, presentan errores menores en las 4 subgráficas de las distintas métricas, tanto en el conjunto de prueba como en el de entrenamiento en comparación con los modelos que predicen para una salida k+7, representados en morado.

Esto confirma que cuanto más lejano sea el horizonte temporal que se intente predecir, peores resultados se obtendrán.

3.6.2 Predicción del consumo horario (A2@d)

Para esta prueba se tratará de predecir únicamente el consumo horario de la hora siguiente, es decir, se utilizará en todo momento una salida con desfase k+1.

En la sección Sección 3.5 se vio que los desfases múltiplos de 24 tenían un valor de correlación lineal muy elevado para el consumo horario, por lo que esperamos tener unos resultados incluso mejores que aquellos que obtuvimos para el consumo diario. Especialmente en los modelos ARX, que se caracterizan por su capacidad para capturar relaciones lineales en los datos.

Se recopila en la Tabla 3.7 los modelos resultantes de cada iteración de la prueba.

Tabla 3.7 Inclusión progresiva de entradas autorregresivas de tipo A2@d (Salida A2@1).

ID ARX	ID CX	Entradas	$SMAPE_{\text{Medio}}$	$SMAPE_{\text{CE}}$	$SMAPE_{\text{CP}}$	ECM_{Medio}	ECM_{CE}	ECM_{CP}
1231	231	-167	0.0253	0.0261	0.0245	0.2353	0.2505	0.2200
1246	246	-167, -23	0.0240	0.0249	0.0232	0.2199	0.2323	0.2075
1254	254	-167, -23, -1	0.0236	0.0245	0.0227	0.2168	0.2295	0.2041
1271	271	-167, -23, -1, -169	0.0202	0.0211	0.0193	0.1936	0.2062	0.1811
1284	284	-167, -23, -1, -169, -25	0.0196	0.0205	0.0188	0.1903	0.2016	0.1790
1295	**295**	**-167, -23, -1, -169, -25, -143**	**0.0196**	**0.0205**	**0.0188**	**0.1888**	**0.2001**	**0.1775**
1297	297	-167, -23, -1, -169, -25, -143, -168	0.0196	0.0204	0.0188	0.1889	0.1999	0.1779
1306	306	-167, -23, -1, -169, -25, -143, -168, -24	0.0196	0.0204	0.0188	0.1890	0.1999	0.1780

Se comprueba que los errores son muchísimo más pequeños en comparación con las pruebas del consumo diario. Incluso para el modelo **1231** resultante de la primera iteración que cuenta únicamente con el desfase -167 como entrada, su ECM Medio es aproximadamente la mitad del valor del ECM Medio que tenía el mejor modelo de predicción del consumo diario del día siguiente.

Se representa el ECM de los modelos en la Figura 3.11, y comprobamos de manera visual que el mejor modelo es el **1295**, resultante de la sexta iteración, creado mediante el **CX 295**, cuyas entradas son: -167, -23, -1, -169, -25, -143. Tiene un **ECM Promedio de 0.188**, un valor significativamente más bajo que los valores alrededor de 0.55 y 0.45 que obtuvimos para el consumo diario.

A pesar de los resultados favorables obtenidos hasta el momento, se llevará a cabo una prueba adicional que implica la inclusión del modelo 1295 en el análisis. En esta prueba, se considerarán las entradas exógenas definidas en la Subsección 3.4.1 con el objetivo de evaluar si es posible mejorar aún más la calidad de las predicciones.

Tabla 3.8 Inclusión progresiva de entradas exogenas (Salida A2@1).

ID ARX	ID CX	Entradas	$SMAPE_{\text{Medio}}$	$SMAPE_{\text{CE}}$	$SMAPE_{\text{CP}}$	ECM_{Medio}	ECM_{CE}	ECM_{CP}
1295	295	A2@d	0.0196	0.0205	0.0188	0.1888	0.2001	0.1775
1315	315	A4@d	0.0196	0.0204	0.0188	0.1882	0.1991	0.1772
1319	319	D1@d	0.0196	0.0204	0.0188	0.1881	0.1990	0.1771
1321	**321**	**A3@d**	**0.0196**	**0.0204**	**0.0188**	**0.1879**	**0.1990**	**0.1769**
1323	323	D2@d	0.0196	0.0204	0.0188	0.1879	0.1990	0.1768

Se puede observar que el SMAPE Medio se mantiene constante con la inclusión de las exógenas, pero en cambio, el ECM Medio se reduce unas centésimas.

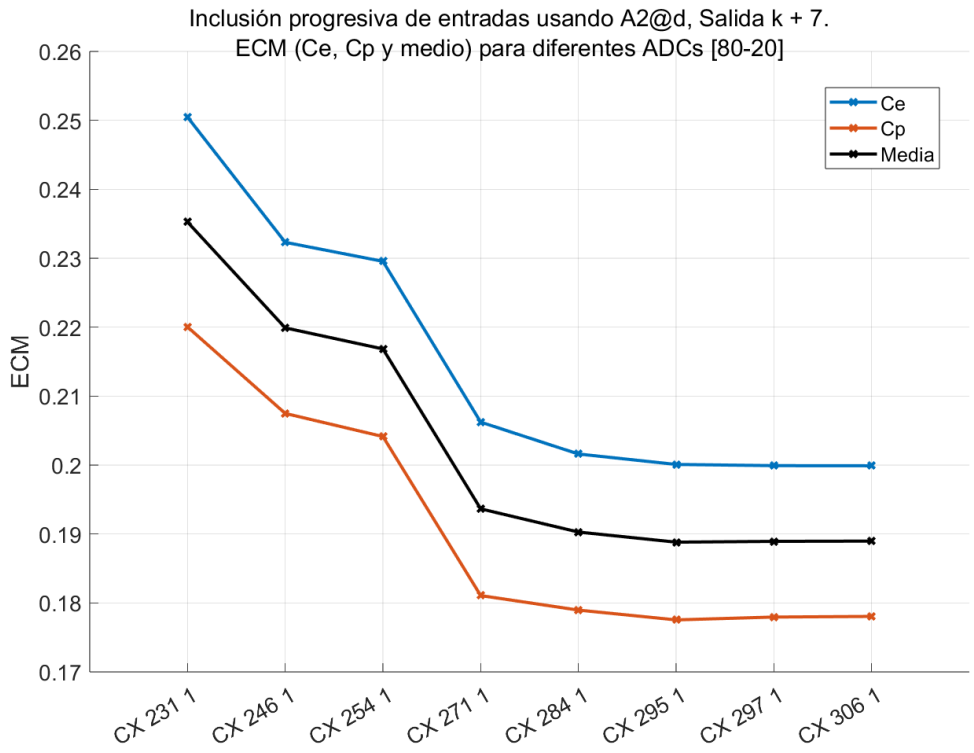


Figura 3.11 Inclusión progresiva de entradas de tipo A2@d (Salida A2@1).

El modelo final resultante de la inclusión de entradas exógenas es el **1321**, el cual exhibe un ECM Medio de **0.1879** y ha sido creado utilizando las entradas:

- **A2@d** con d = -7, -1, -8, -14, -15, -21, -35.
- **A4@d**: Consumo horario del mismo día del año anterior.
- **D1@d**: Seno de la hora del día.
- **A3@d**: Consumo horario del mismo día de la semana anterior.

3.6.3 Comparativa de resultados en la predicción del consumo diario frente al consumo horario

Se representa a continuación, teniendo en cuenta las cuatro métricas de evaluación, en amarillo los errores de los modelos más destacados obtenidos de la predicción del consumo diario para una salida k+1, y en verde los obtenidos de la predicción del consumo horario con una salida k+1.

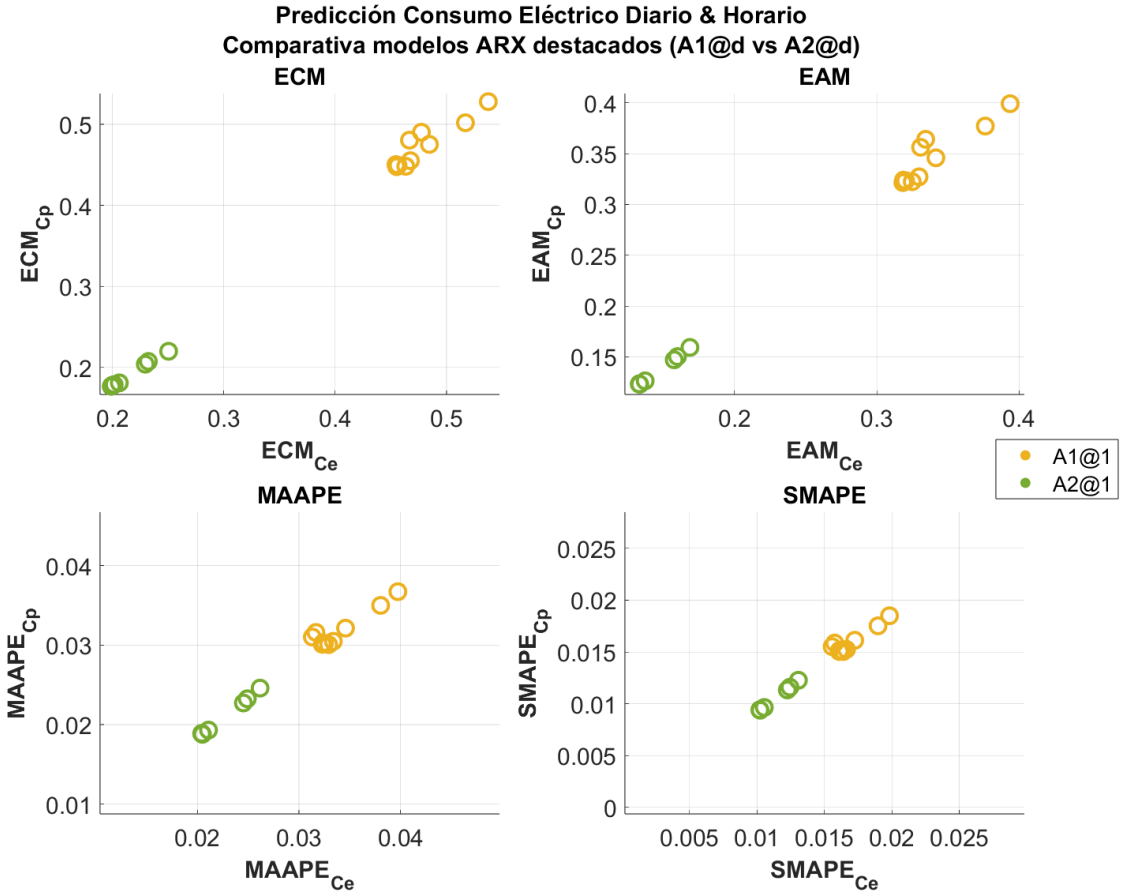


Figura 3.12 Comparativa ARX de consumo diario frente a ARX de consumo horario.

Se verifica una marcada disparidad entre los modelos verdes y amarillos. Esta disparidad se torna especialmente evidente al observar la subgráfica relacionada con el Error Cuadrático Medio (ECM), donde se aprecia una diferencia sustancial de aproximadamente tres décimas.

Es importante destacar que los errores asociados a los modelos destinados a la predicción del consumo horario son notablemente bajos. Esta observación se corrobora en la Figura 3.13, donde se puede apreciar que las líneas verdes correspondientes a las salidas reales, y las negras correspondientes a las predicciones, prácticamente se superponen.

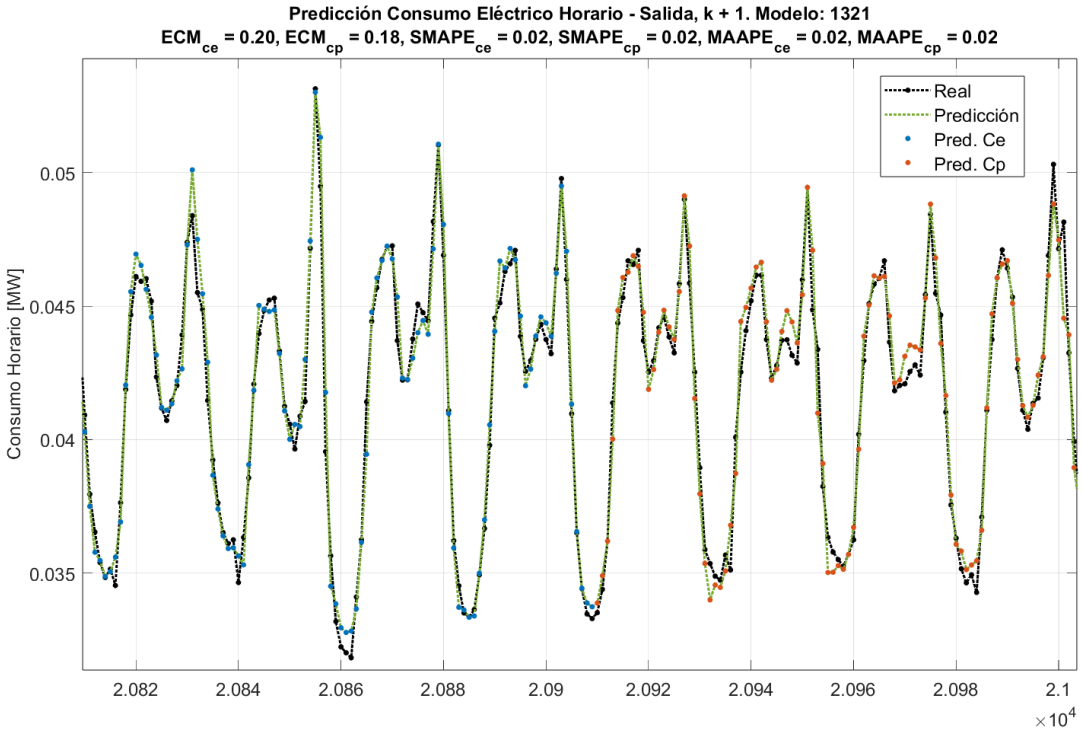


Figura 3.13 Salidas reales frente a las predichas por el modelo ARX 1321.

3.7 Modelado RBF

Como se ha comprobado en la sección anterior, el comportamiento lineal del algoritmo ARX se ha adaptado de manera muy efectiva a esta serie temporal, sobre todo a la hora de predecir el consumo horario. En esta sección, se llevarán a cabo las mismas pruebas con nuestro segundo algoritmo, la red neuronal de base radial. La intención es determinar si esta red neuronal puede obtener resultados aún mejores en comparación con el algoritmo ARX, o si, por el contrario, el ARX es el más adecuado para esta serie.

Al igual que antes, se procede a dividir según estemos tratando de predecir el consumo diario u horario de la serie, y según el horizonte temporal a predecir en la salida.

3.7.1 Predicción del consumo diario (A1@d)

Predicción del consumo diario del día siguiente (Salida A1@1)

En una primera instancia, se lleva a cabo el método iterativo de inclusión progresiva de entradas de la misma forma a como se hizo para los modelos ARX, utilizando los mismos porcentajes para los subconjuntos de entrenamiento y prueba. No obstante, se deben tener en cuenta diferencias significativas en la elección de los parámetros de entrenamiento de los modelos. Mientras que en los modelos ARX se mantuvieron constantes los parámetros factP y gamma en 10 y 1, respectivamente, en el caso de los modelos RBF esta selección se vuelve más compleja.

Se opta por emplear unos parámetros de entrenamiento iniciales para todas las pruebas de inclusión de entradas. Una vez completada esta fase, se identificará el CX con las entradas que hayan proporcionado los mejores resultados y se procederá a realizar un ajuste más detallado de los parámetros. Este proceso de ajuste implicará la creación de conjuntos de modelos variando los valores de los parámetros RBF, con el objetivo de encontrar la combinación de parámetros que minimice el error en las predicciones de manera óptima.

Los parámetros de entrenamiento iniciales que se emplean en esta prueba se presentan en la Tabla 3.9.

Tabla 3.9 Parámetros de entrenamiento iniciales.

Kappa	Nº de Neuronas	Alpha	Nº de Pasadas	Gamma
1 2 3	30 50 75	0.4	40	1

Estos parámetros proporcionan una base sólida para la ejecución de las pruebas de inclusión de entradas en los modelos RBF. La elección de varios valores de kappa y del número de neuronas en esta etapa inicial es una decisión acertada, ya que, de antemano, es complicado determinar con certeza cuál de ellos se adaptará de manera óptima a la serie temporal en cuestión.

Se recogen los modelos resultantes de cada iteración de la prueba en la Tabla 3.10, y se representa graficamente el ECM de cada modelo en la Figura 3.14.

Se puede observar que en la primera iteración, cuando se utiliza un solo desfase de entrada, el

Tabla 3.10 Inclusión progresiva de entradas autorregresivas de tipo A1@d (Salida A1@1).

ID RBF	ID CX	Entradas	$SMAPE_{\text{Medio}}$	$SMAPE_{\text{CE}}$	$SMAPE_{\text{CP}}$	ECM_{Medio}	ECM_{CE}	ECM_{CP}
10080	10005	-6	0.0810	0.0798	0.0823	1.0385	0.9714	1.1057
10294	10017	-6, -1	0.0360	0.0356	0.0364	0.5103	0.4950	0.5256
10336	10033	-6, -1, -8	0.0339	0.0328	0.0349	0.4898	0.4656	0.5140
10453	10048	-6, -1, -8, -13	0.0324	0.0308	0.0340	0.4691	0.4425	0.4957
10619	10054	-6, -1, -8, -13, -7	0.0318	0.0297	0.0339	0.4597	0.4232	0.4961
10716	10071	-6, -1, -8, -13, -7, -2	0.0311	0.0296	0.0326	0.4521	0.4238	0.4804
10771	10081	-6, -1, -8, -13, -7, -2, -15	0.0311	0.0297	0.0326	0.4507	0.4268	0.4747
10871	10090	-6, -1, -8, -13, -7, -2, -15, -35	0.0311	0.0294	0.0329	0.4475	0.4172	0.4778
10935	10098	-6, -1, -8, -13, -7, -2, -15, -35, -5	0.0310	0.0297	0.0324	0.4466	0.4197	0.4734
10984	**10099**	**-6, -1, -8, -13, -7, -2, -15, -35, -5, -14**	**0.0307**	**0.0297**	**0.0317**	**0.4415**	**0.4196**	**0.4634**
11002	10107	-6, -1, -8, -13, -7, -2, -15, -35, -5, -14, -28	0.0325	0.0315	0.0335	0.4592	0.4398	0.4787

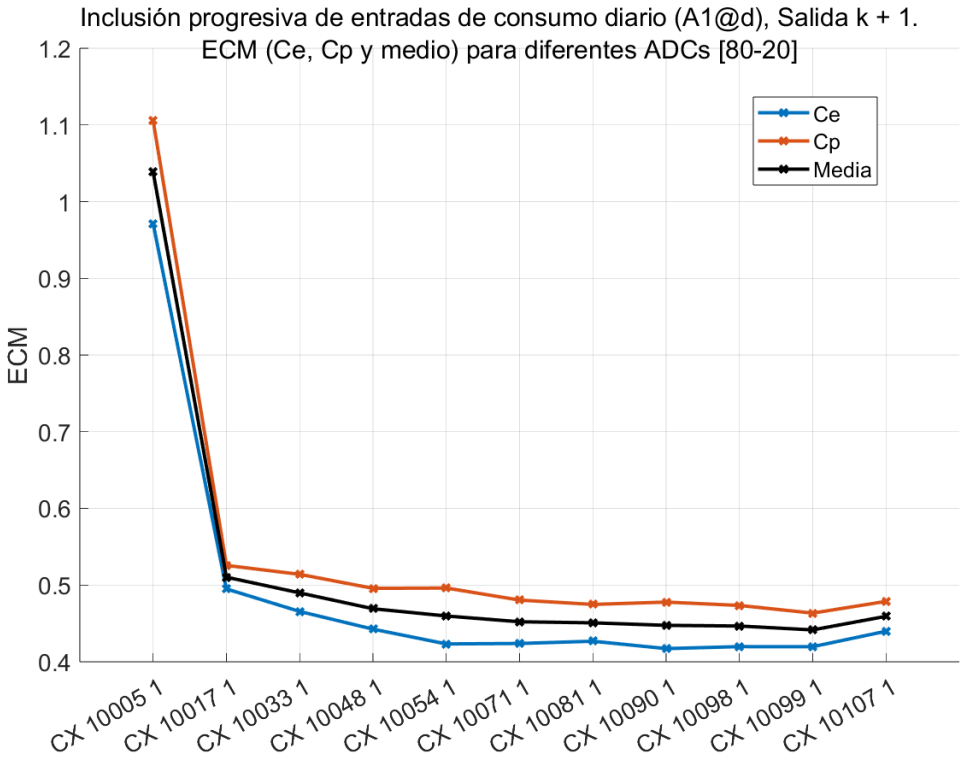


Figura 3.14 Inclusión progresiva de entradas de tipo A1@d (Salida A1@1).

modelo RBF muestra errores notablemente elevados en las predicciones. Sin embargo, estos errores se reducen significativamente en la iteración siguiente al incorporar un segundo desfase de entrada en el modelo.

Este comportamiento sugiere que un único desfase de entrada parece ser insuficiente para capturar los patrones no lineales presentes en la serie, lo que resulta en un error alto. Al agregar más entradas, los modelos RBF comienzan a adaptarse mejor y capturar patrones más complejos.

Los modelos RBF son más flexibles y pueden capturar patrones no lineales más complejos en los datos. Sin embargo, esta flexibilidad también puede requerir más de una entrada para que el modelo pueda ajustarse adecuadamente a los patrones presentes en los datos.

Por otro lado, los modelos ARX son inherentemente lineales y están diseñados para capturar relaciones lineales entre las entradas y las salidas. Por lo tanto, cuando los patrones en los datos son predominantemente lineales, como se ha podido comprobar en la Sección 3.4, los modelos ARX se

adaptan rápidamente y capturan esos patrones en una sola iteración con una única entrada.

Una vez identificado que el mejor modelo es el **10984**, construido a partir del CX **10099**, se procederá a llevar a cabo la búsqueda de la combinación óptima de parámetros de entrenamiento específicamente para dicho CX.

Barrido de parámetros de entrenamiento

Se lleva a cabo una variación conjunta de los valores de kappa y el número de neuronas, ya que son los parámetros que más influyen en el desempeño de los modelos RBF. Específicamente, se crean tandas de modelos en las cuales se varia kappa en un rango de valores que abarca desde 1 hasta 10, y se consideran los valores de 20, 50, 75 y 100 para el número de neuronas. Se mantienen fijos los valores de alpha en 0.4 y el número de pasadas en 40.

La Figura 3.15 presenta tres gráficas que muestran la variación del Error Cuadrático Medio (ECM) en el conjunto de entrenamiento, el conjunto de prueba y el promedio de los modelos en función de los diferentes valores asignados a kappa y al número de neuronas.

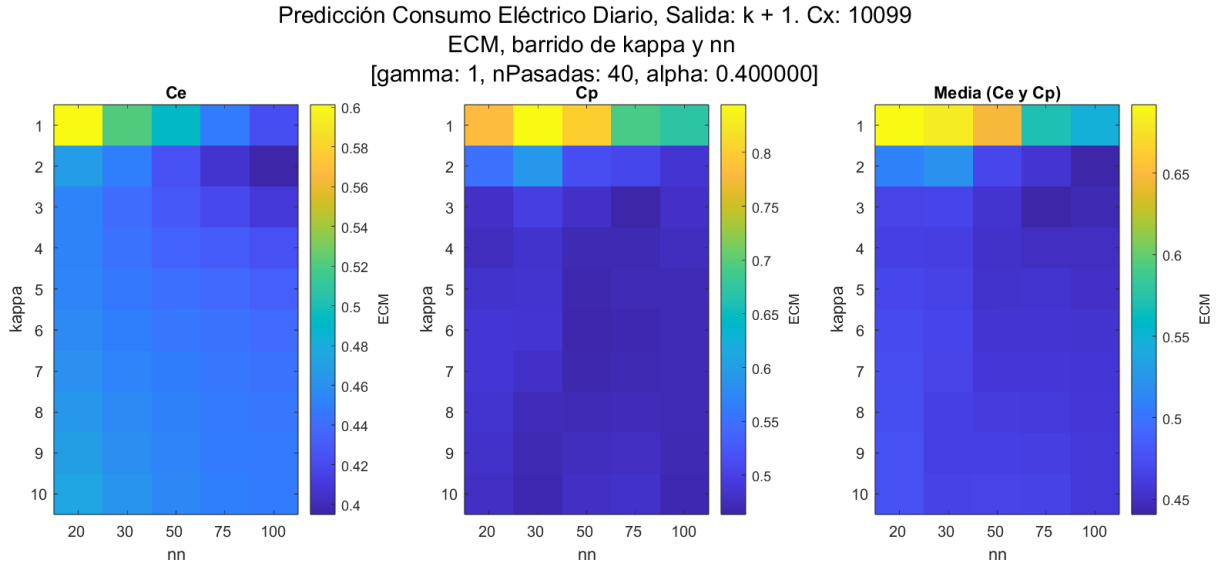


Figura 3.15 Barrido de kappa y el nº de neuronas para el CX 10099.

La observación de que en la gráfica, la zona más oscura (correspondiente a la menor cantidad de error) se encuentra en torno a 75 neuronas y un valor de kappa de 3, es significativa y sugiere una configuración prometedora para estos parámetros.

En la siguiente fase del proceso, se lleva a cabo una búsqueda más detallada al variar nuevamente los mismos parámetros, pero con una resolución mucho mayor. La Figura 3.16 muestra los resultados de esta búsqueda y destaca que los valores ideales para kappa y el número de neuronas son 2.8 y 75, respectivamente.

En la última etapa del proceso, se realiza una exploración de los valores de alpha y el número de pasadas, dos parámetros que generalmente tienen una influencia limitada en los modelos. Después

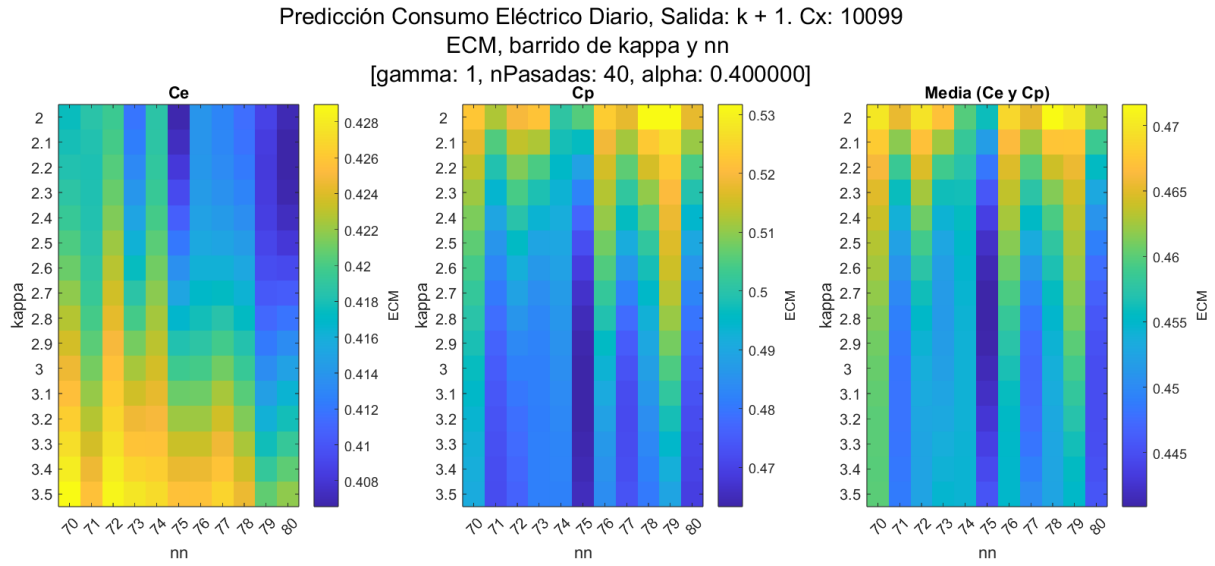


Figura 3.16 Barrido de kappa y el nº de neuronas para el CX 10099.

de crear varias tandas de modelos y analizar los resultados, se observa en la Figura 3.17 que los valores ideales para estos parámetros son 0.4 para alpha y 40 pasadas.

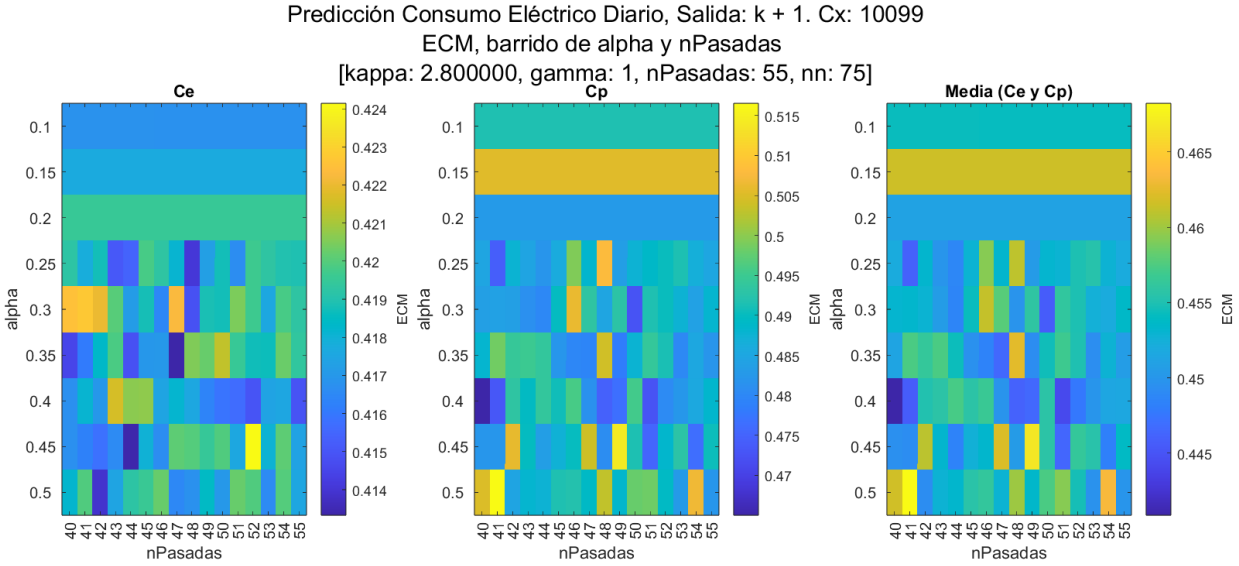


Figura 3.17 Barrido de alpha y el nº de pasadas para el CX 10099.

Por lo tanto, se establecen como parámetros finales del modelo RBF los valores que se presentan en la Tabla 3.11, y el modelo resultante del barrido se identifica como el **11253**, cuyos errores se detallan en la Tabla 3.11. Se observa que presenta un Error Cuadrático Medio (ECM) Medio de **0.4409**, apenas unas centésimas por debajo del modelo resultante previo al barrido de parámetros. En virtud de la mínima reducción de errores evidenciada, se puede concluir que los parámetros iniciales seleccionados para las pruebas se revelan como efectivos, y no se justifica la realización de un barrido de parámetros para cada modelo resultante de las pruebas de esta serie.

Asimismo, como en las pruebas con ARX, se observó que la inclusión de entradas exógenas tampoco redujo significativamente los errores de los modelos, no se considera necesario incorporar estas entradas exógenas en los modelos RBF, ya que no aportan un beneficio sustancial en términos de mejora en la precisión de las predicciones.

Tabla 3.11 Parámetros de entrenamiento del modelo RBF 11253.

Kappa	Nº de Neuronas	Alpha	Nº de Pasadas	Gamma
2.8	75	0.4	40	1

Tabla 3.12 Información del modelo RBF 11253.

ID RBF	ID CX	$SMAPE_{Medio}$	$SMAPE_{CE}$	$SMAPE_{CP}$	ECM_{Medio}	ECM_{CE}	ECM_{CP}
11253	10099	0.0307	0.0296	0.0318	0.4409	0.4167	0.4650

Predicción del consumo diario de dentro de 7 días (Salida A1@7)

Se procede a realizar nuevamente la prueba, utilizando como salida el desfase k+7, con el propósito de predecir los valores del consumo diario de una semana en el futuro.

Tabla 3.13 Inclusión progresiva de entradas autorregresivas de tipo A1@d (Salida A1@7).

ID RBF	ID CX	Entradas	$SMAPE_{\text{Medio}}$	$SMAPE_{\text{CE}}$	$SMAPE_{\text{CP}}$	ECM_{Medio}	ECM_{CE}	ECM_{CP}
11617	10112	-14	0.0810	0.0798	0.0822	1.0384	0.9715	1.1052
11862	10125	-14, -7	0.0460	0.0438	0.0483	0.6212	0.5825	0.6599
11902	10139	-14, -7, -1	0.0446	0.0422	0.0469	0.6047	0.5656	0.6438
12020	10155	-14, -7, -1, -8	0.0432	0.0409	0.0455	0.5832	0.5520	0.6144
12187	10173	-14, -7, -1, -8, -35	0.0420	0.0383	0.0456	0.5572	0.5096	0.6049
12282	**10177**	**-14, -7, -1, -8, -35, -6**	**0.0409**	**0.0384**	**0.0434**	**0.5484**	**0.5098**	**0.5870**
12364	10187	-14, -7, -1, -8, -35, -6, -28	0.0410	0.0386	0.0434	0.5495	0.5112	0.5879
12413	10177	-14, -7, -1, -8, -35, -6, -28, -168	0.0412	0.0379	0.0444	0.5492	0.5050	0.5935

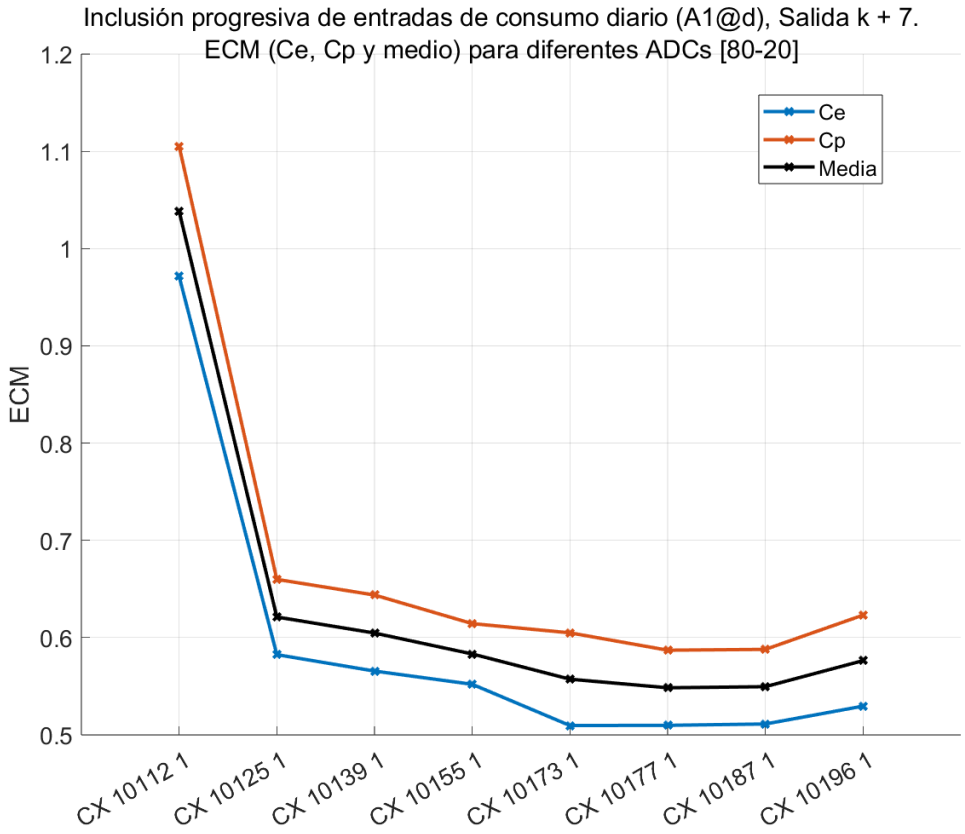


Figura 3.18 Inclusión progresiva de entradas de tipo A1@d (Salida A1@7).

Se observa en la Figura 3.18 un comportamiento similar al de la prueba anterior entre la primera y la segunda iteración, donde se incrementa el número de entradas de una a dos. El modelo óptimo resulta ser el **12282**, construido a partir del CX **10177**, el cual incluye los siguientes desfases: -14, -7, -1, -8, -35 y -6.

Comparativa entre Modelos ARX y RBF en la predicción del consumo diario

Una vez concluidas las pruebas de predicción del consumo diario con los modelos RBF, se procederá a comparar los resultados con aquellos obtenidos con los modelos ARX correspondientes.

En la Figura 3.19, se representan en color rojo los modelos ARX más notables creados para la predicción del consumo diario del día siguiente, mientras que en azul se muestran los modelos RBF del mismo tipo. Esta comparación permite evaluar y contrastar el desempeño relativo de ambos enfoques en términos de precisión de las predicciones.

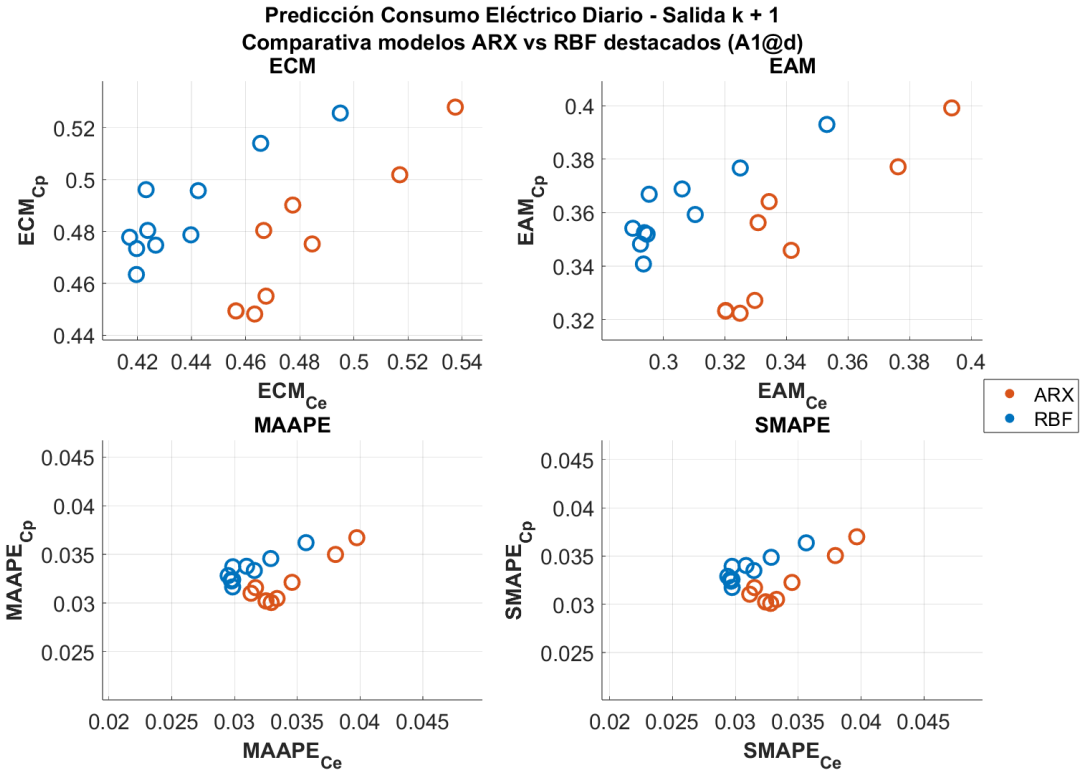


Figura 3.19 Comparativa Modelos ARX vs RBF (Salida A1@1).

En términos generales, se puede observar que los modelos RBF tienden a tener errores superiores en el conjunto de prueba en comparación con los modelos ARX. No obstante, en lo que respecta al conjunto de entrenamiento, los modelos RBF logran una mejor predicción, lo que se refleja en su posición más a la izquierda en las subgráficas.

Cuando la salida se establece en k+7, esta tendencia se acentúa, como se ilustra en la Figura 3.20. Los modelos ARX presentan un mejor rendimiento en la predicción del conjunto de prueba, mientras que los modelos RBF exhiben una mayor precisión en la predicción del conjunto de entrenamiento.

Estos resultados sugieren que, en este contexto específico, los modelos ARX pueden ser más apropiados para la serie, ya que se adaptan mejor al conjunto de prueba.

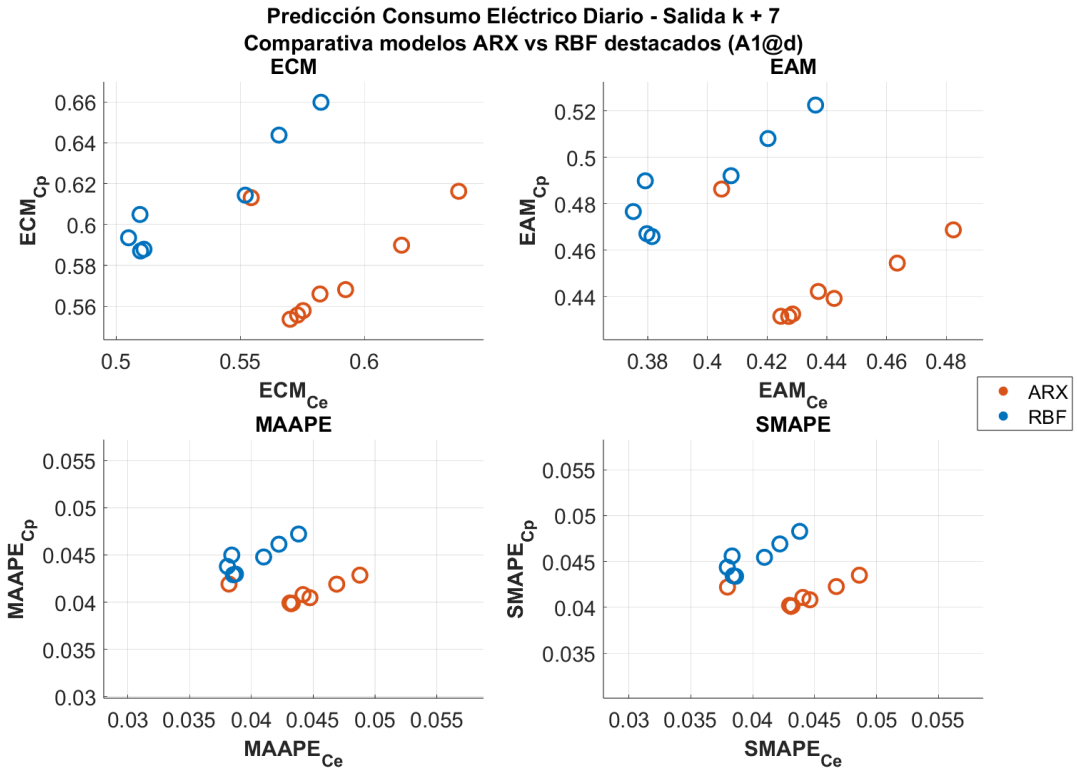


Figura 3.20 Comparativa Modelos ARX vs RBF (Salida A1@7).

3.7.2 Predicción del consumo horario (A2@d)

En la última fase de las pruebas, se han desarrollado modelos RBF con el propósito de predecir el consumo horario de la hora siguiente, utilizando como entrada autorregresiva A2@d. Los resultados de cada iteración se han recopilado en la Tabla 3.14, y en la Tabla 3.14 se presentan los correspondientes Errores Cuadráticos Medios de los modelos.

Tabla 3.14 Inclusión progresiva de entradas autorregresivas de tipo A2@d (Salida A2@1).

ID RBF	ID CX	Entradas	$SMAPE_{\text{Medio}}$	$SMAPE_{\text{CE}}$	$SMAPE_{\text{CP}}$	ECM_{Medio}	ECM_{CE}	ECM_{CP}
20001	20001	-24	0.1241	0.1293	0.1189	0.9734	1.0167	0.9301
20023	20023	-24, -167	0.0243	0.0251	0.0235	0.2258	0.2392	0.2125
20037	20037	-24, -167, -23	0.0230	0.0237	0.0224	0.2134	0.2233	0.2034
20049	20049	-24, -167, -23, -169	0.0230	0.0236	0.0224	0.2125	0.2216	0.2034
20056	**20056**	**-24, -167, -23, -169, -1**	**0.0192**	**0.0200**	**0.0185**	**0.1844**	**0.1943**	**0.1745**
20065	20065	-24, -167, -23, -169, -1,-168	0.0192	0.0200	0.0185	0.1849	0.1947	0.1752
20083	20083	-24, -167, -23, -169, -1,-168, -47	0.0212	0.0221	0.0202	0.1970	0.2088	0.1851

Se observa que en esta prueba, la reducción del error de la primera a la segunda iteración es mucho más significativa en comparación con las pruebas anteriores.

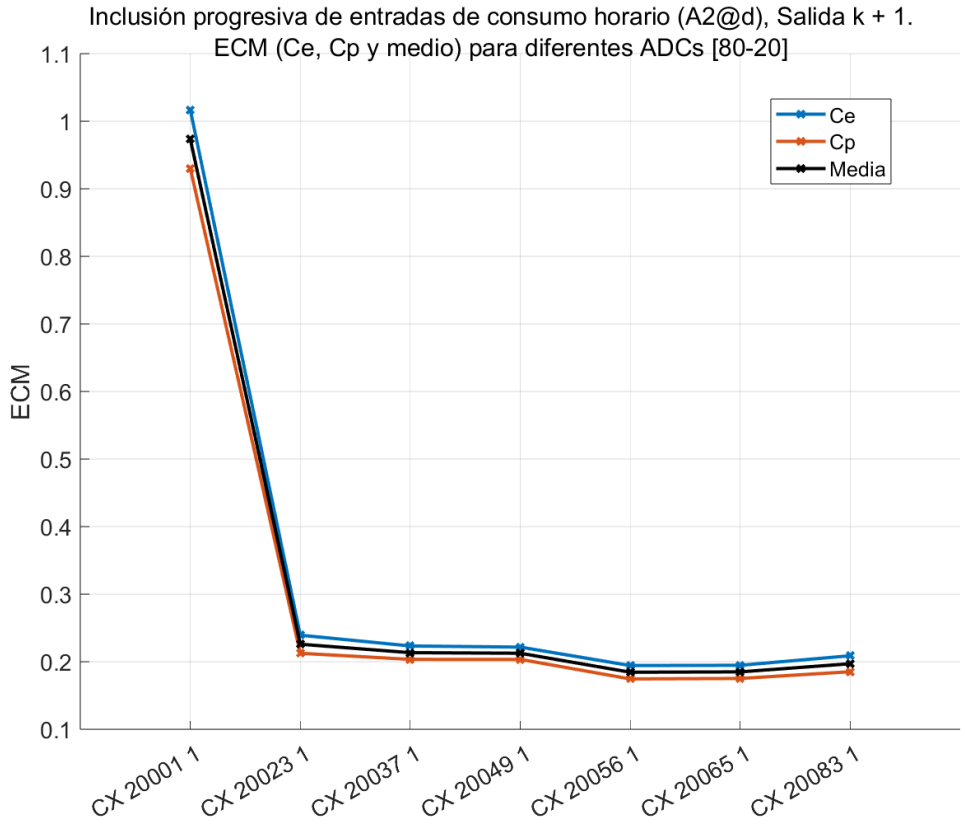


Figura 3.21 Inclusión progresiva de entradas de tipo A2@d (Salida A2@1).

El modelo óptimo es el **20056**, obtenido en la quinta iteración y creado a partir del **CX 20056**, el cual incluye los siguientes desfases: -24, -167, -23, -169 y -1.

Este modelo ha demostrado un desempeño destacado en la predicción del consumo horario de la hora siguiente. Se representan las salidas predichas frente a las reales para este modelo en la Figura 3.22.

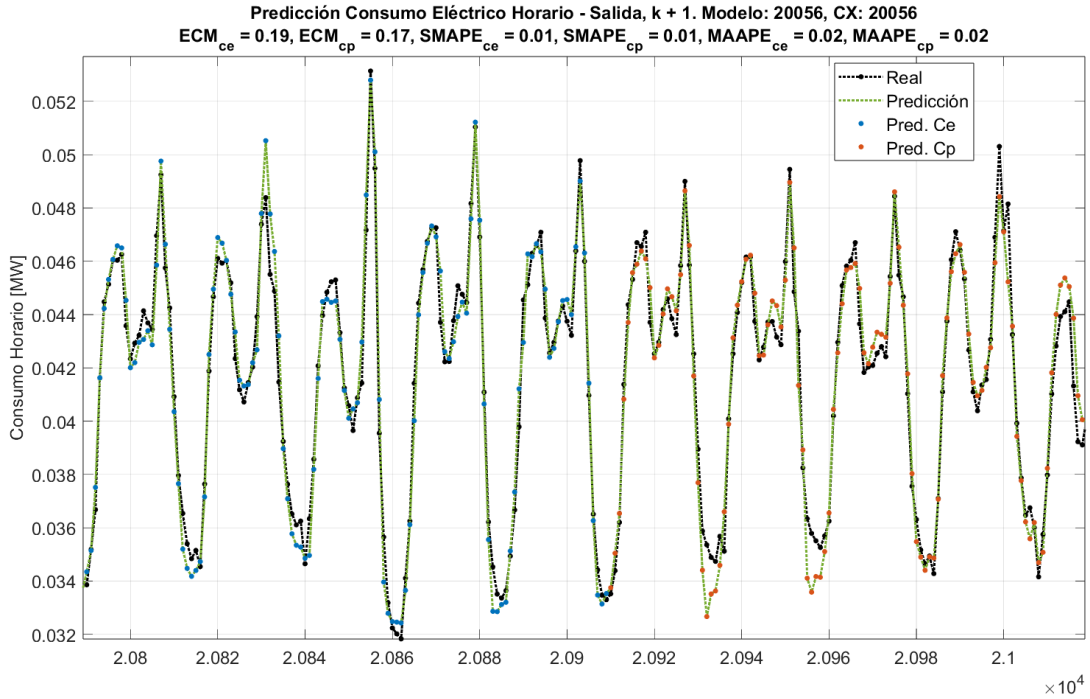


Figura 3.22 Salidas reales frente a las predichas por el modelo RBF 20056.

3.7.3 Comparativa entre Modelos ARX y RBF en la predicción del consumo horario

Al comparar los modelos RBF de esta prueba con los modelos ARX, se observa que sus errores son prácticamente idénticos, lo que dificulta la determinación de cuál enfoque se adapta mejor a la serie en este caso particular. Estos resultados indican que tanto los modelos RBF como los ARX son igualmente efectivos para la tarea de predicción del consumo horario de la hora siguiente en esta serie temporal específica.

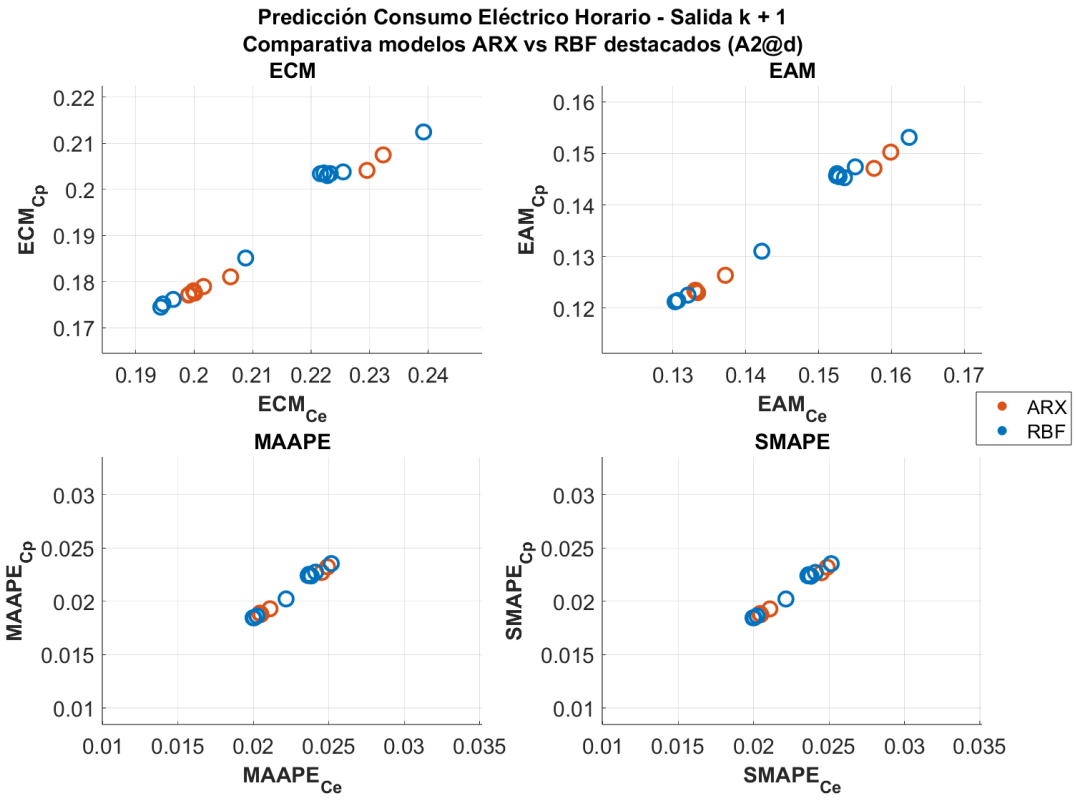


Figura 3.23 Comparativa Modelos ARX vs RBF (Salida A2@1).

La elección entre modelos ARX y RBF depende en gran medida de la naturaleza de los datos y de si los patrones son predominantemente lineales o no lineales. En datos con características predominantemente lineales, los modelos ARX han logrado un buen rendimiento con una única entrada, mientras que los modelos RBF han requerido múltiples entradas para capturar patrones no lineales más complejos.

El enfoque de aproximación mediante redes neuronales de base radial (RBF) es una solución poderosa y flexible que puede adaptarse a una amplia variedad de patrones en los datos, incluyendo patrones no lineales. Sin embargo, en el caso de una serie temporal simple y predominantemente lineal, como la que se ha estado tratando en este capítulo, la complejidad y flexibilidad del modelo RBF pueden no ser necesarias.

En este contexto, el algoritmo ARX ha demostrado ser altamente efectivo al capturar patrones lineales presentes en la serie temporal. Dado que los patrones en los datos son predominantemente lineales, el modelo ARX ha logrado buenos resultados con un enfoque más simple y eficiente. El

uso de modelos RBF podría ser una solución excesiva y compleja para un problema que, como ha quedado demostrado, puede ser abordado de manera efectiva con el algoritmo ARX.

Se puede concluir con que el enfoque del algoritmo ARX ha demostrado ser el más apropiado y eficaz para predecir las series temporales del consumo eléctrico diario y sobre todo, el horario.

4 Aplicación a serie temporal de las transacciones de una entidad financiera

En este nuevo capítulo, al igual que en el anterior, se adapta la metodología de predicción de series temporales a un nuevo conjunto de datos. En esta ocasión, se explora el análisis de una serie temporal de particular interés: las transacciones realizadas en los dispensadores automáticos de una sucursal bancaria, cuyo nombre no se revelará por motivos de confidencialidad.

Cabe destacar que, para garantizar la protección de la información sensible, todas las representaciones gráficas y análisis que se presentan en este capítulo se basan en datos normalizados. Esto se hace con el propósito de preservar la confidencialidad y evitar la divulgación de información financiera real.

Este conjunto de datos proporciona una oportunidad valiosa para aplicar nuestra metodología y evaluar su eficacia en un contexto financiero real. Como se anticipó en el primer capítulo, es altamente plausible que una serie financiera de tales características, albergue atributos y patrones de naturaleza no lineal, lo que dificultará el éxito de las predicciones.

Siguiendo la metodología establecida, se inicia este capítulo abordando las etapas iniciales de recopilación, preprocesamiento y análisis de datos. Estas etapas son fundamentales para comprender la naturaleza de los datos y, por consiguiente, para definir las entradas que se utilizarán en las diversas pruebas de modelado predictivo.

4.1 Recopilación de datos

En primer lugar, los datos en su forma bruta han sido suministrados por el propio banco, ya que este proyecto fue encargado a GAMCO. El fichero de datos que conforma la serie temporal de interés es el de las transacciones realizadas en los dispensadores del banco desde principios de 2019 hasta mediados de 2022.

El banco cuenta con varias sucursales distribuidas por todo el país, y cada sucursal dispone de

uno o varios dispensadores donde los clientes pueden retirar dinero. Cada vez que se realiza una transacción en cualquiera de los dispensadores, ésta se registra de manera automática, junto con la información pertinente, como la fecha y la hora del retiro, la cantidad dispensada, así como otros detalles que no serán revelados por motivos de confidencialidad.

Por lo tanto, el archivo que se encuentra disponible comprende miles de filas, cada una representando una transacción realizada, y numerosas columnas, que contienen información significativa acerca de cada transacción.

Dado que en el país circulan varios tipos de monedas, cada sucursal da lugar a varias series temporales diferentes. Sin embargo, nuestra atención se centra en la predicción de la serie temporal asociada a la moneda nacional de una única sucursal del banco.

4.2 Preprocesamiento de datos

A la hora de llevar a cabo el proceso de filtrado en el extenso archivo de datos, con el objetivo de seleccionar exclusivamente las transacciones asociadas a una sucursal específica, se han tenido en cuenta dos factores principales:

- En primer lugar, es esencial contar con una gran cantidad de datos disponibles para lograr una predicción precisa. Por lo tanto, se buscará una sucursal que cuente con un gran volumen de datos disponibles.

- Por otro lado, se examinará el fichero en busca de datos faltantes, ya que la presencia de un intervalo de fechas significativo con datos faltantes puede entorpecer la predicción.

Dado que resulta complicado determinar de antemano qué sucursales podrían ser las mejores candidatas para la predicción, se ha optado por postergar esta decisión hasta la etapa de análisis subsiguiente. En consecuencia, se procede con el preprocesamiento del archivo de datos completo, y más adelante se centra en la identificación y aislamiento de los datos pertinentes a la sucursal y la moneda seleccionada.

Se conservan todos los campos de la matriz de datos, ya que cada uno de ellos puede proporcionar información valiosa, y se reliza el mapeo de la siguiente manera: para los campos que se encuentren en formato de texto, como el tipo de moneda, se les asigna un número único a cada moneda. A modo de ejemplo, podría asignarse el número 1 a los euros y el número 2 a los dólares. Esta conversión es esencial para asegurar que todos los campos estén en formato numérico, lo cual nos habilitapara transformar la matriz a un formato MAT, compatible con Matlab, donde se llevarán a cabo los filtrados necesarios y demás operaciones de forma automática.

Posteriormente, se asegura la disposición cronológica de los datos, y se lleva a cabo el proceso de limpieza y procesamiento de la matriz de datos de todas las sucursales, eliminando los valores atípicos e inservibles.

Una vez finalizada esta etapa, se pueden llevar a cabo todos los análisis estadísticos pertinentes y crear representaciones gráficas. Esto será fundamental para la selección de una sucursal específica sobre la cual llevar a cabo la predicción en relación a su serie temporal de transacciones.

4.3 Análisis de los datos

Se inicia con un análisis estadístico general de los datos, sin focalizarse en ningún dispensador o sucursal en particular. Con el propósito de seleccionar sucursales candidatas para la predicción, se ha calculado el número total de transacciones realizadas en cada sucursal, así como su proporción en relación con el número total de transacciones de todas las sucursales. Del mismo modo, se ha realizado este cálculo para los dispensadores.

Al realizar un análisis minucioso del archivo, se ha identificado un problema de considerable magnitud: **la ausencia de una gran cantidad de datos** en la mayoría de las series temporales correspondientes a cada sucursal. Esta ausencia de datos posiblemente se deba a una falla en el sistema de registro de las transacciones en los dispensadores. Se ha observado que estos intervalos de fechas con valores vacíos son más frecuentes durante los meses de junio a noviembre de 2021. Es imperativo destacar que este problema es de gran relevancia y no puede pasarse por alto.

Después de realizar diversos cálculos estadísticos en relación con la serie completa de transacciones, finalmente se ha optado por seleccionar una de las sucursales que presenta un mayor volumen de datos disponibles. Además, en esta sucursal, la falta de datos se concentra a partir de un punto específico en el tiempo. Esta circunstancia brinda la oportunidad de recortar la matriz de datos de la sucursal y enfocarse exclusivamente en el período con datos disponibles, que sigue siendo suficientemente grande. De esta manera, se aborda de manera integral el problema de la ausencia de datos en dicho período.

Además, la sucursal cuenta con un único dispensador, lo cual permite simplificar el proceso ya que ahorra el cálculo de tener que sumar los montos dispensados por cada dispensador perteneciente a la sucursal para obtener su serie temporal.

Adicionalmente, es importante destacar que la sucursal seleccionada dispone de un único dispensador. Esta circunstancia simplifica el proceso, ya que no es necesario calcular la suma de los montos dispensados por varios dispensadores pertenecientes a la sucursal para obtener su serie temporal.

4.3.1 Análisis gráfico

En primer lugar, se representa la serie temporal correspondiente a la suma de los saldos de las transacciones diarias de la sucursal elegida en su totalidad, sin llevar a cabo ningún recorte. Esto se hace con el propósito de visualizar el intervalo de fechas en el que se registra la ausencia de datos, tal como se mencionó anteriormente.

Es posible observar en la Figura 4.1 que desde finales de junio hasta noviembre de 2021 existe un período en el cual no se han registrado datos de transacciones. Como se mencionó previamente, este problema se ha solucionado recortando la matriz de datos hasta el 25 de junio de 2021.

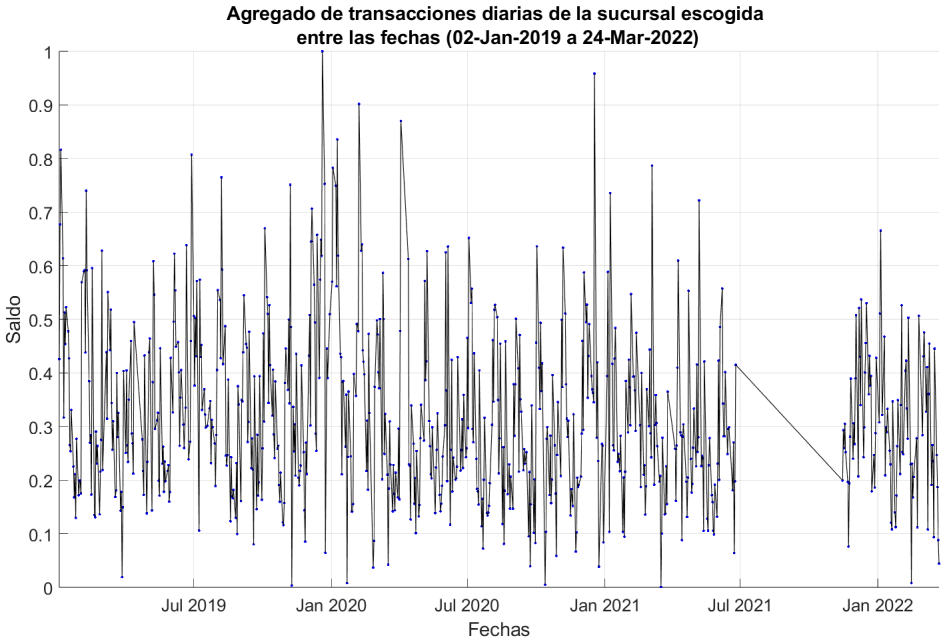


Figura 4.1 Suma de transacciones diarias de la sucursal entre el 02-01-2019 y el 24-03-2022.

En la Figura 4.2 se muestra la versión final de la serie temporal que se utilizará para predecir los saldos de transacciones diarias de la sucursal.

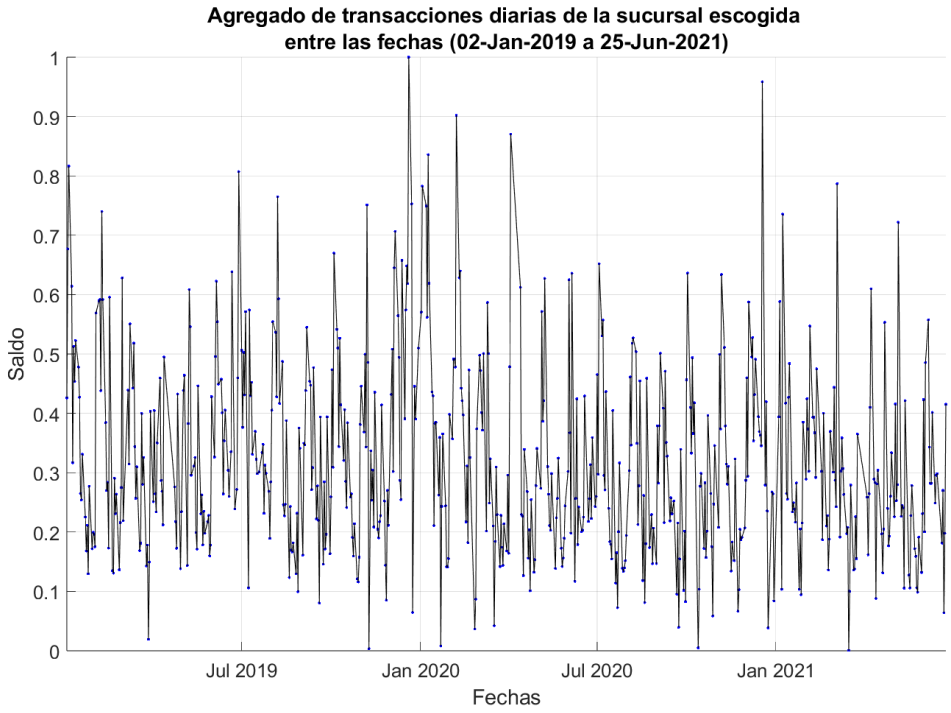


Figura 4.2 Suma de transacciones diarias de la sucursal entre el 02-01-2019 y el 25-06-2021.

Continuando con el análisis de la serie temporal, se procede a realizar una serie de representaciones gráficas con el objetivo de obtener una comprensión más profunda de los patrones de comportamiento de los clientes. Estas visualizaciones serán útiles en etapas posteriores para definir las diversas entradas.

En primer lugar, se representan tanto el número normalizado de transacciones como la suma total de transacciones normalizadas en función de la hora del día.

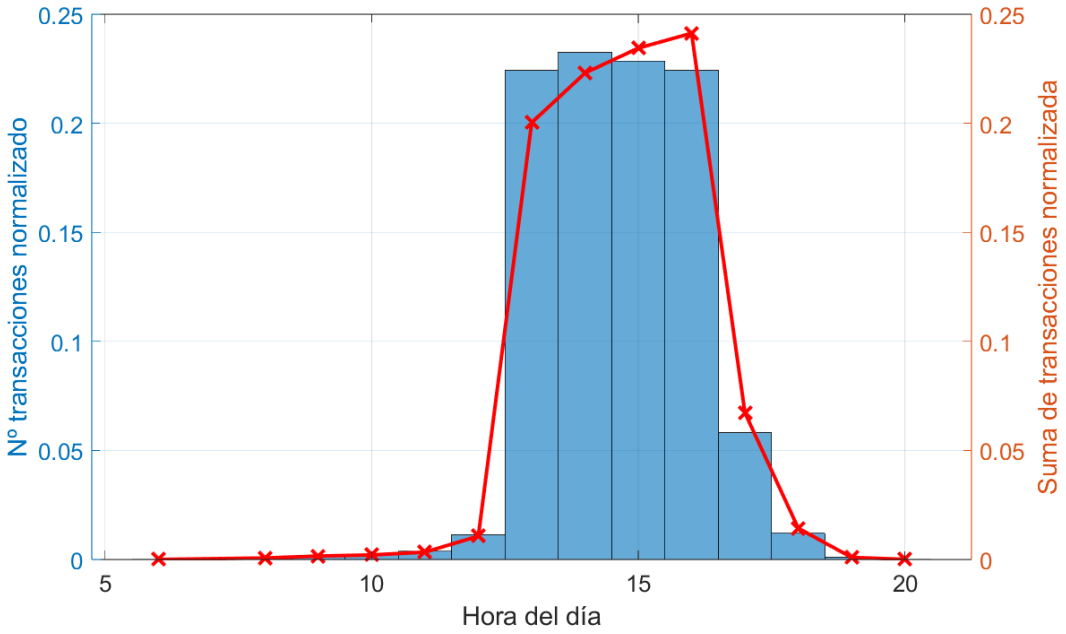


Figura 4.3 Número y suma de transacciones normalizadas por hora.

Esta visualización nos permite examinar los patrones de uso de los dispensadores en función de la hora. Lo que es evidente es que el uso de los dispensadores para esta sucursal muestra una concentración significativa en el rango de horas que va desde las 13:00 PM hasta las 16:00 PM. Durante estas horas, se observa un aumento notable tanto en el número como en la suma total de transacciones.

A lo largo de las demás horas, los valores son considerablemente más bajos. Esta información inicial nos da una idea de los momentos del día en los que se produce una mayor actividad en los dispensadores de la sucursal, lo cual puede ser valioso a la hora de definir entradas exógenas basadas en los horarios de consumo.

A continuación, se procede a representar las mismas variables, es decir, el número normalizado de transacciones y la suma total de transacciones normalizadas, en relación con el día de la semana.

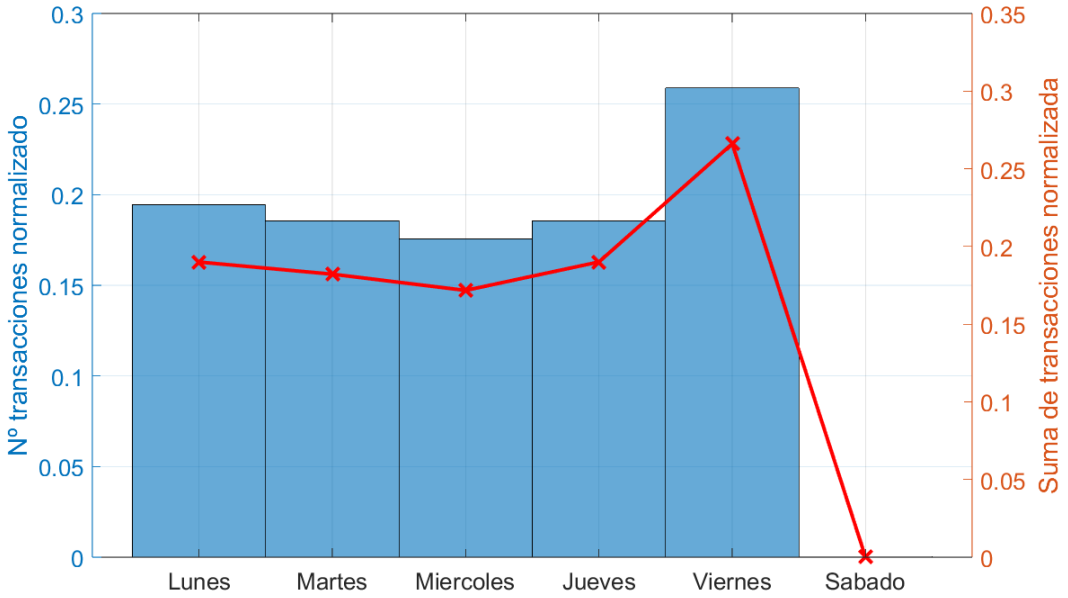


Figura 4.4 Número y suma de transacciones normalizadas por día de la semana.

Esta visualización nos brinda una visión clara de cómo varía la actividad de los dispensadores a lo largo de los diferentes días de la semana.

En primer lugar, es evidente que los datos para el día sábado son prácticamente nulos, mientras que para el día domingo ni siquiera aparecen en la representación debido a la ausencia total de datos. Esto sugiere fuertemente que los dispensadores han de estar cerrados durante los fines de semana debido a la falta de actividad durante estos días.

Además, se destaca un patrón interesante: el pico máximo de uso se registra con considerable diferencia en los días viernes. Este aumento en la actividad sugiere que los clientes tienden a utilizar los dispensadores en mayor medida los días previos al fin de semana, lo que podría estar relacionado con la necesidad de efectivo para sus actividades y gastos durante esos días.

Esta información desempeñará un papel crucial en la definición de las entradas que reflejen estos patrones. En particular, como se explicará más adelante, esta observación será fundamental para la creación de una entrada exógena de tipo calendario denominada "**tipo de día**". A través de esta entrada, se categorizarán los días de la semana en diferentes tipos en función de sus niveles de actividad. A cada tipo de día se le asignará un coeficiente que reflejará la actividad registrada en ese día. Cuanto más activo sea el día, mayor será el coeficiente asignado al tipo de día correspondiente.

De esta manera, esta entrada permitirá al modelo adaptarse de manera más precisa a las variaciones en la actividad de los dispensadores en diferentes días de la semana, contribuyendo así a una mayor precisión en las predicciones.

Siguiendo con el análisis, a continuación se procede a representar tanto el número como la suma de transacciones en función del día del mes.

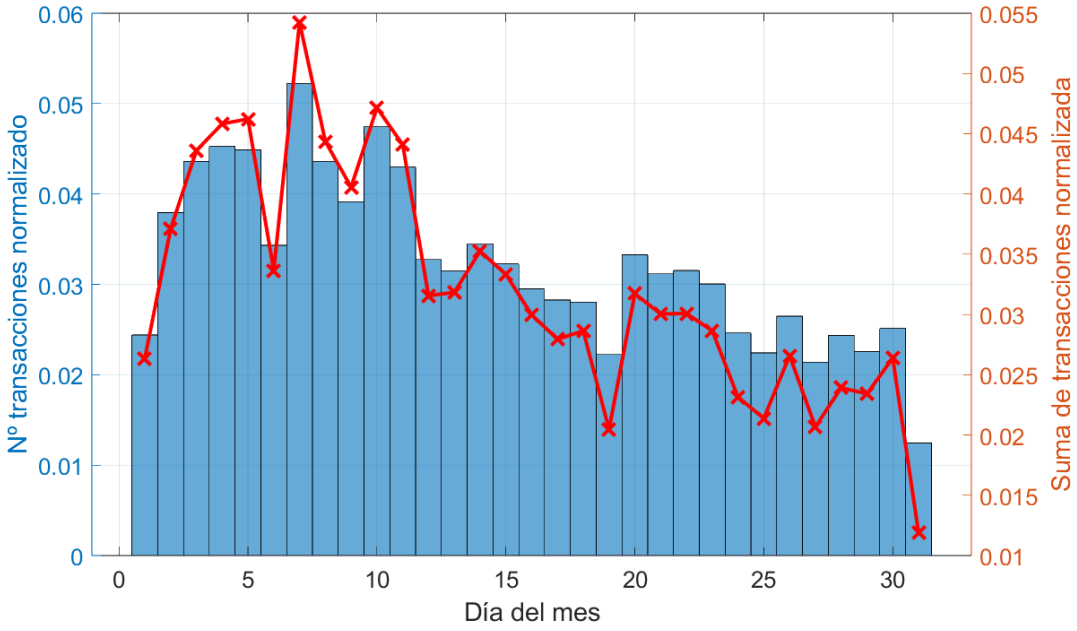


Figura 4.5 Número y suma de transacciones normalizadas por día del mes.

A través de esta visualización, se percibe un patrón caracterizado por un incremento gradual en el número de transacciones durante el primer tercio del mes, seguido por un descenso progresivo a partir de dicho punto. No obstante, también se presentan unos pequeños picos intermitentes de actividad a lo largo del resto del mes.

Esta variación puede sugerir una posible influencia de factores mensuales, como los ciclos de pago o gastos habituales de los usuarios, en los patrones de uso de los dispensadores.

Más adelante, se definen las entradas de tipo calendario que contribuyen a una mejor adaptación del modelo a las fluctuaciones en la demanda de transacciones a lo largo del mes.

Finalmente, se representa el número y la suma de transacciones en función del mes.

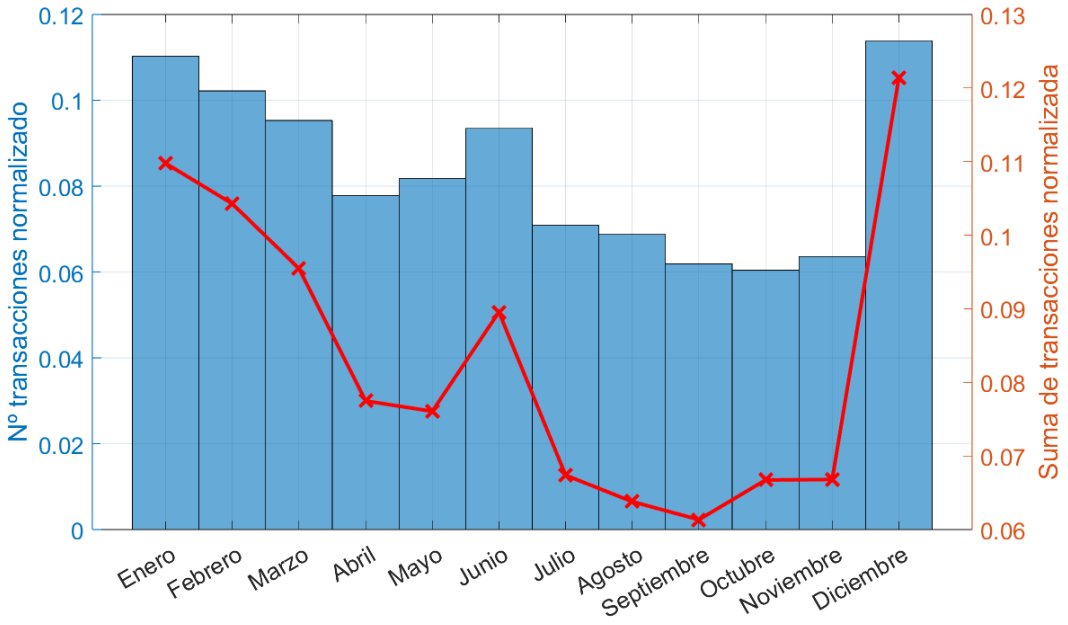


Figura 4.6 Número y suma de transacciones normalizadas por mes.

Se destaca que el número de transacciones es más elevado en diciembre, donde alcanza su punto máximo. Además, los primeros tres meses del año también presentan un mayor número de transacciones en comparación con los demás meses del año. Este comportamiento se debe a varios factores.

En primer lugar, estos meses coinciden con las festividades navideñas, un período en el que el consumo tiende a aumentar considerablemente.

Además, estos meses se alinean con el período de verano en el país donde se encuentra ubicado el banco objeto de nuestras pruebas. Dado que este país se encuentra en el hemisferio sur, es natural que las transacciones se incrementen durante el periodo vacacional, ya que muchas personas aprovechan estas fechas para realizar actividades y compras relacionadas con las vacaciones.

Durante el resto del año, las transacciones se mantienen en niveles similares, con un aumento notorio en el mes de junio, coincidente con la llegada del invierno.

Estos hallazgos resaltan la influencia de las estaciones del año en los patrones de uso de los dispensadores y nos brindan información valiosa para definir entradas que consideren esta variabilidad estacional.

4.4 Definición de entradas

La conclusión más destacada que se puede extraer de estas representaciones gráficas es la **ausencia de datos en los fines de semana**. Esto influye de manera significativa en la definición de las entradas autorregresivas, ya que si estamos utilizando un desfase temporal como entrada o como salida del modelo que corresponda a un fin de semana o a un día festivo, y no existen transacciones en esos días, se registraría un valor nulo. Esto podría entorpecer considerablemente la capacidad de predicción del modelo.

Para abordar este desafío, se ha desarrollado una solución que implica la creación de dos tipos de entradas autorregresivas:

1. **B1@d:** esta entrada tomará el agregado de la cantidad diaria dispensada en los dispensadores de la sucursal el día k+d. En caso de no haber transacciones registradas en el día k+d, se tomará entonces el día anterior más cercano con transacciones.

2. **B2@d:** esta entrada también tomará el agregado de la cantidad diaria dispensada en los dispensadores de la sucursal el día k+d, pero en caso de no existir transacciones para el día k+d, se tomará el día anterior del mismo tipo (por ejemplo, otro viernes) más próximo con transacciones.

Para definir la entrada B2@d, es necesario identificar los tipos de días. Esta distinción puede llevarse a cabo al observar la gráfica de análisis mostrada en la Figura 4.4. En dicha gráfica, se hace evidente que el comportamiento de los viernes es notablemente diferente al de los otros días de la semana. Por lo tanto, se designará al **viernes** como un tipo de día único, mientras que los **lunes, martes, miércoles y jueves** serán considerados como otro tipo de día común.

Estas definiciones permiten preservar la autenticidad de la serie temporal y aseguran que su comportamiento no se vea distorsionado. Ambas entradas serán las principales entradas autorregresivas que se utilizarán como base tanto para el modelado ARX como para el RBF.

En contraste con la serie temporal de la demanda eléctrica, en esta serie se utilizará exclusivamente el desfase k+7 como salida (utilizando la entrada B1@d como base de cálculo para evitar la predicción de valores nulos que puedan confundir al modelo). Esto se debe a que en este contexto es de interés predecir al menos una semana en el futuro, proporcionando al banco un margen de maniobra adecuado para tomar decisiones relacionadas con su actividad.

Definición de entradas exógenas

A continuación, se definen una serie de entradas exógenas que desempeñarán un papel fundamental en la mejora de los resultados de la predicción. En contraste con la serie del consumo eléctrico, debido a la magnitud y complejidad de esta serie, estas entradas exógenas serán cruciales y tendrán un impacto significativo en la capacidad de los modelos para obtener resultados precisos y efectivos.

1. **Entradas que aportan información adicional sobre la propia serie de transacciones (B)**

    - **B3@d:** Número de transacciones en los dispensadores de la sucursal el día k+d.
    - **B4@d:** Media de la cantidad diaria dispensada en los dispensadores de la sucursal los últimos 5 días con transacciones previos al día k+d, incluido el día k+d.

- **B5@d:** Mínima cantidad diaria dispensada en los dispensadores de la sucursal los últimos 5 días con transacciones previos al día k+d, incluido el día k+d.
- **B6@d:** Máxima cantidad diaria dispensada en los dispensadores de la sucursal los últimos 5 días con transacciones previos al día k+d, incluido el día k+d.

La entrada B3@d aporta valiosa información, permitiendo al modelo capturar los patrones de actividad diaria de los clientes. Esto le brinda la capacidad de adaptarse a los días con una mayor o menor demanda de transacciones.

Por otro lado, la entrada B4@d ofrece información que puede contribuir a la comprensión de las tendencias de gasto de los clientes y cómo estas varían en función de los días anteriores.

En cuanto a las entradas B5@d y B6@d, su inclusión puede ser de utilidad al modelo para capturar tanto los valores mínimos como máximos de actividad. Esto es especialmente relevante, ya que estas cifras extremas también ejercen influencia en las predicciones del modelo.

Los días que no tienen transacciones no se cuentan para hacer la media o para coger la cantidad máxima o mínima, por lo que habrá que tomar tantos días anteriores como sea necesario.

2. **Entradas de tipo calendario (C)**

- **C1@d:** Identificador asignado a cada día de la semana.
- **C2@d:** Identificador asignado a cada día del mes.
- **C3@d:** Identificador asignado a cada mes.
- **C4@d:** Coeficiente aplicado a cada tipo de día para ponderar su importancia relativa. Específicamente, se asigna 0.7 para los viernes y 0.4 para el resto de días de la semana.

Estas cuatro entradas adicionales, al igual que en la serie del consumo eléctrico, permiten al modelo considerar a qué día de la semana, día del mes, mes o tipo de día corresponde la salida que se va a predecir. La incorporación de estas entradas permite tener en cuenta la influencia de ciertos días de la semana, meses o estaciones del año en el comportamiento de la serie temporal, lo que enriquece la capacidad predictiva del modelo.

3. **Entradas derivadas (D)**

- **D1@d y D2@d:** seno y coseno del día de la semana respectivamente.
- **D3@d y D4@d:** seno y coseno del día del mes respectivamente.
- **D5@d y D6@d:** seno y coseno del mes respectivamente.
- **D7@d:** semana del mes.
- **D8@d y D9@d:** seno y coseno de la semana del mes respectivamente.

Al igual que para la serie temporal del consumo eléctrico, definimos este tipo de entradas cuyo cálculo se deriva de las entradas de tipo calendario. Están basadas en funciones trigonométricas que permiten al modelo capturar patrones temporales cíclicos de manera más precisa y suave, lo que puede mejorar la capacidad para predecir las fluctuaciones regulares presentes en la serie temporal.

Estas entradas, al igual que las de calendario, únicamente se utilizarán con desfase **7**, pues lo que interesa es que proporcionen información sobre la salida a predecir.

4.5 Análisis de correlaciones de Pearson

En esta sección, se han calculado las correlaciones entre la salida a predecir y los instantes de tiempo desfasados de la serie. Para realizar este cálculo, se han utilizado las dos entradas autorregresivas definidas en la Sección 4.4. Específicamente, se han utilizado los desfases desde el -2 hasta el -60, lo que abarca aproximadamente dos meses de datos. Al limitar el análisis a dos meses, se trata de capturar los patrones más recientes que puedan influir en la variable objetivo. Esto se debe a que a medida que los datos se alejan en el tiempo, es posible que las relaciones entre las variables cambien, y por tanto, siempre es apropiado otorgar una mayor relevancia a los eventos recientes.

Este análisis de correlaciones permite identificar qué desfases temporales tienen una mayor influencia en la salida. En primer lugar, se representará la correlación de Pearson de estos desfases utilizando la entrada autorregresiva B1@d como cálculo.

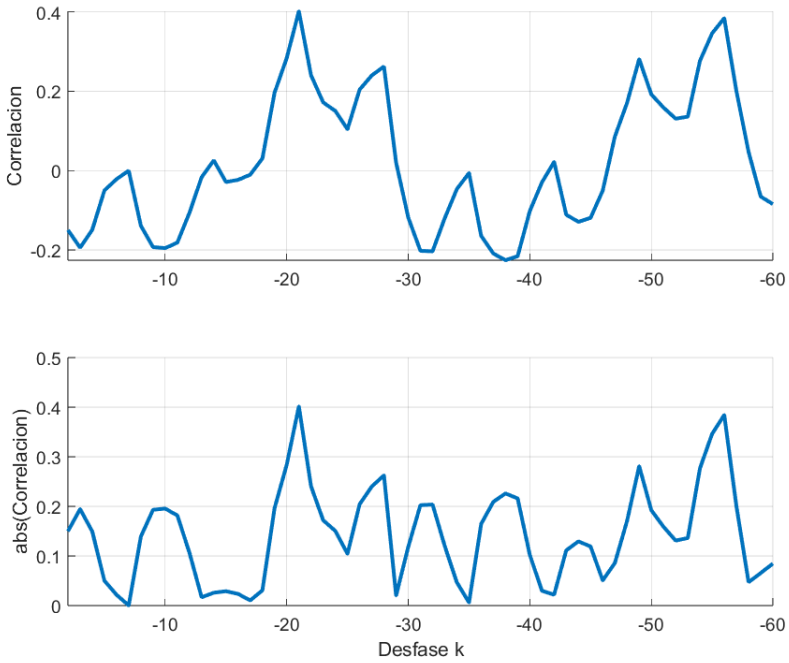


Figura 4.7 Correlaciones entre saldo(k+7) y saldo(k+d), con d = [-2,-60] utilizando B1@d.

En la Figura 4.7 se observan varios picos altos de correlación en torno a los desfases -20 y -55, pero no se aprecia ningún patrón ni tendencia evidente.

Además, es importante destacar que los picos de correlación están en torno a 0.4, un valor significativamente más bajo que el que observamos en la serie temporal de la energía eléctrica. Esto indica una tendencia lineal más baja en los datos de las transacciones bancarias.

A continuación, se representan las correlaciones de los desfases desde el -2 hasta el -60 utilizando la entrada B2@d.

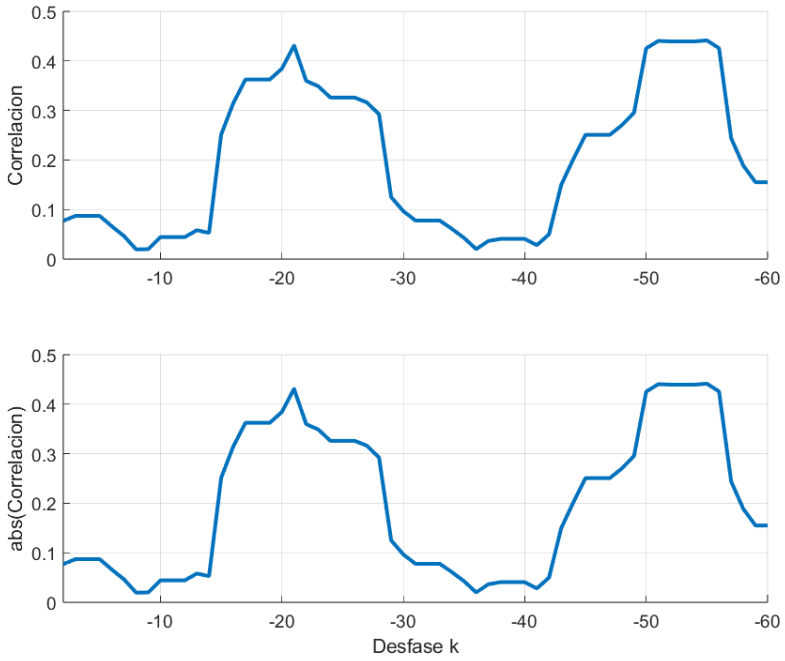


Figura 4.8 Correlaciones entre saldo(k+7) y saldo(k+d), con d = [-2,-60] utilizando B2@d.

En la gráfica de la correlación utilizando la entrada B2@d, representada en la Figura 4.8, se observa una mayor periodicidad en los valores de correlación de los desfases. Desde el desfase -15 hasta el -30, se encuentra un rango de mayor correlación lineal, y esta tendencia parece repetirse cada 30 días. Se recogen los 15 desfases con mayor correlación para ambas entradas en la Tabla 4.1.

Tabla 4.1 Tabla de Correlaciones de B1@d y B2@d.

B1@d	
d	**abs(Corr. k+7)**
-21	0.4016
-56	0.3843
-55	0.3460
-20	0.2839
-49	0.2810
-54	0.2766
-28	0.2625
-22	0.2404
-27	0.2400
-38	0.2260
-39	0.2159
-37	0.2091
-26	0.2043
-32	0.2036
-31	0.2023

B2@d	
d	**abs(Corr. k+7)**
-55	0.4414
-51	0.4404
-52	0.4395
-53	0.4395
-54	0.4395
-21	0.4310
-56	0.4257
-50	0.4253
-20	0.3843
-17	0.3624
-18	0.3624
-19	0.3624
-22	0.3597
-23	0.3488
-24	0.3259

Se puede observar en la Tabla 4.1 que al utilizar la entrada B2@d, se obtiene un poco de mayor correlación lineal entre los desfases y la salida. Sin embargo, es importante destacar que esta mayor correlación no necesariamente se traducirá en mejores resultados al utilizar dicha entrada. En el caso de una serie tan compleja como esta, las correlaciones lineales pueden no proporcionar una comprensión completa, ya que es probable que la serie esté influenciada por numerosas relaciones no lineales que no son detectadas por la correlación de Pearson.

4.5.1 Análisis de correlación de Pearon de las Entradas Exógenas

También se calcula la correlación lineal con la salida utilizando las entradas exógenas generadas a partir de la propia serie temporal de transacciones (B) con el mismo fin de determinar que desfases temporales se han de considerar para su inclusión en los modelos. Se recuerda que estas entradas son:

- **B3@d:** el número de transacciones del día k+d,
- **B4@d:** la media de la cantidad dispensada en los últimos 5 días anteriores a k+d.
- **B5@d y B6@d:** los valores mínimos y máximos de los últimos 5 días anteriores a k+d.

En la Tabla 4.2 se recogen los 15 desfases de mayor correlación obtenidos para cada una de estas entradas.

Tabla 4.2 Tabla de Correlaciones de B3@d, B4@d, B5@d y B6@d.

B3@d		B4@d		B5@d		B6@d	
d	**abs(Corr. k+7)**	**d**	**abs(Corr. k+7)**	**d**	**abs(Corr. k+7)**	**d**	**abs(Corr. k+7)**
-21	0.3833	-50	0.3778	-49	0.2797	-50	0.3190
-41	0.3713	-21	0.3700	-21	0.2703	-49	0.2959
-56	0.3709	-20	0.3682	-50	0.2640	-51	0.2946
-13	0.3212	-49	0.3629	-20	0.2401	-53	0.2724
-6	0.3157	-51	0.3584	-22	0.2385	-20	0.2720
-34	0.3061	-19	0.3329	-51	0.2259	-21	0.2716
-40	0.2942	-52	0.3283	-19	0.2255	-52	0.2713
-24	0.2649	-22	0.3173	-48	0.2097	-54	0.2422
-12	0.2635	-18	0.3070	-34	0.2049	-48	0.2402
-53	0.2598	-53	0.3056	-33	0.2038	-22	0.2255
-49	0.2592	-48	0.3044	-3	0.2032	-19	0.2219
-28	0.2525	-54	0.2879	-52	0.2031	-55	0.2069
-52	0.2473	-17	0.2736	-18	0.2028	-18	0.1961
-5	0.2452	-23	0.2547	-32	0.2020	-47	0.1783
-33	0.2415	-47	0.2480	-35	0.1981	-23	0.1766

Se observa que las entradas con las correlaciones más altas son B3@d y B4@d, con una diferencia de aproximadamente diez décimas en comparación con las correlaciones de B5@d y B6@d. Sin embargo, en general, los niveles de correlación lineal para esta serie son bastante bajos, especialmente al compararlos con la serie de la demanda eléctrica.

4.6 Modelado ARX

Como se hizo con la serie temporal del capítulo anterior, se procede con el método de inclusión progresiva de entradas, utilizando inicialmente los 15 valores de desfase temporal que demostraron la mayor correlación lineal al emplear la entrada B1@d como cálculo autorregresivo. Se recuerda que esta entrada, en caso de no haber transacciones registradas en el día k+d, se tomará el día anterior más cercano con transacciones. Posteriormente, se realizará lo mismo utilizando la entrada B2@d. El objetivo es determinar, para ambos tipos de entrada, a partir de qué iteración se obtiene el modelo con el menor error.

Al igual que en las pruebas realizadas para la serie anterior, los datos se dividen en un subconjunto de entrenamiento, que abarca el 80 % de los datos, y un conjunto de prueba con el 20 % restante. Los parámetros FactP y Gamma de los modelos ARX, al igual que en la serie del consumo eléctrico, se mantendrán en 10 y 1, respectivamente.

4.6.1 Prueba con entradas autorregresivas de tipo B1@d

Se representan gráficamente los Errores Cuadráticos Medios de los modelos resultantes de cada iteración, lo que permite apreciar de manera más clara la progresión a lo largo de las iteraciones.

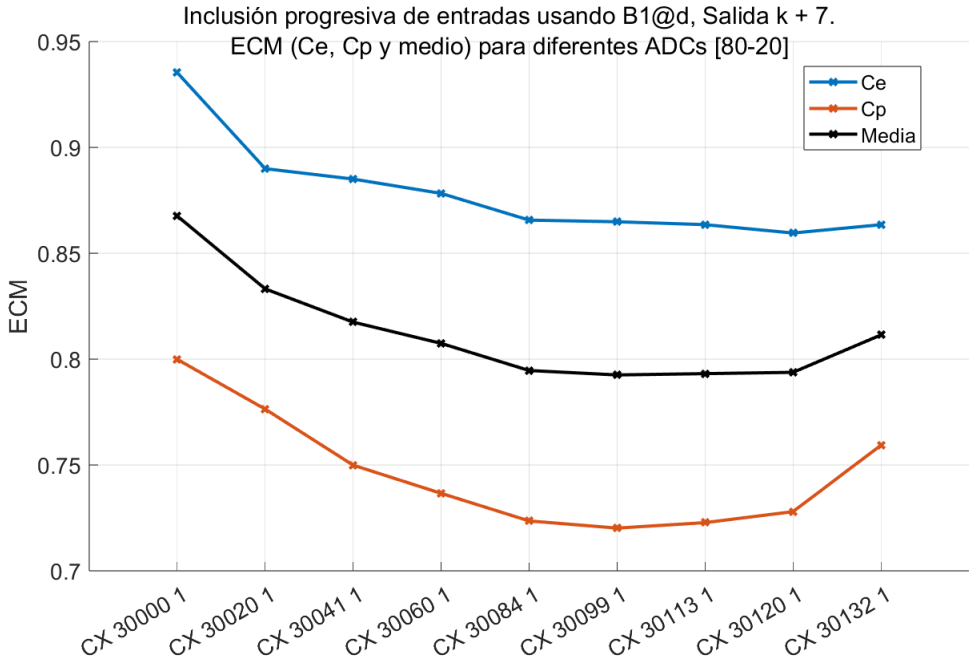


Figura 4.9 Inclusión progresiva de entradas autorregresivas de tipo B1@d.

En la representación gráfica de la Figura 4.9, se observa que el Error Cuadrático Medio prácticamente se estabiliza en la quinta iteración y, de hecho, comienza a incrementar su valor a partir de la octava. El modelo que ofrece el mejor rendimiento es el **60099**, creado mediante el **CX 30099**, que incluye los desfases: -21, -56, -49, -28, -38 y -27 como entradas autorregresivas de tipo B1@d. Dicho modelo posee un ECM medio de **0.7925**, lo cual es un valor bastante superior a los que se obtuvieron para la serie del consumo eléctrico.

Además, es evidente una notoria disparidad entre los valores del ECM del conjunto de entrenamiento (CE), que se sitúan en un valor torno a 0.88, y el ECM del conjunto de prueba (CP) de los modelos, que se aproxima a 0.73. También, cabe destacar que resulta paradójico que el error en el conjunto de prueba sea inferior al del conjunto de entrenamiento, lo cual es inusual y sugiere una falta de adecuación del algoritmo ARX a esta serie temporal. Esta discrepancia de más de 10 décimas entre ambos errores indica que el algoritmo ARX no está logrando un ajuste óptimo a esta serie de datos temporal.

Esto puede ser debido a que al ser el enfoque ARX un enfoque inherentemente lineal, se limita su capacidad para capturar las complejas relaciones no lineales presentes en la serie temporal, lo que se refleja en los errores elevados que se observan en estos modelos.

Se representa en la Figura 4.10 las salidas reales frente a las salidas predichas por el modelo **600999**, y se observa claramente cómo las predicciones, especialmente para los valores del conjunto de entrenamiento, se encuentran considerablemente alejadas de los valores reales.

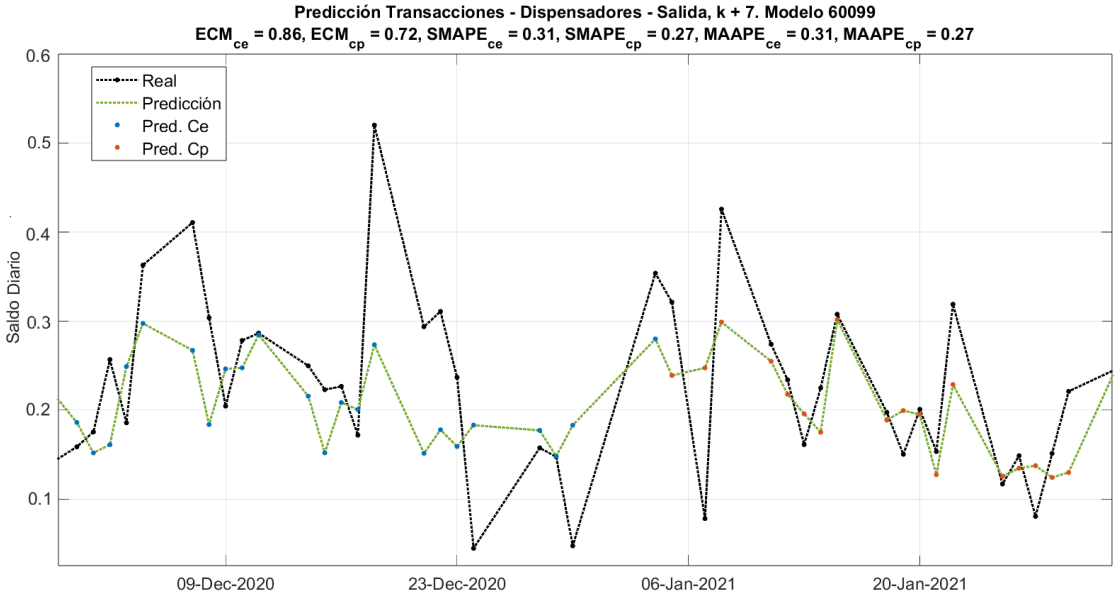


Figura 4.10 Salidas reales frente a las predichas por el modelo ARX 60099.

Se puede concluir que el algoritmo ARX no ha logrado adaptarse eficazmente a esta serie temporal. En consecuencia, se enfocará en realizar pruebas exclusivamente con el segundo tipo de entradas autorregresiva para determinar cuál de ellas funciona mejor en este tipo de modelos, sin incorporar las entradas exógenas definidas anteriormente.

4.6.2 Inclusión progresiva de entradas autorregresivas de tipo B2@d

Se realiza la misma prueba con los mismos parámetros y porcentajes para los subconjuntos de prueba y entrenamiento, para la entrada B2@d, la cual, si se encuentra con que no hay transacciones registradas en el día k+d, tomará los datos del mismo tipo de día anterior más cercano con transacciones.

Se recuerda que los tipos de días que se han definido son:

1. Lunes, martes, miércoles y jueves
2. Viernes

Se representa el ECM de los modelos resultantes de la inclusión progresiva de entradas en la Figura 4.11.

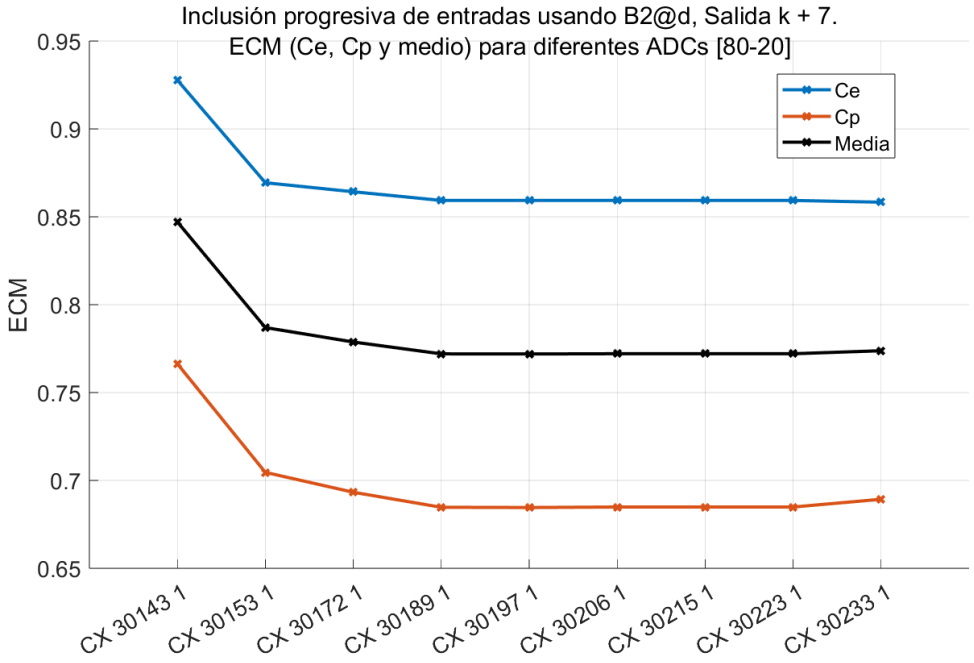


Figura 4.11 Inclusión progresiva de entradas autorregresivas de tipo B2@d.

Se observa un comportamiento bastante similar al de la anterior prueba. A partir de la cuarta iteración por más que se agreguen entradas a los modelos, el ECM no disminuye. El ECM Medio alcanza un valor mínimo de **0.772** para el modelo **60197**, creado con el **CX 30197** que incluye las entradas: -21, -55, -50, -22, -20.

A pesar de que el ECM medio sea unas 2 décimas inferior al mejor modelo que utilizaba la entrada B1@d, no se puede concluir aún que esta entrada produce, en general, mejores resultados.

4.6.3 Comparativa entre los dos tipos de entradas autorregresivas

En esta sección, se representan los cuatro tipos de errores de los modelos más destacados de las dos pruebas anteriores para obtener una comprensión más precisa de cuál de las entradas autorregresivas, B1@d o B2@d, ha funcionado mejor para los modelos ARX. En las gráficas, se utilizá el color amarillo para representar los modelos creados con la entrada B1@d y el color violeta para los modelos que utilizan B2@d. Estos análisis permiten evaluar de manera más completa y detallada el desempeño de ambos tipos de entrada.

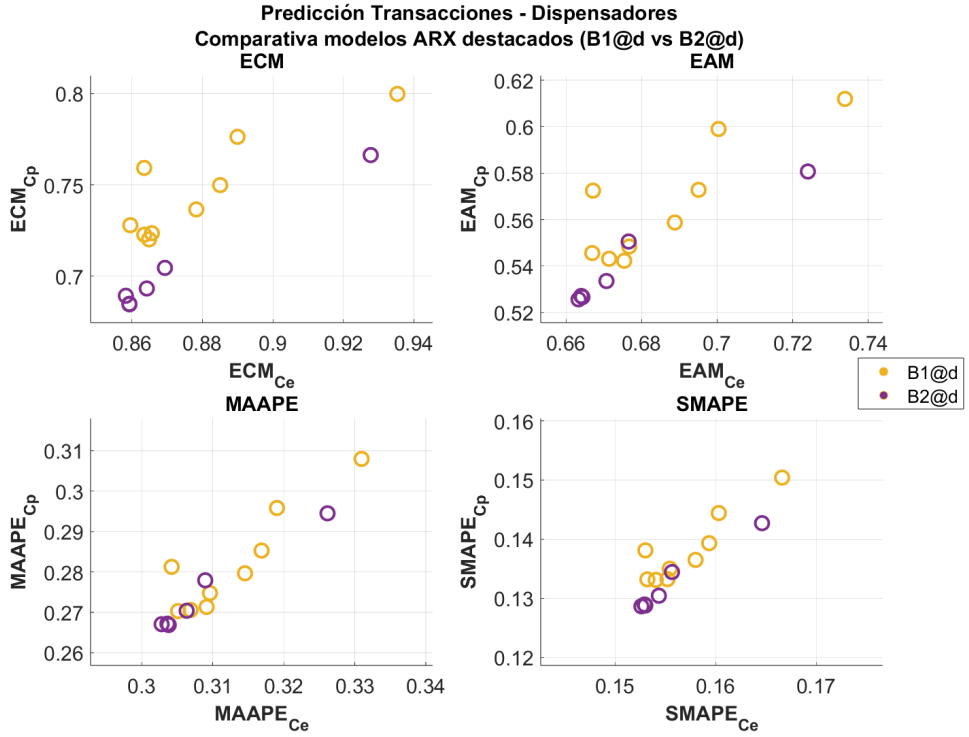


Figura 4.12 Comparativa entre modelos ARX usando diferentes tipos de entradas autorregresivas.

En la Figura 4.12 se puede observar que los errores de los dos tipos de entradas autorregresivas son, en su mayoría, muy similares. Sin embargo, es cierto que parece que los modelos representados en color violeta (B2@d) tienden más hacia la esquina inferior izquierda en las cuatro subgráficas.

A pesar de que los resultados para ambas entradas no han sido muy favorables, se puede afirmar afirmar que la entrada autorregresiva B2@d ha funcionado mejor al usar modelos ARX. Esto es coherente con lo que vimos en la Sección 4.5, donde se encontraron valores de correlación lineal más altos para esta entrada.

Sin embargo, es importante señalar que la diferencia de rendimiento entre las entradas autorregresivas B1@d y B2@d en los modelos ARX no necesariamente se traducirá en el mismo patrón para los modelos RBF. Dado que está claro que el enfoque lineal de los modelos ARX no ha logrado adaptarse a la complejidad de la serie temporal, la verdadera prueba llegará en la siguiente sección con los modelos RBF, donde además se incorporarán las entradas exógenas definidas en la introducción de este capítulo.

4.7 Modelado RBF

Se inician las pruebas del modelado RBF con la inclusión progresiva de las entradas autorregresivas de tipo B1@d. El mejor modelo obtenido en esta prueba se utilizará como punto de partida para incorporar posteriormente las entradas exógenas, con la esperanza de lograr una reducción significativa en los errores de predicción.

Por último, se repetirá todo el proceso para la entrada autorregresiva B2@d. De esta manera, se podrá comparar el rendimiento de ambas entradas y determinar cuál de ellas funciona mejor cuando se incorporan las entradas exógenas, y cuál de las dos entradas proporciona un mejor ajuste a la serie de datos y, por lo tanto, es más adecuada para las predicciones utilizando modelos RBF.

4.7.1 Prueba con entradas autorregresivas de tipo B1@d

Los parámetros iniciales de entrenamiento que se han empleado para esta prueba se presentan en la Tabla 4.3.

Tabla 4.3 Parámetros de entrenamiento iniciales.

Kappa	Nº de Neuronas	Alpha	Nº de Pasadas	Gamma
1 2 3	50 75 100	0.01	40	1

Una vez concluida la fase de inclusión progresiva de las entradas autorregresivas y se obtenga el conjunto de entradas que resulte en los errores más bajos, se seguirá el mismo procedimiento que se aplicó en la serie de la demanda de energía eléctrica. Es decir, se realizará un barrido exhaustivo de los parámetros de entrenamiento del modelo RBF con el objetivo de encontrar la combinación óptima que minimice los errores de las predicciones.

Los modelos resultantes de la inclusión progresiva de entradas aparecen recopilados en la Tabla 4.4, y en la Figura 4.13 se representa gráficamente el ECM los mismos.

Tabla 4.4 Inclusión progresiva de entradas de tipo B1@d.

ID RBF	ID CX	Entradas	$SMAPE_{CE}$	$SMAPE_{CP}$	ECM_{CE}	ECM_{CP}
100002	40002	-55	0.3650	0.3419	1.0155	0.9308
100163	40015	-55, -21	0.3010	0.2989	0.8476	0.8223
100333	40036	-55, -21, -38	0.2971	0.3038	0.8281	0.8115
100485	40053	-55, -21, -38, -31	0.2861	0.2825	0.8027	0.8077
100566	40057	-55, -21, -38, -31, -54	0.2751	0.2894	0.7703	0.8126
100634	40074	-55, -21, -38, -31, -54, -32	0.2880	0.2908	0.8028	0.8124
100696	40078	-55, -21, -38, -31, -54, -32, -28	0.2961	0.2905	0.8212	0.7962
100770	40090	-55, -21, -38, -31, -54, -32, -28, -37	0.2901	0.2585	0.8122	0.7589
100837	40094	-55, -21, -38, -31, -54, -32, -28, -37, -49	0.2945	0.2697	0.8137	0.7564
100923	40101	-55, -21, -38, -31, -54, -32, -28, -37, -49, -22	0.2882	0.2767	0.7847	0.7765
100955	40105	-55, -21, -38, -31, -54, -32, -28, -37, -49, -22, -56	0.2915	0.2878	0.8040	0.7598
100998	40110	-55, -21, -38, -31, -54, -32, -28, -37, -49, -22, -56, -20	0.2914	0.2858	0.8012	0.7591
101052	**40116**	**-55, -21, -38, -31, -54, -32, -28, -37, -49, -22, -56, -20, -26**	**0.2841**	**0.2963**	**0.7786**	**0.7826**
101063	40117	-55, -21, -38, -31, -54, -32, -28, -37, -49, -22, -56, -20, -26, -27	0.2904	0.2920	0.7960	0.7825
101073	40119	-55, -21, -38, -31, -54, -32, -28, -37, -49, -22, -56, -20, -26, -27, -39	0.2912	0.2963	0.7984	0.8019

A diferencia de lo que ocurría para los modelos ARX, en la Figura 4.13 se observa una mayor

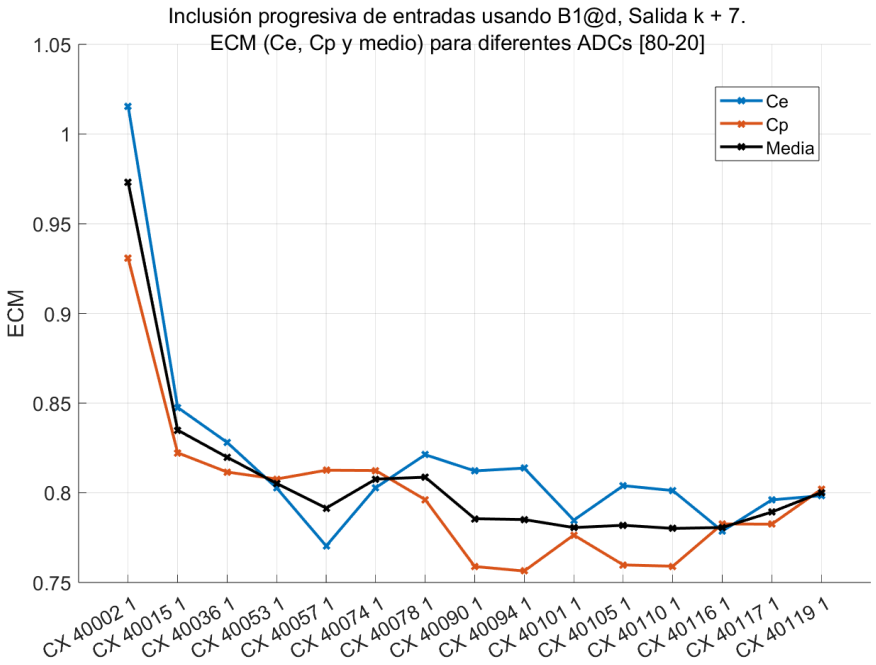


Figura 4.13 Inclusión progresiva de entradas autorregresivas de tipo B1@d.

disminución del ECM a medida que se añaden entradas, además de valores más similares entre los conjuntos de prueba y entrenamiento.

El menor ECM medio es 0.78 y se alcanza en la decimotercera iteración en el modelo **101052**, cuyo conjunto de entradas es el **CX 40116** que incluye los desfases:-55, -21, -38, -31, -54, -32, -28, -37, -49, -22, -56, -20 y -26.

Se realiza una exploración de parámetros para este conjunto de entradas. Inicialmente, se han variado los valores de kappa y el número de neuronas hasta alcanzar los valores óptimos de 2.5 y 101 neuronas, respectivamente, como se muestra en la Figura 4.14.

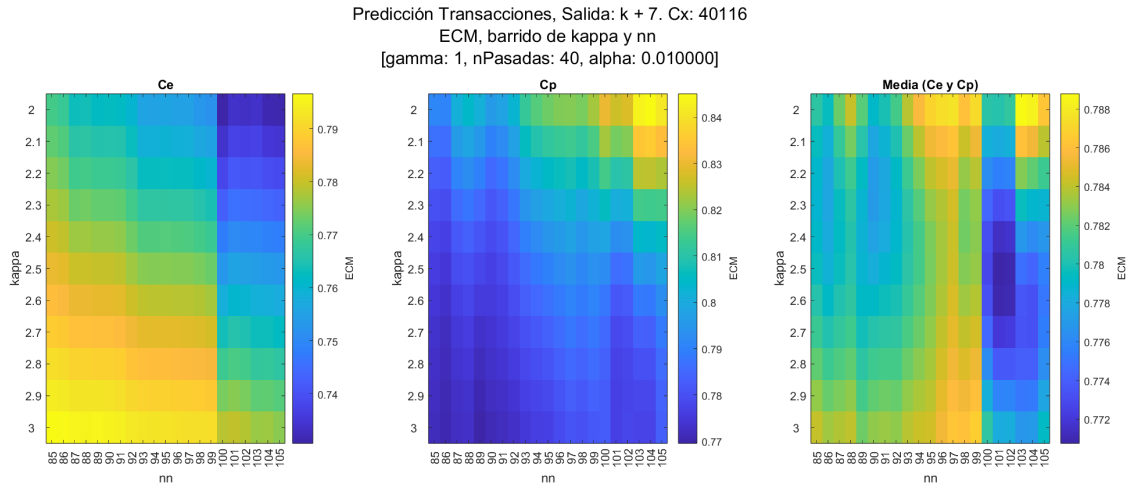


Figura 4.14 Barrido de kappa y el nº de neuronas para el CX 40116.

Y posteriormente se han variado los valores de alpha y el número de pasadas hasta llegar a 100 pasadas y alpha=0.003 como se puede apreciar en la Figura 4.15.

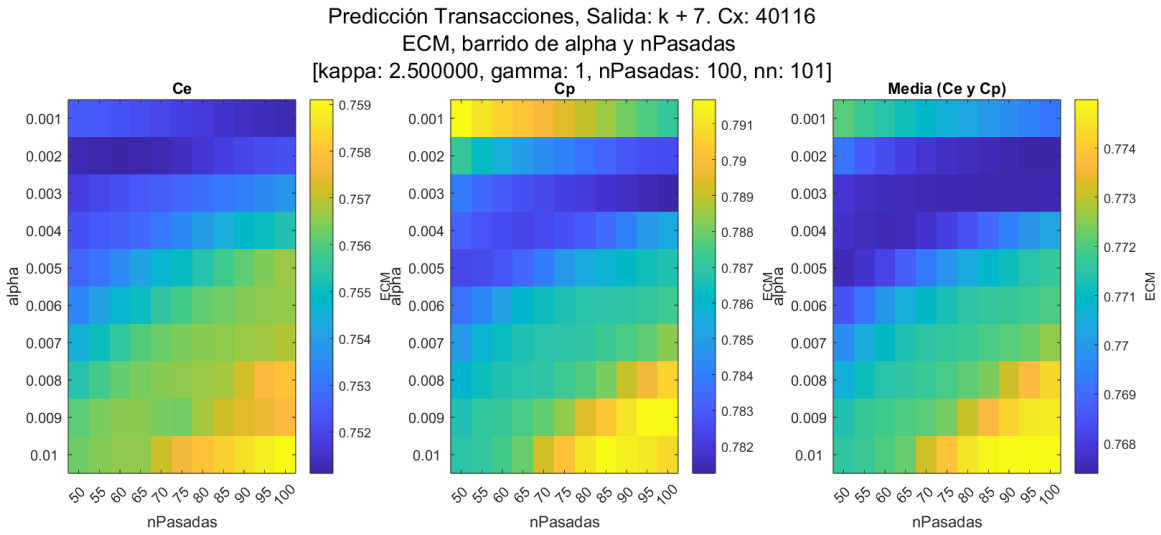


Figura 4.15 Barrido de alpha y el nº de pasadas para el CX 40116.

Una vez elegido el conjunto de desfases de entrada (Tabla 4.5) y la combinación óptima de parámetros RBF (Tabla 4.6) se genera el modelo **101767** cuyos errores aparecen en la Tabla 4.7.

Tabla 4.5 Información del CX 40116.

CX	Tipo de entrada	Desfases
40116	B1@d	[-55, -21, -38, -31, -54, -32, -28, -37, -49, -22, -56, -20, -26]

Tabla 4.6 Parámetros de entrenamiento del modelo RBF 101767.

Kappa	Nº de Neuronas	Alpha	Nº de Pasadas	Gamma
2.5	101	0.0030	100	1

Tabla 4.7 Información del modelo RBF 101767.

ID RBF	ID CX	$SMAPE_{Medio}$	$SMAPE_{CE}$	$SMAPE_{CP}$	ECM_{Medio}	ECM_{CE}	ECM_{CP}
101767	40116	0.2829	0.2760	0.2897	0.7675	0.7537	0.7812

Este modelo **101767** servirá como punto de partida para incorporar progresivamente cada uno de los tipos de entradas exógenas definidos para esta serie. El objetivo es reducir los errores y mejorar la calidad de las predicciones de manera significativa.

Inclusión de entradas exógenas en el modelo

1. **Inclusión de entradas exógenas que aportan información adicional sobre la propia serie de transacciones (B)**

   - **B3@d:** Número de transacciones en los dispensadores de la sucursal el día k+d.

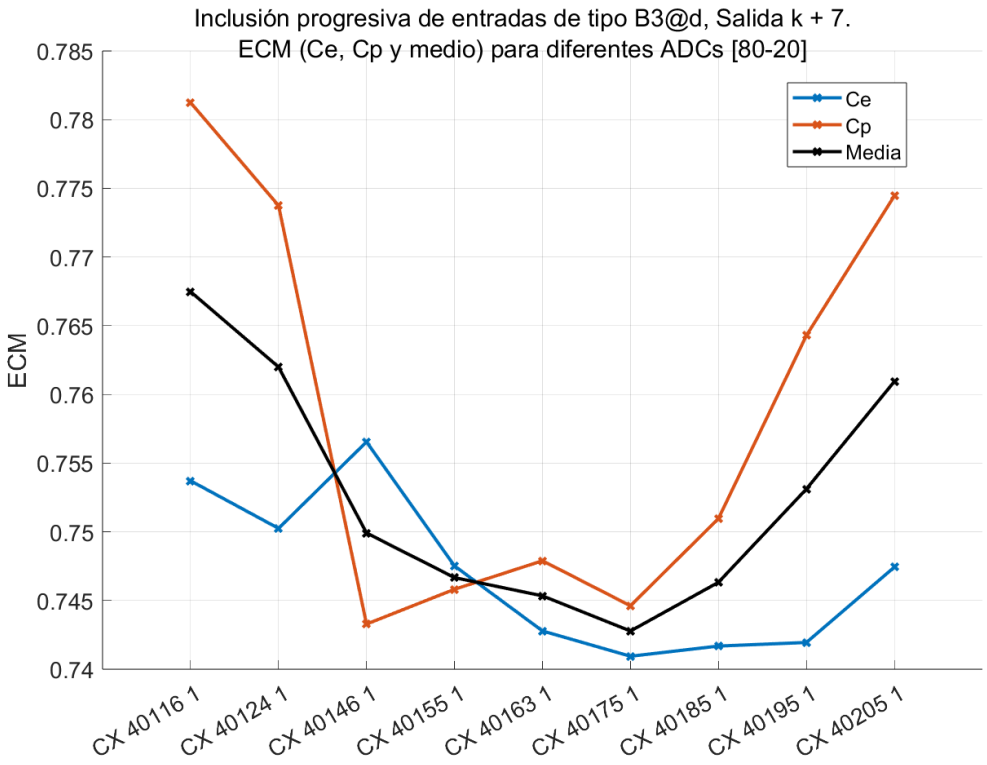


Figura 4.16 Inclusión progresiva de entradas exógenas de tipo B3@d.

En la Figura 4.16 se parte del modelo inicial 101767 creado con el conjunto de desfases de entradas CX 10116. Luego, se procede a añadir de uno en uno los desfases de la entrada B3@d. Se puede observar en el gráfico que esta entrada ha funcionado de manera efectiva para el modelo, ya que se ha logrado una reducción considerable del Error Cuadrático Medio desde un valor inicial de 0.7675 hasta alcanzar un valor de 0.7428 en la quinta iteración de esta prueba (CX 40175).

El CX **40175** incluye las siguientes entradas:

- **B1@d** con d = -55, -21, -38, -31, -54, -32, -28, -37, -49, -22, -56, -20, -26.
- **B3@d** con d = -6, -52, -24, -56, -41.

El conjunto de entradas resultante de cada prueba será la base a partir de la cual se iniciará la siguiente prueba de inclusión de entradas exógenas.

- **B4@d:** Media de la cantidad diaria dispensada en los dispensadores de la sucursal los últimos 5 días con transacciones previos al día k+d.

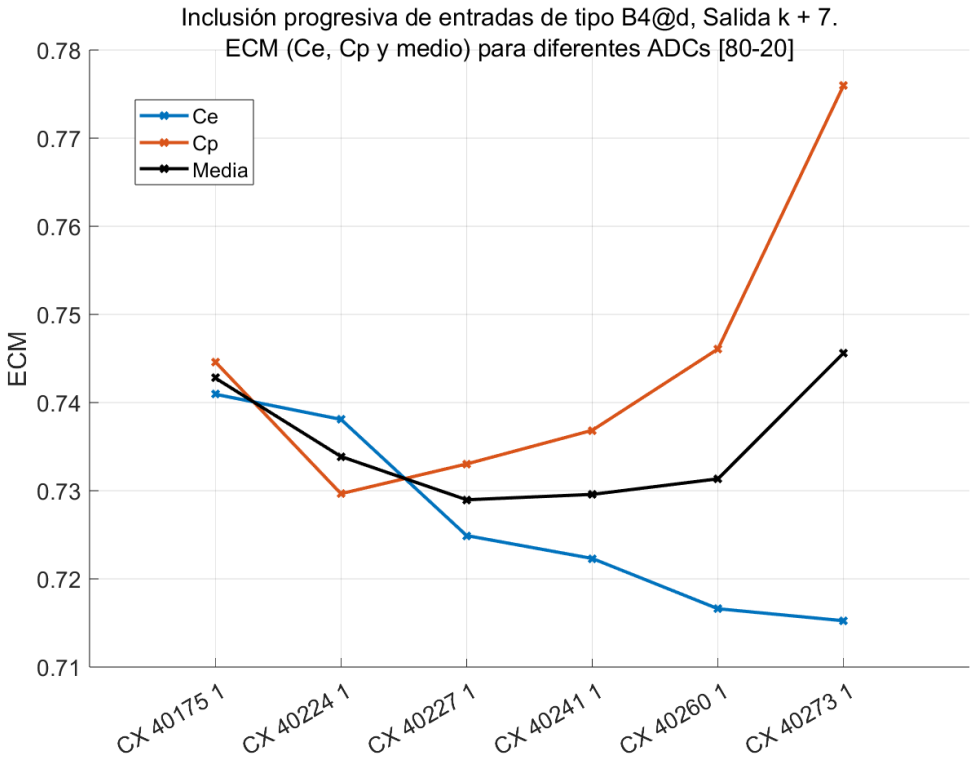


Figura 4.17 Inclusión progresiva de entradas exógenas de tipo B4@d.

Se observa que la inclusión de la entrada que representa la media de las cantidades de los últimos 5 días ha resultado eficaz para el modelo, ya que ha conducido a una disminución del Error Cuadrático Medio promedio a 0.729 en la segunda iteración (CX 40227).

El CX **40227** incluye las siguientes entradas:

- **B1@d** con d = -55, -21, -38, -31, -54, -32, -28, -37, -49, -22, -56, -20, -26.
- **B3@d** con d = -6, -52, -24, -56, -41.
- **B4@d** con d = -17, -50.

- **B5@d:** Mínima cantidad diaria dispensada en los dispensadores de la sucursal los últimos 5 días con transacciones previos al día k+d.

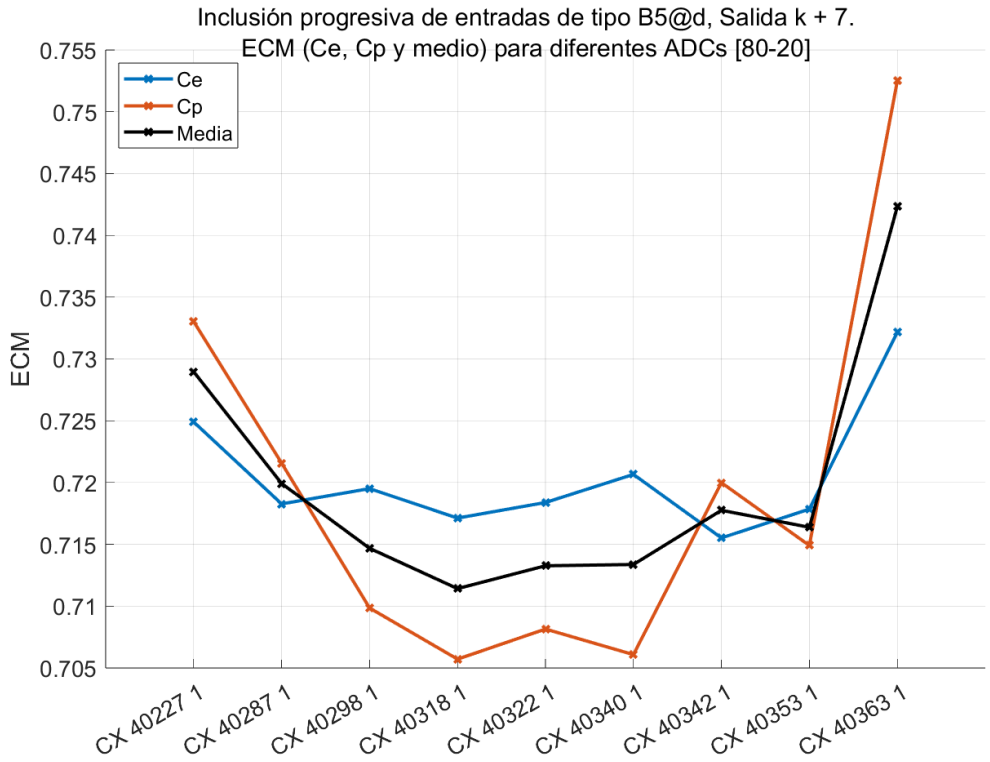


Figura 4.18 Inclusión progresiva de entradas exógenas de tipo B5@d.

Una vez más, la entrada exógena B5@d parece funcionar al mejorar las predicciones al reducir el Error Cuadrático Medio promedio a 0.7114 en la tercera iteración (CX 40318).

El CX **40318** incluye las siguientes entradas:

- **B1@d** con d = -55, -21, -38, -31, -54, -32, -28, -37, -49, -22, -56, -20, -26.
- **B3@d** con d = -6, -52, -24, -56, -41.
- **B4@d** con d = -17, -50.
- **B5@d** con d = -3, -19, -35.

- **B6@d:** Máxima cantidad diaria dispensada en los dispensadores de la sucursal los últimos 5 días con transacciones previos al día k+d.

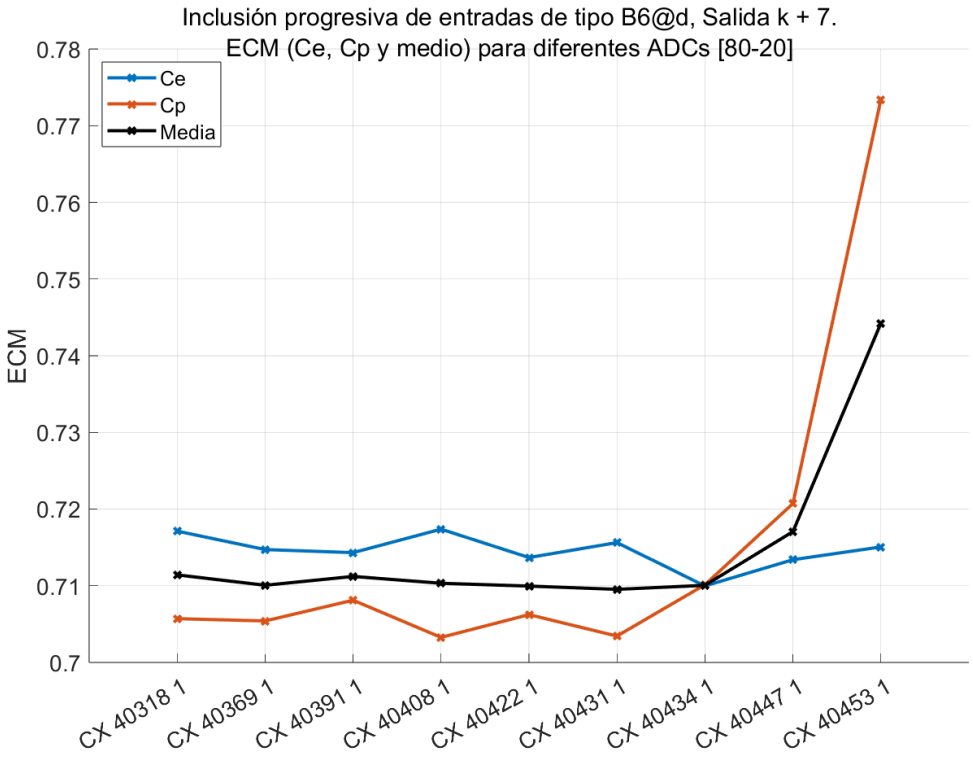


Figura 4.19 Inclusión progresiva de entradas exógenas de tipo B6@d.

Y por último, al incluir la entrada B6@d se comprueba que el ECM Medio se mantiene practicamente constante e incluso se dispara al añadir varias entradas alrededor de la séptima iteración.

Pese a esto, se logra reducir ligeramente el ECM medio a 0.71 en la segunda iteración (CX 40369) al agregar un único desfase de entrada.

El CX **40369** incluye las siguientes entradas:

- **B1@d** con d = -55, -21, -38, -31, -54, -32, -28, -37, -49, -22, -56, -20, -26.
- **B3@d** con d = -6, -52, -24, -56, -41.
- **B4@d** con d = -17, -50.
- **B5@d** con d = -3, -19, -35.
- **B6@d** con d = -50.

Después de llevar a cabo esta serie de pruebas, se puede afirmar que la inclusión de las entradas exógenas basadas en la propia serie temporal de transacciones ha resultado altamente efectiva en la mejora de los modelos. Los resultados demuestran que han logrado reducir significativamente el Error Cuadrático Medio de los modelos, disminuyéndolo desde un valor inicial de 0.7675 hasta un valor final de 0.71. Esta disminución de aproximadamente seis décimas representa una mejora considerable en la capacidad de predicción de los modelos.

2. Inclusión de entradas exógenas de tipo Calendario (C) y Derivadas (D)

En última instancia, se procederá a la inclusión conjunta de las entradas exógenas basadas en el calendario y sus entradas derivadas basadas en funciones trigonométricas.

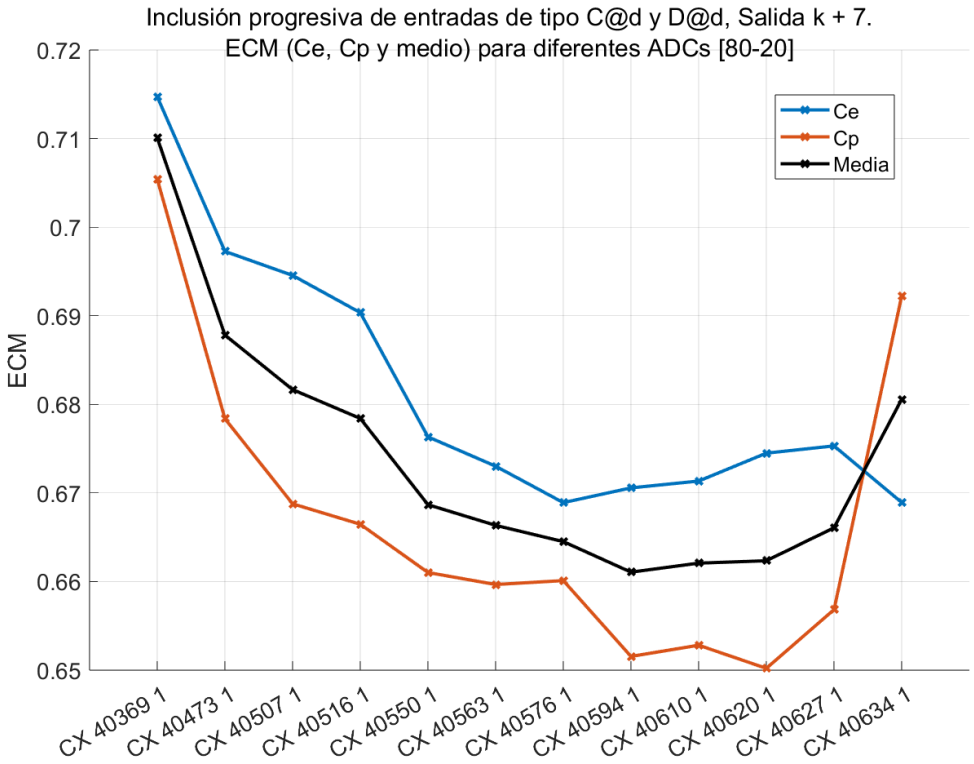


Figura 4.20 Inclusión progresiva de entradas exógenas de tipo C@d y D@D.

En la figura Figura 4.20 se aprecia una reducción significativa en el ECM de los modelos. El valor mínimo del error medio se alcanza en la séptima iteración (CX 40594) y es de 0.661. Esto representa una disminución de aproximadamente **5 décimas** en comparación con el modelo resultante de la inclusión de entradas exógenas de tipo B y una mejora de casi **11 décimas** en relación con el modelo inicial sin entradas exógenas.

Las entradas incorporadas al modelo resultante final (CX 40594) son:

- **B1@d** con d = -55, -21, -38, -31, -54, -32, -28, -37, -49, -22, -56, -20, -26.
- **B3@d** con d = -6, -52, -24, -56, -41.
- **B4@d** con d = -17, -50.
- **B5@d** con d = -3, -19, -35.
- **B6@d** con d = -50.
- **D3@7**: seno del día del mes al que pertenece la salida.
- **D8@7**: seno de la semana del mes a la que pertenece la salida.
- **C4@7**: coeficiente ponderador aplicado tipo de día al que corresponde la salida.
- **D6@7**: coseno del mes al que pertenece al salida.
- **D2@7**: coseno del día del mes al que pertenece la salida.
- **C1@7**: identificador asignado al día de la semana al que pertenece la salida.
- **C2@7**: identificador asignado al día del mes al que pertenece la salida.

Estos resultados respaldan aún más la utilidad de las entradas exógenas en la predicción de la serie temporal de transacciones bancarias.

A continuación, a modo de resumen, se recopilan en la Figura 4.21 cada uno de los modelos resultantes de las pruebas de inclusión de entradas exógenas, y en la Tabla 4.8 se indica qué tipo de entrada exógena ha sido incorporada a cada modelo, así como los errores correspondientes de cada uno de ellos.

El primer modelo de la Figura 4.21 corresponde al modelo inicial sobre el cual se han realizado los barridos de parámetros y posteriormente se han incorporado los distintos tipos de entradas exógenas.

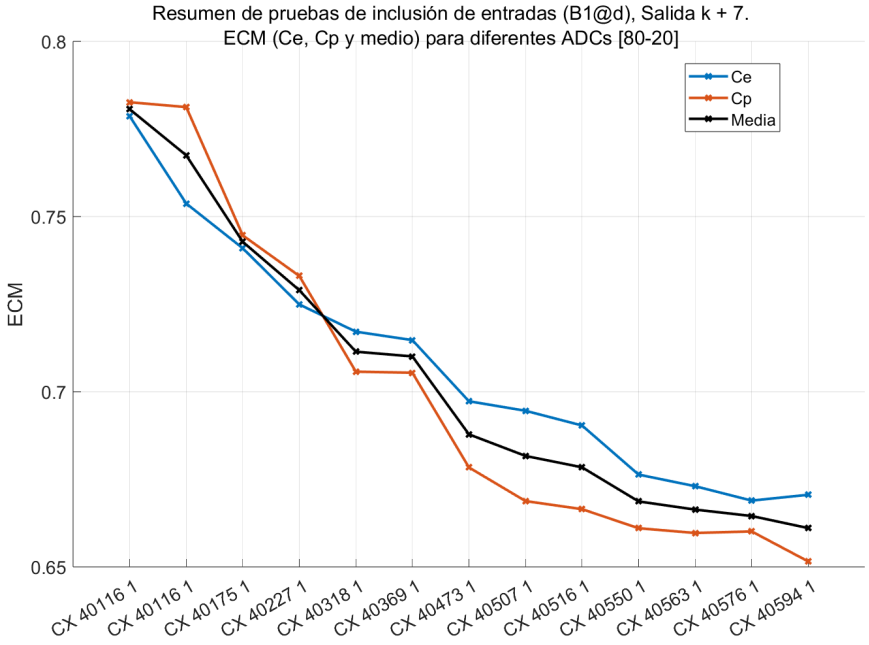


Figura 4.21 Resumen de resultados de las pruebas de inclusión de entradas exógenas.

Tabla 4.8 Resumen de resultados de las pruebas de inclusión de entradas exógenas.

ID RBF	ID CX	Entradas	$SMAPE_{\text{Medio}}$	$SMAPE_{\text{CE}}$	$SMAPE_{\text{CP}}$	ECM_{Medio}	ECM_{CE}	ECM_{CP}
101052	40116	B1@d	0.2902	0.2841	0.2963	0.7806	0.7786	0.7826
101767	40116	B1@d	0.2829	0.2760	0.2897	0.7675	0.7537	0.7812
101900	40175	B3@d	0.2721	0.2719	0.2723	0.7428	0.7409	0.7446
101952	40227	B4@d	0.2685	0.2627	0.2743	0.7290	0.7249	0.7330
102043	40318	B5@d	0.2623	0.2607	0.2639	0.7114	0.7171	0.7057
102094	40369	B6@d	0.2620	0.2597	0.2643	0.7101	0.7147	0.7054
102198	40473	D3@7	0.2562	0.2533	0.2591	0.6878	0.6973	0.6784
102232	40507	D8@7	0.2553	0.2515	0.2590	0.6817	0.6946	0.6688
102241	40516	C4@7	0.2561	0.2507	0.2615	0.6784	0.6904	0.6665
102275	40550	D6@7	0.2466	0.2475	0.2456	0.6687	0.6763	0.6610
102288	40563	D2@7	0.2462	0.2485	0.2439	0.6663	0.6730	0.6597
102301	40576	C1@7	0.2453	0.2464	0.2441	0.6645	0.6689	0.6601
102319	**40594**	**C2@7**	**0.2437**	**0.2499**	**0.2375**	**0.6611**	**0.6706**	**0.6515**

4.7.2 Comparativa entre modelos ARX y RBF usando la entrada B1@d

En la Figura 4.22 se representan los errores de los modelos RBF resultantes junto con los modelos ARX obtenidos para la misma prueba. Es evidente que los modelos RBF han producido errores más bajos, ya que se adaptaron mejor a las características no lineales de la serie temporal.

Además, al incorporar las entradas exógenas en los modelos RBF, se ha logrado reducir aún más los errores y mejorar significativamente la calidad de las predicciones.

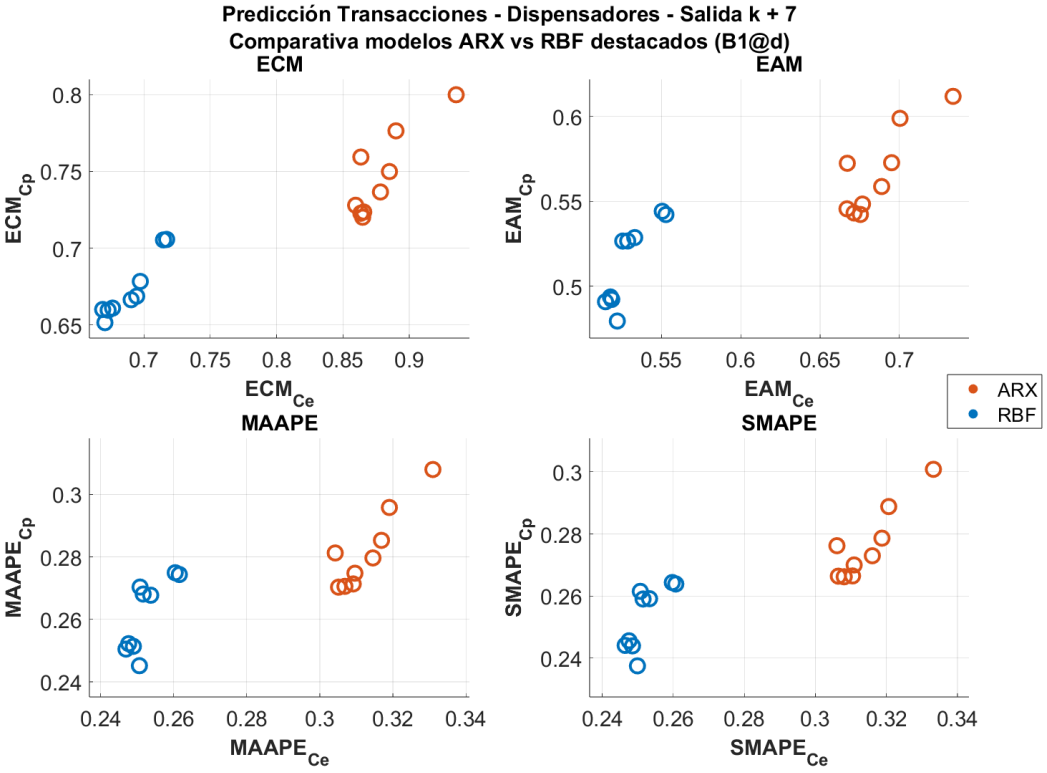


Figura 4.22 Comparativa entre modelos ARX y RBF usando la entrada B1@d.

4.7.3 Prueba con entradas autorregresivas de tipo B2@d

Para la entrada autorregresiva B2@d, cuyo cálculo consiste en tomar los datos del mismo tipo de día anterior más cercano con transacciones cuando se encuentra con un día sin datos, se realizan las mismas pruebas que para la B1@d.

Se comienza con la prueba de inclusión de entradas puramente autorregresivas utilizando la definición de B2@d. Luego, se lleva a cabo un barrido de parámetros para el conjunto de entradas resultante y, finalmente, se realizan pruebas para incorporar de manera incremental cada uno de los tipos de entradas exógenas al modelo resultante.

En la tabla Tabla 4.9, se han recopilado los modelos resultantes de cada iteración de la primera prueba de inclusión de entradas autorregresivas de tipo B2@d, y en la Figura 4.23, se han representado sus errores cuadráticos medios.

Tabla 4.9 Inclusión progresiva de entradas para B2@d.

ID RBF	ID CX	Entradas	$SMAPE_{CE}$	$SMAPE_{CP}$	ECM_{CE}	ECM_{CP}
200007	50007	-50	0.3650	0.3419	1.0155	0.9308
200191	50015	-50, -55	0.2988	0.2998	0.8194	0.7883
200291	50033	-50, -55, -21	0.2944	0.2936	0.8140	0.7751
200406	50046	-50, -55, -21, -56	0.2928	0.2831	0.8123	0.7549
200508	50054	-50, -55, -21, -56, -51	0.2887	0.2885	0.8048	0.7629
200612	50072	-50, -55, -21, -56, -51, -22	0.2923	0.2790	0.8161	0.7526
200701	50083	-50, -55, -21, -56, -51, -22, -24	0.2884	0.2824	0.8119	0.7534
200796	**50084**	**-50, -55, -21, -56, -51, -22, -24, -52**	**0.2771**	**0.2841**	**0.7888**	**0.7656**
200866	50095	-50, -55, -21, -56, -51, -22, -24, -52, -17	0.2795	0.2855	0.7886	0.7700
200922	50100	-50, -55, -21, -56, -51, -22, -24, -52, -17, -53	0.2785	0.2870	0.7845	0.7807
200970	50105	-50, -55, -21, -56, -51, -22, -24, -52, -17, -53, -54	0.2737	0.2881	0.7809	0.7790
201013	50113	-50, -55, -21, -56, -51, -22, -24, -52, -17, -53, -54, -23	0.2764	0.2903	0.7841	0.7802
201033	50115	-50, -55, -21, -56, -51, -22, -24, -52, -17, -53, -54, -23, -18	0.2864	0.2898	0.8105	0.7793
201058	50118	-50, -55, -21, -56, -51, -22, -24, -52, -17, -53, -54, -23, -18, -19	0.2876	0.2856	0.8157	0.7676
201073	50119	-50, -55, -21, -56, -51, -22, -24, -52, -17, -53, -54, -23, -18, -19, -20	0.2863	0.2850	0.8093	0.7740

El mejor modelo resultante de la prueba de inclusión de entradas autorregresivas es el **200796**, obtenido en la octava iteración. Su conjunto de entradas es el **CX 50084**, que incluye los desfases: -50, -55, -21, -56, -51, -22, -24 y -52. Su ECM Medio es de 0.777, pero tras un barrido exhaustivo de los parámetros de entrenamiento, se logró reducir el ECM en una décima.

El modelo resultante del barrido es el **201786**, y sus parámetros de entrenamiento se presentan en la Tabla 4.10, y sus errores en la Tabla 4.11.

Tabla 4.10 Parámetros de entrenamiento del modelo RBF 201786.

Kappa	Nº de Neuronas	Alpha	Nº de Pasadas	Gamma
2.4	86	0.0028	18	1

Tabla 4.11 Información del modelo RBF 201786.

ID RBF	ID CX	$SMAPE_{Medio}$	$SMAPE_{CE}$	$SMAPE_{CP}$	ECM_{Medio}	ECM_{CE}	ECM_{CP}
201786	40116	0.2831	0.2698	0.2963	0.7670	0.7571	0.7768

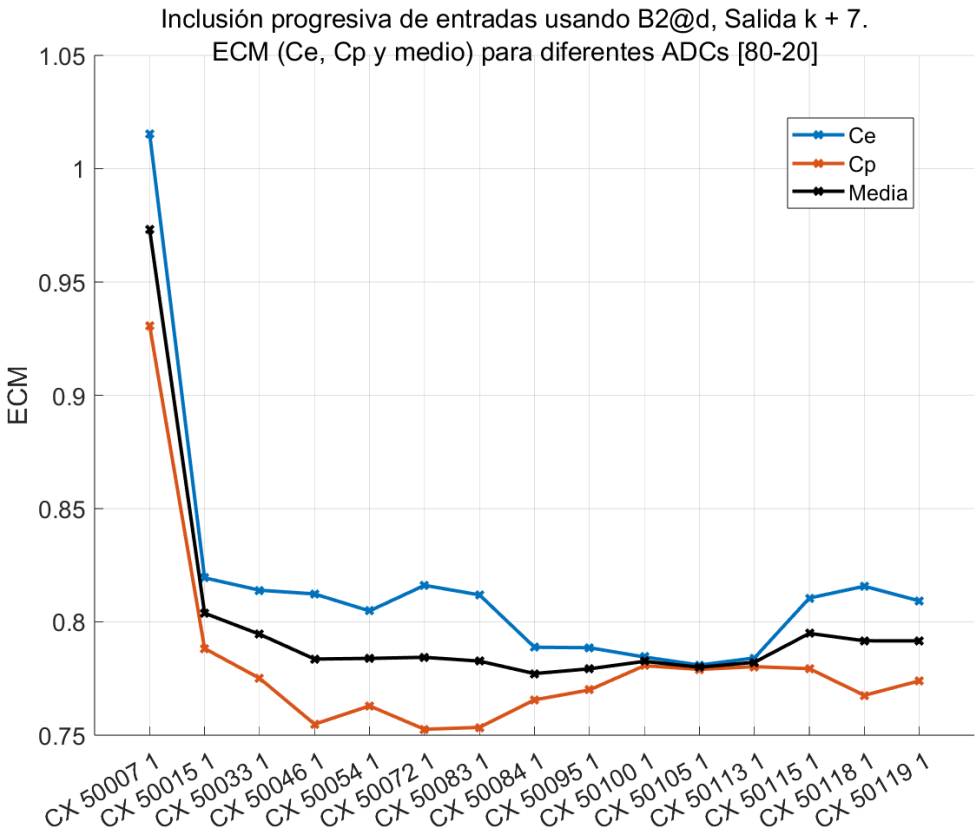


Figura 4.23 Inclusión progresiva de entradas de tipo B2@d.

Los errores del modelo 201786 son del mismo orden que los errores del modelo obtenido para la misma prueba usando la entrada B1@d. La principal diferencia se encuentra en el número de neuronas que es algo menor (86 neuronas frente a las 101 de la anterior prueba), y especialmente en el número de pasadas (18 frente a las 100 de la anterior prueba). Este modelo será el modelo inicial sobre el que se añadirán las entradas exógenas a continuación.

Inclusión de entradas exógenas

Para evitar extender esta sección, se presentarán directamente en la tabla Tabla 4.12 todos los modelos que han logrado reducir los errores en las pruebas de inclusión de entradas exógenas.

Tabla 4.12 Resumen de resultados de las pruebas de inclusión de entradas exógenas.

ID RBF	ID CX	Entradas	$SMAPE_{\text{Medio}}$	$SMAPE_{\text{CE}}$	$SMAPE_{\text{CP}}$	ECM_{Medio}	ECM_{CE}	ECM_{CP}
200796	50084	B2@d	0.2806	0.2771	0.2841	0.7772	0.7888	0.7656
201786	50084	B2@d	0.2831	0.2698	0.2963	0.7670	0.7571	0.7768
201982	50187	B3@d	0.2660	0.2654	0.2666	0.7320	0.7317	0.7323
202062	50268	B4@d	0.2548	0.2553	0.2544	0.7051	0.7129	0.6973
202195	50401	C2@7	0.2476	0.2541	0.2411	0.6901	0.7191	0.6611
202222	50428	C3@7	0.2455	0.2571	0.2339	0.6670	0.7104	0.6236
202251	50457	D5@7	0.2479	0.2578	0.2380	0.6666	0.7095	0.6238
202271	50477	D9@7	0.2454	0.2587	0.2321	0.6593	0.7069	0.6116
202275	50481	D3@7	0.2434	0.2579	0.2288	0.6540	0.7006	0.6074
202289	50495	D4@7	0.2431	0.2567	0.2295	0.6494	0.7002	0.5986
202312	50518	D8@7	0.2427	0.2554	0.2300	0.6493	0.6954	0.6032
202324	50530	D6@7	0.2400	0.2546	0.2255	0.6507	0.6931	0.6083
202328	50534	C4@7	0.2402	0.2539	0.2264	0.6491	0.6939	0.6043
202341	**50547**	**D1@7**	**0.2386**	**0.2551**	**0.2221**	**0.6449**	**0.6957**	**0.5941**

Recordamos que cada entrada presente en cada fila se va añadiendo a la entrada del modelo anterior. Por lo tanto, el modelo final resultante es el **202341** cuyo conjunto de entradas (CX 50547) incluye:

- **B2@d** con d = -50, -55, -21, -56, -51, -22, -24, -52.
- **B3@d** con d = -41, -28, -24, -34, -53, -13.
- **B4@d** con d = -50, -54, -19.
- **C2@7**: identificador asignado al día del mes al que pertenece la salida.
- **C3@7**: identificador asignado al mes al que pertenece la salida.
- **D5@7**: seno del mes a la que pertenece la salida.
- **D9@7**: coseno de la semana del mes a la que pertenece la salida.
- **D3@7**: seno del día del mes al que pertenece la salida.
- **D4@7**: coseno del día del mes al que pertenece la salida.
- **D8@7**: seno de la semana del mes a la que pertenece la salida.
- **D6@7**: coseno del mes al que pertenece al salida.
- **C4@7**: coeficiente ponderador aplicado tipo de día al que corresponde la salida.
- **D1@7**: seno del día del mes al que pertenece la salida.

Una diferencia notable en esta prueba en comparación con el uso de la entrada B1@d es que las entradas exógenas que representan los valores máximos y mínimos de las transacciones de los últimos 5 días (B5@d y B6@d) no han contribuido a mejorar las predicciones.

Se observa que el ECM Medio del modelo resultante de esta inclusión de entradas exógenas es de 0.6449, lo que representa una disminución de aproximadamente **12 décimas** en comparación con el valor del error del modelo inicial.

Este gran descenso en el ECM se visualiza mejor en la Figura 4.24 donde se representa el ECM de todos los modelos de la tabla Tabla 4.12.

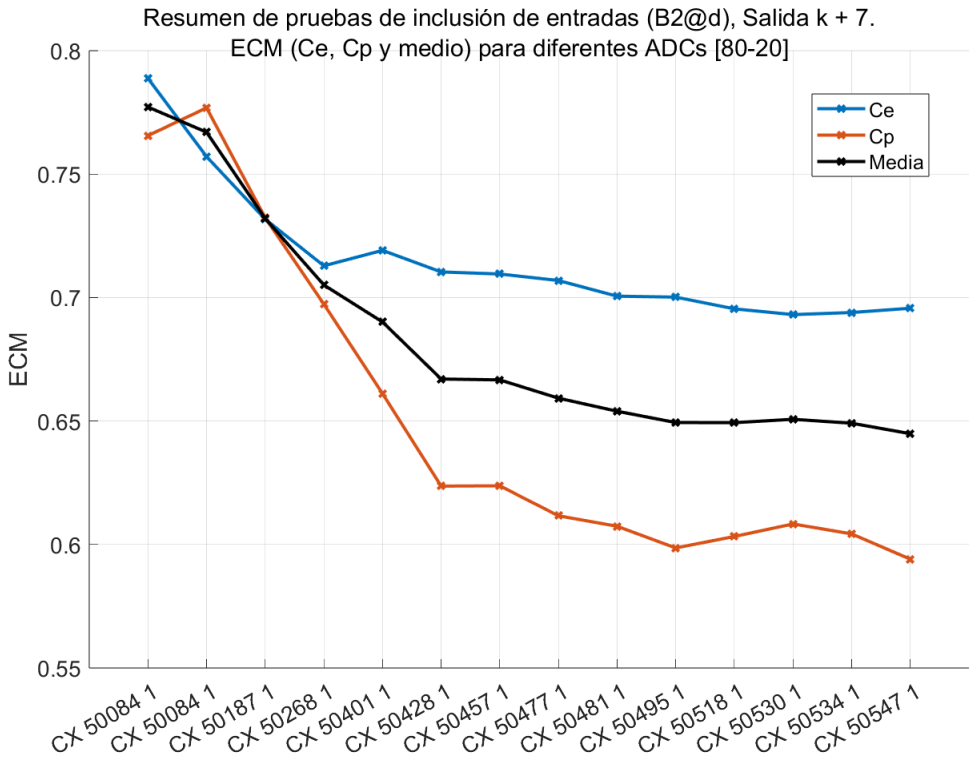


Figura 4.24 Resumen de resultados de las pruebas de inclusión de entradas exógenas.

4.7.4 Comparativa entre modelos ARX y RBF usando la entrada B2@d

En la Figura Figura 4.25, se pueden observar los modelos RBF basados en la entrada autorregresiva B2@d, representados en azul, junto con los modelos ARX que emplean la misma entrada, representados en rojo. En este caso, la discrepancia entre ambos tipos de modelos parece ser aún más marcada que cuando se utilizaba la entrada B1@d.

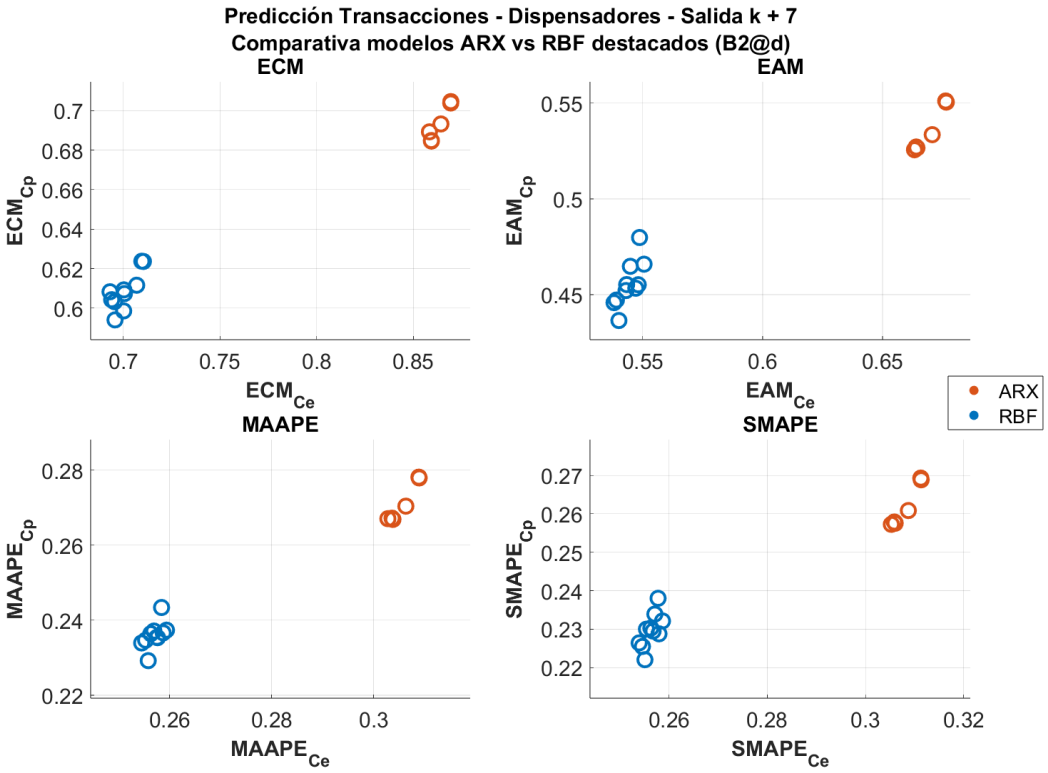


Figura 4.25 Comparativa entre modelos ARX y RBF usando la entrada B2@d.

4.8 Comparativa entre modelos RBF que utilizan B1@d y B2@d

Una vez determinado que los modelos RBF son claramente superiores a los ARX para esta serie, queda por determinar cuál de las dos entradas autorregresivas que hemos definido ha proporcionado mejores resultados.

Para llevar a cabo esta comparación, se representan los modelos resultantes de ambas pruebas en la Figura 4.26. Se utiliza el color amarillo para representar los modelos que emplean la entrada B1@d y el color violeta para los que utilizan la B2@d.

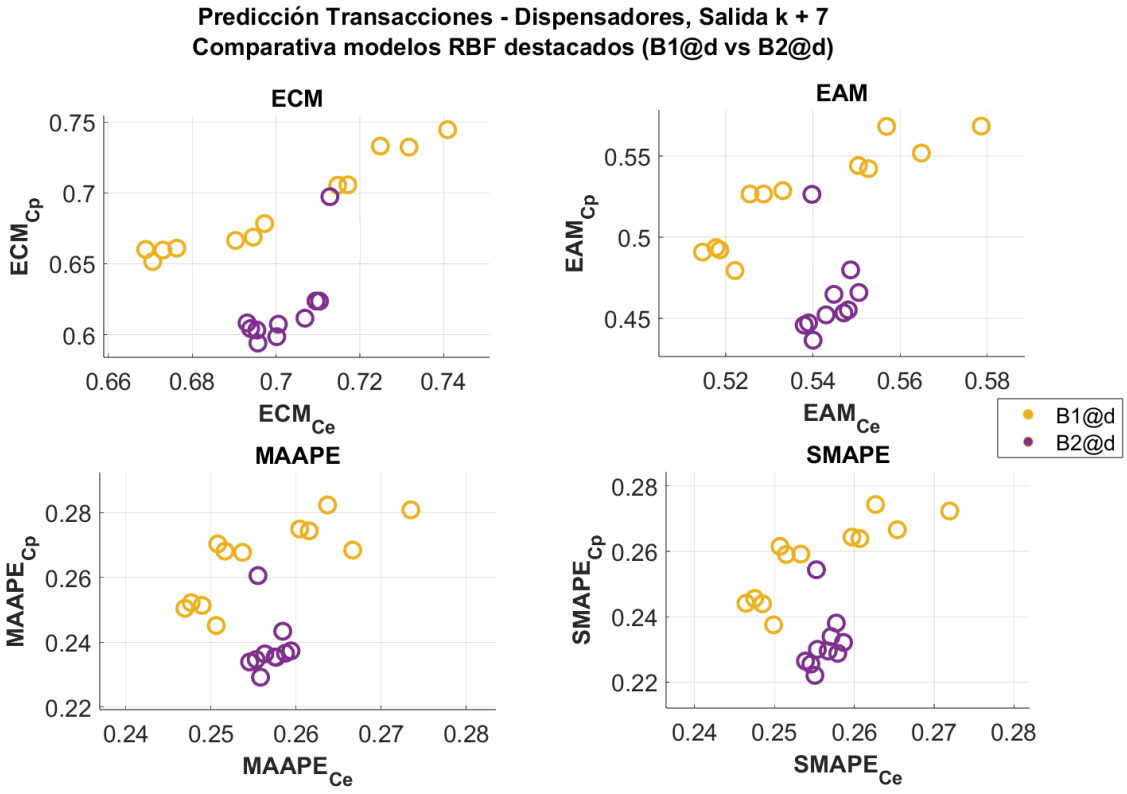


Figura 4.26 Comparativa entre modelos RBF usando entradas B1@d y B2@d.

Al analizar detenidamente los resultados, se pueden identificar varias tendencias importantes que nos permiten evaluar la eficacia de los modelos basados en las entradas B1@d y B2@d en esta serie temporal.

En primer lugar, se puede observar que los modelos basados en la entrada autorregresiva B2@d tienden a presentar errores más concentrados en comparación con aquellos que utilizan la entrada B1@d.

Además, es importante destacar que, en general, los modelos representados en amarillo parecen alcanzar un menor error en sus conjuntos de entrenamiento, ya que están situados más a la derecha en las cuatro subgráficas. Sin embargo, los modelos violetas logran tener errores más bajos en el conjunto de prueba.

Para obtener una comprensión más completa, se han representado las salidas reales en comparación con las predicciones generadas por los mejores modelos de cada tipo. En la Figura 4.27, se muestra un pequeño rango de las predicciones realizadas por el mejor modelo basado en la entrada B1@d, mientras que en la Figura 4.28, se presenta el mismo rango de predicciones generadas por el mejor modelo de las pruebas con B2@d.

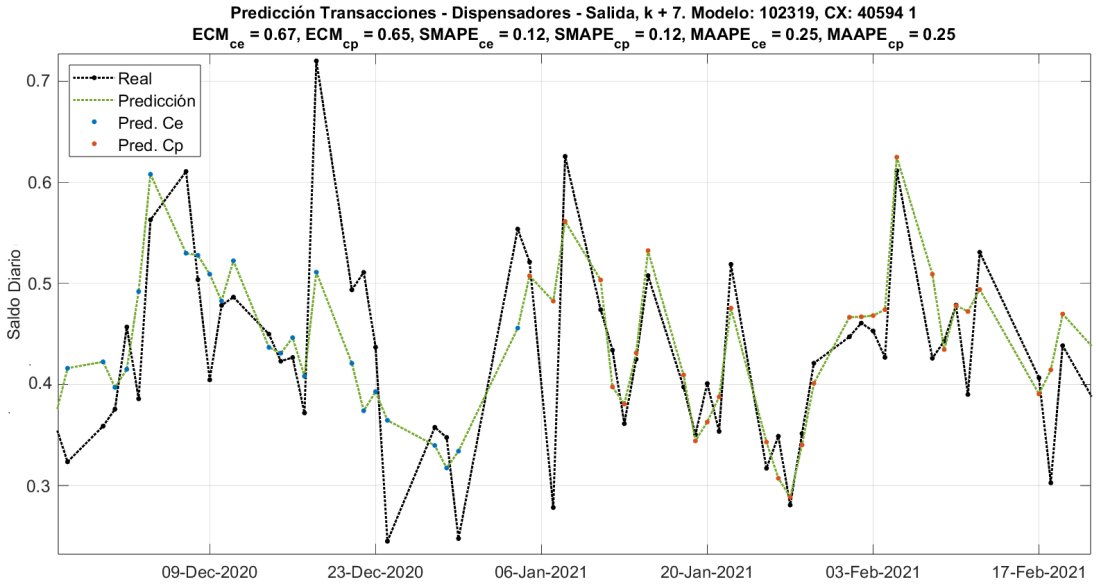


Figura 4.27 Salidas reales frente a las predichas por el modelo RBF 102319.

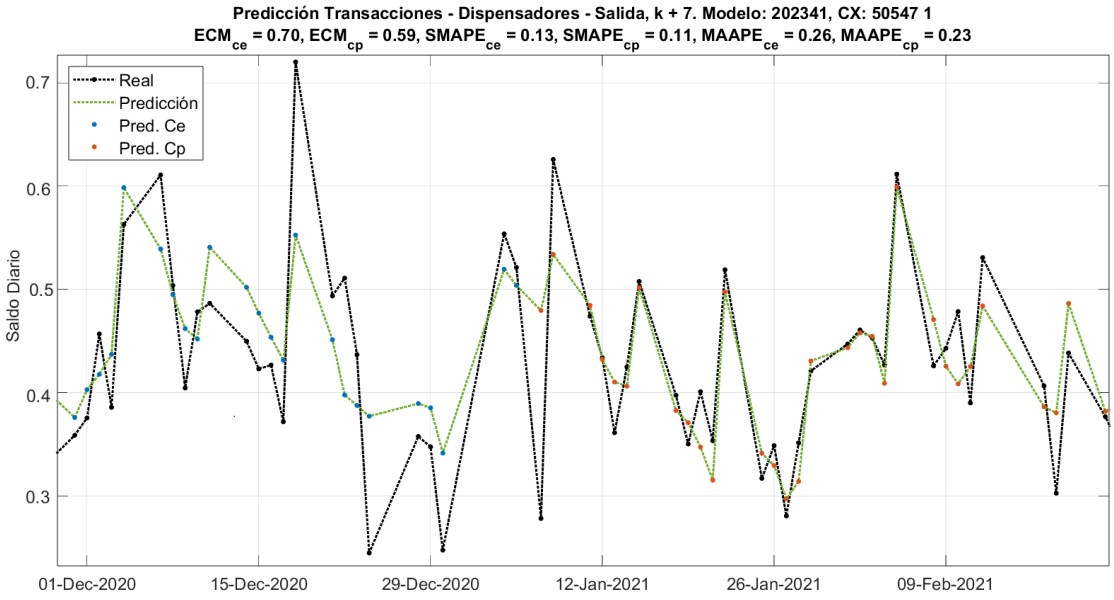


Figura 4.28 Salidas reales frente a las predichas por el modelo RBF 202341.

Se observa que ambas variantes de modelos parecen proporcionar resultados similares. No obstante, los modelos que utilizan la entrada B2@d pueden considerarse superiores para esta serie

temporal por las siguientes razones:

1. **Menores errores en el conjunto de prueba:** Los modelos con entrada B2@d presentan mejores predicciones para el conjunto de prueba. Esto se puede observar en las gráficas que presentan las salidas reales frente a las predichas por ambos tipos de modelos. Esta mejora del rendimiento en el conjunto de prueba es esencial ya que demuestra la capacidad de los modelos para generalizar y comportarse de manera efectiva en datos no vistos. Esta capacidad de generalización indica que estos modelos están capturando patrones más significativos y relevantes en la serie temporal.

2. **Optimización de recursos:** Aunque los modelos con entrada B1@d pueden mostrar un menor error en el conjunto de entrenamiento, esto puede deberse a un posible sobreajuste a los datos de entrenamiento. Estos modelos utilizan un mayor número de neuronas y pasadas, lo que no es deseable en términos de eficiencia y generalización.

Estos resultados respaldan la elección de la entrada B2@d como la más adecuada para esta serie temporal.

5 Conclusiones

En este Trabajo de Fin de Grado, se ha llevado a cabo un análisis exhaustivo de dos series temporales altamente contrastantes: la demanda de energía eléctrica y las transacciones bancarias. Los hallazgos de nuestras investigaciones arrojan luz sobre la idoneidad de diversos algoritmos de predicción y la relevancia de las entradas autorregresivas y exógenas en el proceso de predicción de series temporales.

Asimismo, se destaca la importancia de las etapas de preprocesamiento y análisis de datos, que sirven como base sólida para la construcción de modelos de predicción efectivos. El preprocesamiento garantiza la calidad y coherencia de los datos, un aspecto esencial para obtener resultados confiables. Por otro lado, el análisis de datos permite una comprensión profunda de la naturaleza de las series, la identificación de patrones, tendencias y datos faltantes, y, en consecuencia, la definición de entradas autorregresivas exógenas adecuadas para cada serie.

En el caso de la serie de demanda de energía eléctrica, se ha adoptado un enfoque dual al considerar tanto el consumo diario como el consumo horario. Este enfoque ha permitido una comprensión más completa de las variaciones en la demanda eléctrica a diferentes niveles de granularidad temporal. Los resultados han indicado una naturaleza predominantemente lineal, especialmente en la serie del consumo horario, lo que ha llevado a la elección del algoritmo ARX como la mejor opción para las predicciones. Este algoritmo ha demostrado ser efectivo al proporcionar resultados sólidos sin requerir una gran capacidad computacional ni recursos adicionales. Además, la inclusión de entradas exógenas no se ha considerado esencial, ya que las entradas autorregresivas por sí solas han producido predicciones precisas. En este contexto, los modelos RBF, a pesar de su potencia y complejidad, no han ofrecido ventajas sustanciales en comparación con el ARX.

Por otro lado, en el caso de la serie temporal de transacciones bancarias, se ha presentado un escenario completamente diferente. Los modelos ARX han exhibido un rendimiento deficiente, con errores notoriamente altos que no han permitido obtener predicciones útiles. En contraste, los modelos RBF se han adaptado mejor a las características y tendencias no lineales de esta serie. Además, el descubrimiento de una significativa falta de datos en la serie durante la etapa de análisis ha llevado a la creación de dos tipos distintos de entradas autorregresivas como una solución efectiva para abordar el problema. La inclusión de entradas exógenas para capturar los hábitos de consumo de los clientes y las tendencias estacionales también ha conducido a mejoras sustanciales en las predicciones.

Es importante destacar que, aunque los modelos RBF han mejorado significativamente las predicciones en la serie de transacciones bancarias, los errores aún no alcanzan un nivel adecuado para su uso en la toma de decisiones críticas para el banco, como la gestión logística. No obstante, estos modelos se han convertido en un punto de referencia valioso para algoritmos más avanzados, como los mapas autoorganizativos (SOM) y los algoritmos genéticos. La información obtenida sobre las entradas autorregresivas y exógenas que han contribuido a las mejores predicciones será fundamental para aplicaciones futuras más avanzadas en el campo de la predicción de series temporales.

En resumen, este Trabajo de Fin de Grado ha subrayado la importancia de seleccionar el algoritmo de predicción adecuado según las características de la serie temporal y ha resaltado el papel crucial de las entradas autorregresivas y exógenas en la mejora de las predicciones. Los resultados obtenidos han proporcionado una base sólida para futuras investigaciones y aplicaciones avanzadas en el ámbito de la predicción de series temporales.

Índice de Figuras

Índice de Tablas

Bibliografía

[1] Stuart J. Russell and Peter Norvig. *Artificial intelligence: a modern approach*. Prentice Hall series in artificial intelligence. Pearson, Boston Columbus Indianapolis New York San Francisco Upper Saddle River Amsterdam Cape Town Dubai London Madrid Milan Munich Paris Montreal Toronto Delhi Mexico City Sao Paulo Sydney Hong Kong Seoul Singapore Taipei Tokyo, third edition, global edition edition, 2016.

[2] Trevor Hastie, Robert Tibshirani, and J. H. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer series in statistics. Springer, New York, NY, 2nd ed edition, 2009.

[3] Tom M. Mitchell. *Machine learning*. McGraw-Hill series in Computer Science. McGraw-Hill, New York, nachdr. edition, 1997.

[4] BBVA. Te contamos qué es el 'machine learning'y cómo funciona.

[5] A. L. Samuel. Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, 3(3):210–229, July 1959.

[6] Machine Learning: Los orígenes y la evolución hasta la actualidad, September 2020.

[7] Ruel V. Churchill and James Ward Brown. *Complex variables and applications*. Brown and churchill series. McGraw-Hill Education, New York, NY, ninth edition edition, 2014.

[8] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015.

[9] International Neural Network Society and IEEE Computational Intelligence Society, editors. *2015 International Joint Conference on Neural Networks (IJCNN 2015): Killarney, Ireland, 12 - 17 July 2015*. IEEE, Piscataway, NJ, 2015.

[10] Ligdi González. ¿Qué son las Redes Neuronales Artificiales?, November 2021.

[11] Red neuronal artificial, August 2023. Page Version ID: 153414927.

[12] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, November 1997.

[13] Everette S. Gardner and Ed. Mckenzie. Forecasting Trends in Time Series. *Management Science*, 31(10):1237–1246, October 1985.

[14] Peter R. Winters. Forecasting Sales by Exponentially Weighted Moving Averages. *Management Science*, 6(3):324–342, April 1960.

[15] James M. Dobbie. Forecasting Periodic Trends by Exponential Smoothing. *Operations Research*, 11(6):908–918, December 1963.

[16] Jimeng Shi, Mahek Jain, and Giri Narasimhan. Time Series Forecasting (TSF) Using Various Deep Learning Models, April 2022.

[17] ¿Qué son las redes neuronales recurrentes? | IBM.

[18] Dst Team. Recurrent Neural Network (RNN): ¿de qué se trata?, December 2021.

[19] Problema del desvanecimiento del gradiente (vanishing gradient problem), May 2018.

[20] G.Peter Zhang. Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, 50:159–175, January 2003.

[21] Arquitectura de las RNNs y modelos LSTM · Aprendizaje Profundo.

[22] Red neuronal recurrente, July 2023. Page Version ID: 152427344.

[23] Rohitash Chandra, Ayush Jain, and Divyanshu Singh Chauhan. Deep learning via LSTM models for COVID-19 infection forecasting in India. *PLOS ONE*, 17(1):e0262708, 2022.

[24] Seouk Jun Kim. The Effectiveness of Feed-forward Neural Networks in Trend-based Trading (1), August 2020.

[25] Red neuronal prealimentada, October 2022. Page Version ID: 146859609.

[26] Kaivan Kamali. Deep Learning (Part 1) - Feedforward neural networks (FNN), March 2023.

[27] Ervin Ceperic, Vladimir Ceperic, and Adrijan Baric. A Strategy for Short-Term Load Forecasting by Support Vector Regression Machines. *IEEE Transactions on Power Systems*, 28(4):4356–4364, November 2013.

[28] Sina E. Charandabi and Kamyar Kamyar. Using A Feed Forward Neural Network Algorithm to Predict Prices of Multiple Cryptocurrencies. *European Journal of Business and Management Research*, 6(5):15–19, September 2021.

[29] Qué es ARX Concepto y definición. Glosario.

[30] Juan M. Gutierrez. ¿Qué es el Modelo ARMAX y la diferencia con el modelo ARIMA?, May 2020.

[31] Redes de funcion de base radial (RBF).

[32] MSE - Mean Squared Error — Permetrics 1.4.3 documentation.

[33] MAE - Mean Absolute Error — Permetrics 1.4.3 documentation.

[34] MAAPE - Mean Arctangent Absolute Percentage Error — Permetrics 1.4.3 documentation.

[35] SMAPE - Symmetric Mean Absolute Percentage Error — Permetrics 1.4.3 documentation.

Entradas autorregresivas del consumo eléctrico (A):

- **A1@d**: se encargará de tomar los datos del consumo eléctrico total de cada día.
- **A2@d**: se encargará de tomar los datos del consumo eléctrico de cada hora.

Entradas exógenas del consumo eléctrico

- **A3@d:** Consumo horario del mismo día de la semana anterior.
- **A4@d:** Consumo horario del mismo día del año anterior.
- **T1@d:** Temperatura media del día anterior.

Entradas autorregresivas de las transacciones de los dispensadores del banco (B):

- **B1@d:** tomará el agregado de la cantidad diaria dispensada en los dispensadores de la sucursal el día k+d. En caso de no haber transacciones registradas en el día k+d, se tomará entonces el día anterior más cercano con transacciones.
- **B2@d:** tomará el agregado de la cantidad diaria dispensada en los dispensadores de la sucursal el día k+d, pero en caso de no existir transacciones para el día k+d, se tomará el día anterior del mismo tipo (por ejemplo, otro viernes) más próximo con transacciones.

Entradas exógenas de las transacciones de los dispensadores del banco

- **B3@d:** Número de transacciones en los dispensadores de la sucursal el día k+d.
- **B4@d:** Media de la cantidad diaria dispensada en los dispensadores de la sucursal los últimos 5 días con transacciones previos al día k+d, incluido el día k+d.
- **B5@d:** Mínima cantidad diaria dispensada en los dispensadores de la sucursal los últimos 5 días con transacciones previos al día k+d, incluido el día k+d.
- **B6@d:** Máxima cantidad diaria dispensada en los dispensadores de la sucursal los últimos 5 días con transacciones previos al día k+d, incluido el día k+d.

Entradas exógenas de tipo calendario (C)

- **C1@d:** Identificador asignado a cada día de la semana.
- **C2@d:** Identificador asignado a cada día del mes.
- **C3@d:** Identificador asignado a cada mes.
- **C4@d:** Coeficiente aplicado a cada tipo de día para ponderar su importancia relativa. Específicamente, se asigna 0.7 para los viernes y 0.4 para el resto de los días de la semana en la serie de transacciones.

Entradas exógenas derivadas (D)

- **D1@d y D2@d:** seno y coseno del día de la semana respectivamente.
- **D3@d y D4@d:** seno y coseno del día del mes respectivamente.
- **D5@d y D6@d:** seno y coseno del mes respectivamente.
- **D7@d:** semana del mes.
- **D8@d y D9@d:** seno y coseno de la semana del mes respectivamente.