Trabajo Fin de Máster

Modelos de inflación: predicción y análisis de impactos

Ana Blanco Bugueiro

Máster en Técnicas Estadísticas

Curso 2023-2024

Propuesta de Trabajo Fin de Máster

Título en galego: Modelos de inflación: predición e análise de impactos
Título en español: Modelos de inflación: predicción y análisis de impactos
English title: Inflation models: prediction and impact analysis
Modalidad: Modalidad B
Autora: Ana Blanco Bugueiro, Universidad de Santiago de Compostela
Director: Guillermo López Taboada, Universidad de La Coruña;
Tutores: Teresa Veiga Rodríguez, ABANCA; Sergio Díaz Canosa, ABANCA
Breve resumen del trabajo: En el área de Planificación Estratégica y PMO de ABANCA se lleva el análisis y seguimiento del entorno macroeconómico. Esto incluye el desarrollo de modelos para proyectar indicadores macroeconómicos. Un aspecto clave es la inflación, y se ha identificado la necesidad de un modelo específico para predecirla a un año vista, además de analizar impactos de otras variables.
Recomendaciones: Este trabajo requiere conocimientos de técnicas de tratamiento de datos, herramientas de modelización estadística, así como de análisis de series temporales. Se valorarán conocimientos y experiencia en la utilización de algún software de análisis de datos (R, R shiny, IBM SPSS Modeler...) así como de manejo de bases de datos (SQL,...). Interés por trabajar en un banco innovador y comprometido con su entorno. Capacidad de trabajo autónomo y habilidades comunicativas. Interés y disposición para trabajar en un equipo multidisciplinar de alto desempeño.

Don Guillermo López Taboada, Catedrático de la Universidad de La Coruña, doña Teresa Veiga Rodríguez, Especialista de Planificación y Estudios de ABANCA, y don Sergio Díaz Canosa, Especialista de Planificación y Estudios de ABANCA, informan que el Trabajo Fin de Máster titulado

Modelos de inflación: predicción y análisis de impactos

fue realizado bajo su dirección por doña Ana Blanco Bugueiro para el Máster en Técnicas Estadísticas. Estimando que el trabajo está terminado, dan su conformidad para su presentación y defensa ante un tribunal.

En A Coruña, a 30 de mayo de 2024.

El director:

LOPEZ TABOADA GUILLERMO - 44088774M

Firmado digitalmente por LOPEZ TABOADA GUILLERMO - 44088774M
Fecha: 2024.05.30 08:42:47 +02'00'

Don Guillermo López Taboada

La tutora:

Doña Teresa Veiga Rodríguez

El tutor:

Don Sergio Díaz Canosa

La autora:

Doña Ana Blanco Bugueiro

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

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

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

Agradecimientos

En primer lugar, agradecer a ABANCA la oportunidad de llevar a cabo mi Trabajo de Fin de Máster en esta empresa. Esta experiencia me ha permitido aplicar y ampliar los conocimientos adquiridos a lo largo de mis años de formación.

Especialmente, me gustaría agradecer a mis tutores, Teresa Veiga Rodríguez y Sergio Díaz Canosa, por su valioso asesoramiento, su confianza y el apoyo brindado durante estos meses. Su orientación ha sido fundamental para el éxito de mi proyecto.

Asimismo, agradecer también a Guillermo López Taboada, el director del proyecto, por acompañarme y orientarme en la realización del mismo.

Por último, agradecer a mi familia, y en especial, a mis padres y a mi hermano, por su apoyo constante, cariño y consejos. Gracias también a mis amigos, por todos los momentos compartidos estos años y, sobre todo, a Daniela y a Alfonso, por acompañarme y apoyarme incondicionalmente.

Índice general

Índice de figuras xii

Índice de tablas xv

Resumen xix

Prefacio xxi

I Metodología 1

1. La inflación 3

1.1. El IPC . . . 3

1.1.1. Desagregaciones del IPC general . . . 4

1.2. Cálculo del índice . . . 5

1.2.1. Índices elementales y agregados . . . 6

1.2.2. Repercusiones . . . 7

1.3. Variables influyentes . . . 7

2. Análisis univariante de series temporales 9

2.1. Conceptos previos . . . 9

2.2. Modelos Box-Jenkins . . . 11

2.2.1. Series estacionarias . . . 11

2.2.2. Series no estacionarias . . . 13

2.3. Modelos de regresión con series temporales . . . 15

2.3.1. Relación lineal entre series de tiempo . . . 15

2.3.2. Modelo para los errores . . . 16

2.4. Selección, estimación, diagnosis y predicción del modelo . . . 17
2.4.1. Selección del modelo generador de la serie . . . 17
2.4.2. Estimación del modelo seleccionado . . . 18
2.4.3. Diagnosis . . . 19
2.4.4. Predicción . . . 20

3. Análisis multivariante de series temporales **21**
3.1. Modelo VAR . . . 21
3.1.1. Estimación de los coeficientes . . . 23
3.1.2. Especificación del modelo . . . 24
3.1.3. Diagnosis . . . 25
3.1.4. Predicción . . . 29
3.2. Modelo VECM . . . 30
3.2.1. Estimación y especificación del modelo . . . 31
3.2.2. Diagnosis y predicción . . . 33
3.3. Funciones impulso-respuesta (IRF) . . . 33
3.4. Descomposición de la varianza del error de predicción (FEVD) . . . 34

4. Redes neuronales *LSTM* **35**
4.1. Conceptos previos sobre redes neuronales . . . 35
4.1.1. Redes neuronales artificiales (*ANN*) . . . 36
4.1.2. Redes neuronales recurrentes (*RNN*) . . . 37
4.2. *LSTM* . . . 40
4.3. Selección de hiperparámetros . . . 41

II Aplicación práctica **43**

5. Análisis exploratorio de los datos **45**
5.1. Variables empleadas en la modelización . . . 45
5.1.1. Automatización en la descarga de los datos . . . 47
5.2. Análisis de las variables . . . 51
5.2.1. Corrección de estacionalidad . . . 53
5.2.2. Estacionariedad . . . 55

5.2.3. Análisis de causalidad y correlación entre las variables . . . 57
5.3. Hipótesis de alto nivel sobre las variables regresoras . . . 59

6. Modelos de regresión dinámica para el IPC 61
6.1. Selección de los modelos . . . 61
6.2. Validación . . . 63
6.3. Estimación y ajuste . . . 65

7. Modelos VAR y VECM para el IPC 69
7.1. Modelo VAR . . . 69
7.1.1. Selección . . . 69
7.1.2. Validación . . . 70
7.1.3. Estimación y ajuste . . . 72
7.1.4. IRF y FEVD . . . 74
7.2. Modelo VECM . . . 76
7.2.1. Selección . . . 76
7.2.2. Validación . . . 77
7.2.3. Estimación y ajuste . . . 77
7.2.4. IRF y FEVD . . . 80

8. Redes neuronales LSTM para el IPC 83
8.1. Implementación de las redes LSTM . . . 83
8.2. Preprocesamiento de los datos . . . 84
8.3. Selección del modelo . . . 85
8.4. Evaluación en la muestra de entrenamiento y de test . . . 86
8.5. Ajuste del modelo . . . 87

9. Comparativa y rendimiento de los modelos ajustados 91
9.1. Rendimiento de los modelos ajustados . . . 91
9.1.1. Predicciones . . . 91
9.1.2. *Backtesting* . . . 94
9.2. Análisis de sensibilidad a impactos . . . 94

10. Aplicación para el análisis de la inflación 97

10.1. Estructura de construcción básica de una aplicación en Shiny 97
10.2. Aplicación desarrollada para la inflación 98
10.3. Ejecución de la app en cualquier ordenador 101

11. Conclusiones y líneas futuras **103**

11.1. Conclusiones 103
11.2. Líneas futuras 104

Índice de figuras

1. Porcentajes de variación interanual del IPC y su descomposición de grupos especiales desde 2020 hasta 2023. XXII

4.1. Esquema representativo de una red neuronal con una única capa oculta, en el que cada nodo circular representa una neurona y cada flecha representa el enlace desde la salida de una neurona a la entrada de otra. 36

4.2. Esquema representativo del proceso que ocurre en un nodo de la capa oculta de una red neuronal recurrente atendiendo a Torres (2020). 38

4.3. Esquema representativo del algoritmo de *backpropagation through time* (BPTT) en una red neuronal recurrente de acuerdo con Pascanu et al. (2013). 38

4.4. Esquema representativo de la célula de memoria de una red neuronal *LSTM*. 41

5.1. Ventana de automatización con la librería `taskscheduleR`. 50

5.2. Series del IPC general y de sus componentes de grupos especiales. 51

5.3. Variaciones interanuales de las series del IPC general y sus componentes de grupos especiales. 52

5.4. De izquierda a derecha: serie del precio del petróleo en euros, serie del precio de las gasolinas en España (teniendo en cuenta el descuento del Gobierno), serie del índice FAO, serie del número de pernoctaciones en España y serie de precios de la electricidad en España. 53

5.5. Serie del IPC de bienes industriales en España original y serie corregida de estacionalidad. 54

5.6. Gráfico de autocorrelación del logaritmo de la serie del IPC de Servicios corregida de estacionalidad. 55

5.7. Correlaciones contemporáneas entre las series de las componentes del IPC sin componente estacional. 57

5.8. Correlaciones contemporáneas entre las series de las componentes del IPC sin componente estacional, estacionarias y preblanqueadas. 57

5.9. Correlaciones cruzadas entre el IPC de alimentos sin elaboración y el IPC de alimentos con elaboración, bebidas y tabaco, tras el proceso de preblanqueado. 58

5.10. Correlaciones cruzadas entre el precio de los carburantes en España y el IPC de productos energéticos, tras el proceso de preblanqueado. 58

6.1. Precio del petróleo por las variaciones interanuales positivas y negativas del precio del mismo. 62

6.2. Esquema de las variables implicadas en los modelos. 62

6.3. Funciones de autocorrelación simple de los residuos de los modelos. 64

6.4. Funciones de autocorrelación parcial de los residuos de los modelos. 65

6.5. Series reales y valores ajustados de los modelos de regresión. 67

6.6. Series del IPC reconstruido y de sus variaciones interanuales junto con el ajuste a partir de los modelos de regresión dinámica seleccionados. 68

7.1. Variación de los criterios de información de Akaike (AIC), de Hamilton-Quinn (HQ) y de Schwarz (SC) según el orden del modelo VAR. 70

7.2. Ajustes de *OLS-CUSUM* para comprobar la estabilidad estructural de cada una de las variables endógenas en el modelo $VAR(4)$ ajustado. 71

7.3. Series reales y ajustes del modelo $VAR(4)$. 73

7.4. Series del IPC reconstruido y de sus variaciones interanuales, junto con el ajuste resultante del modelo $VAR(4)$ por componentes. 74

7.5. Funciones de impulso-respuesta con impulso el IPC de productos energéticos con el modelo $VAR(4)$ ajustado. 75

7.6. Descomposición de la varianza del error del modelo $VAR(4)$. 75

7.7. Series reales y ajustes del modelo VECM. 79

7.8. Series del IPC reconstruido y de sus variaciones interanuales, junto con el ajuste resultante del modelo $VECM(3)$ con 3 ecuaciones de cointegración por componentes. 79

7.9. Funciones de impulso-respuesta con impulso el IPC de productos energéticos con el modelo $VECM(3)$ con 3 ecuaciones de cointegración ajustado. 80

7.10. Descomposición de la varianza del error del modelo VECM. 81

8.1. Error cuadrático medio en la muestra de entrenamiento y de test en cada paso (*epoch*) del entrenamiento del modelo. 87

8.2. Series reales y ajustes del modelo de redes neuronales $LSTM$ seleccionado. 88

8.3. Series del IPC reconstruido y de sus variaciones interanuales junto con el ajuste resultante del modelo $LSTM$ por componentes. 89

9.1. Serie del IPC general junto con las predicciones obtenidas con los distintos modelos ajustados en niveles (a la izquierda) y en porcentaje de variación interanual (a la derecha). 92

9.2. Serie del IPC general desde 2023 junto con las predicciones obtenidas con los distintos modelos ajustados en niveles (a la izquierda) y en porcentaje de variación interanual (a la derecha). 92

10.1. Pestaña de inicio de la aplicación. 99

10.2. Pestaña de análisis de la inflación de la aplicación. 99

10.3. Pestaña de los modelos de la aplicación. 100

10.4. Pestaña de comparativa y previsiones de la aplicación. 100

10.5. Pestaña de análisis de impactos de la aplicación. 101

10.6. Captura de pantalla de la carpeta "`AppIPC`" con la aplicación para el análisis de la inflación lista para ser ejecutada o compartida a cualquier usuario. 102

Índice de tablas

1.1. Ponderaciones de las componentes de grupos especiales del IPC en 2023 y 2024 publicadas por el INE. 5

5.1. Organismos proveedores de los datos, abreviaturas empleadas y referencia de los mismos. 45

5.2. Información principal sobre las variables implicadas en los diferentes modelos. 46

5.3. P-valores resultantes de los tests de estacionalidad sobre las series del IPC. 53

5.4. P-valores resultantes de los tests de estacionalidad sobre las series asociadas a las variables regresoras. 54

5.5. P-valores resultantes de los tests de estacionariedad de Dickey-Fuller sobre las series del IPC. 55

5.6. P-valores resultantes de los tests de estacionariedad de Dickey-Fuller sobre las series del IPC transformadas mediante un logaritmo y una diferencia regular. 56

5.7. P-valores resultantes de los tests de estacionariedad de Dickey-Fuller sobre las series asociadas a las variables regresoras. 56

5.8. P-valores resultantes de los tests de estacionariedad de Dickey-Fuller sobre las series asociadas a las variables regresoras tras aplicar logaritmos y una diferencia regular. . . . 56

6.1. Detalles de los modelos ajustados: variables implicadas, transformaciones aplicadas, período de inicio y fin, órdenes del modelo y BIC. 63

6.2. P-valores resultantes de la fase de validación de los residuos de los modelos e indicador de significación de los coeficientes de los mismos. 63

6.3. Ajuste de los modelos de regresión dinámica: estimación y significación de sus coeficientes. 66

7.1. Valores de los criterios de información de Akaike (AIC), de Hamilton-Quinn (HQ) y de Schwarz (SC) según el orden m del modelo VAR, tras descartar algunas variables exógenas. 70

7.2. Estadísticos y p-valores resultantes de la fase de validación de los residuos del modelo VAR. 71

7.3. Estadísticos y p-valores resultantes de los contrastes de causalidad en el sentido de Granger (1969) en el modelo $VAR(4)$. 72

7.4. Estadísticos y cuantiles teóricos asociados al test del rango de cointegración de Johansen (1995) para el modelo $VECM(3)$. 77

7.5. Estadísticos y p-valores resultantes de la fase de validación de los residuos del modelo $VECM(3)$. 77

8.1. Posibles valores considerados para los hiperparámetros implicados en la red neuronal LSTM. 85

8.2. Error cuadrático medio en la muestra de entrenamiento y de test con la red neuronal LSTM escogida. 87

9.1. Predicciones del IPC y de sus componentes de grupos especiales para 2024 en variaciones interanuales obtenidas. por un lado. con los modelos ajustados y por otro, a partir de otras fuentes. 93

9.2. Predicciones del IPC para el primer trimestre del 2024 en variaciones interanuales obtenidas con los modelos ajustados y los correspondientes datos reales. 93

9.3. Porcentajes de variación de la media de las predicciones en los meses de octubre, noviembre y diciembre de los modelos recortados hasta septiembre de 2023 frente a la media de los datos reales en esos tres meses. 94

9.4. Aumento en las predicciones del precio de los carburantes en 2024 suponiendo que el precio del petróleo es un 10 %, 50 % y 100 % superior a las hipótesis de alto nivel fijadas para dicho precio. 95

9.5. Predicciones del IPC y de sus componentes de grupos especiales para 2024 en variaciones interanuales suponiendo que el precio del petróleo es un 10 %, 50 % y 100 % superior a las hipótesis de alto nivel fijadas por los expertos del área para dicho precio en 2024. . . 95

Resumen

Resumen

En el área de Planificación Estratégica y PMO de ABANCA se lleva a cabo un análisis y seguimiento continuo del entorno macroeconómico. Además, esto se complementa con el desarrollo de modelos estadísticos que contribuyen a la elaboración de previsiones de las principales variables macroeconómicas. En este marco, dada la alta volatilidad de los precios en el mercado de los últimos años, surge la necesidad de analizar la inflación, medida por el IPC en España.

Para ello, en este proyecto se ajustan modelos de regresión dinámica, modelos VAR, modelos VECM y redes neuronales LSTM, que sirven no solo para modelizar el IPC y construir predicciones, sino también para analizar el efecto de posibles impactos en la economía española sobre el IPC. Los modelos se ajustan desagregando el IPC por las cinco componentes de grupos especiales de alimentos sin elaboración, alimentos con elaboración, bebidas y tabaco, productos energéticos, servicios y bienes industriales sin productos energéticos; e incluyendo a su vez otras variables económicas exógenas en los mismos. Como paso previo, se analizan todas las variables implicadas, aplicando las pertinentes correcciones o transformaciones en cada caso.

Las predicciones de los modelos se comparan entre sí, con las previsiones publicadas por otros organismos y con los datos reales disponibles en el momento previo a la entrega de esta memoria, obteniéndose resultados satisfactorios. Adicionalmente, se desarrolla una aplicación web, que facilita a los analistas del banco interactuar con los modelos y visualizar los resultados obtenidos, facilitando de este modo su uso, sin requerir conocimientos de las herramientas utilizadas en el desarrollo del proyecto, entre ellas R Studio.

Resumo

Na área de Planificación Estratéxica e PMO de ABANCA lévase a cabo unha análise e seguimento continuo da contorna macroeconómica. Ademais, isto compleméntase co desenvolvemento de modelos estatísticos que contribúen á elaboración de previsións das principais variables macroeconómicas. Neste marco, dada a alta volatilidade dos prezos no mercado dos últimos anos, xorde a necesidade de analizar a inflación, medida polo IPC en España.

Para iso, neste proxecto axústanse modelos de regresión dinámica, modelos VAR, modelos VECM e redes neuronais LSTM, que serven non só para modelizar o IPC e construír predicións, senón tamén para analizar o efecto de posibles impactos na economía española sobre o IPC. Os modelos axústanse desagregando o IPC polas cinco compoñentes de grupos especiais de alimentos sen elaboración, alimentos con elaboración, bebidas e tabaco, produtos enerxéticos, servizos e bens industriais sen produtos enerxéticos; e incluíndo á súa vez outras variables económicas esóxenas nos mesmos. Como paso previo,

analízanse todas as variables implicadas, aplicando as pertinentes correccións ou transformacións en cada caso.

As predicións dos modelos compáranse entre si, coas previsións publicadas por outros organismos e cos datos reais dispoñibles no momento previo á entrega desta memoria, obténdose resultados satisfactorios. Adicionalmente, desenvolveuse unha aplicación web, que facilita aos analistas do banco interactuar cos modelos e visualizar os resultados obtidos, facilitando deste modo o seu uso, sen requerir de coñecementos das ferramentas empregadas no desenvolvemento do proxecto, entre elas R Studio.

Abstract

In the Strategic Planning and PMO area at ABANCA, continuous analysis and monitoring of the macroeconomic environment are carried out. In addition, this is complemented by the development of statistical models which contribute to the forecasting of key macroeconomic variables. Within this framework, given the high volatility of prices in the market in recent years, there is a need to analyze inflation, measured by the CPI in Spain.

In order to achieve this, this project adjusts dynamic regression models, VAR models, VECM models and LSTM neural networks, which serve not only to model the CPI and build predictions, but also to analyze the effect of potencial impacts of the Spanish economy on the CPI. The models are adjusted by disaggregating the CPI by the five components of special groups of unprocessed food, processed food, beverages and tobacco, energy products, services and industrial goods without energy products; and including other exogenous economic variables in them. As a preliminary step, all the variables involved are analyzed, applying necessary corrections or transformations as needed.

The model predictions are compared with each other, as well as with forecasts published by other organizations and real data available prior to the submission of this report, resulting in satisfactory outcomes. Additionally, a web application is developed, which makes it easier for the bank analysts to interact with the models and visualize the obtained results, thus facilitating its use, whitout requiring knowledge of the tools used in the project development, including R Studio.

Prefacio

El área de Planificación Estratégica y PMO (*Project Management Office*) de ABANCA se encarga de la elaboración y seguimiento del plan estratégico de la entidad. Más concretamente, sus principales funciones son las de definición de la estrategia, elaborándose un plan anual; control de seguimiento del plan; gobierno del dato; y seguimiento de proyectos, coordinando a todos los equipos implicados y asegurando el cumplimiento de todos los objetivos del banco.

Específicamente, para la planificación estratégica se lleva a cabo un análisis y seguimiento continuo del entorno macroeconómico, a partir de modelos estadísticos que contribuyen a la elaboración de previsiones de las principales variables macroeconómicas. Este análisis se focaliza en la economía española y gallega en particular, al estar la mayoría del negocio de la entidad concentrado en esta región.

En este contexto, se recurre al análisis estadístico de series temporales, empleándose modelos para predecir variables como el Producto Interior Bruto (PIB), la tasa de paro, las importaciones y exportaciones o el precio de la vivienda, entre otras.

En los últimos años, la inflación ha tomado una especial relevancia debido a la alta volatilidad de los precios en el mercado. Así, desde la entidad surge la necesidad de desarrollar nuevos modelos o metodologías que permitan explicar la inflación aprovechando información de alta frecuencia, como indicadores o variables diarias. Este enfoque permite no solo predecir esta variable a largo plazo, sino también aprovechar esta información disponible para anticipar la inflación en el mes en curso.

En la Figura 1 pueden verse los ritmos de variación interanual de la inflación en España por componentes en los últimos tres años. Se puede observar cómo después de un período deflacionario en el año 2020, marcado por la pandemia, los precios comienzan a subir de manera sostenida a partir de marzo de 2021. Analizando las componentes, es claro que esta subida se debe especialmente a los precios de la energía, y más concretamente, de los combustibles y la electricidad; y que después, por efecto contagio, se va trasladando al resto de componentes.

Tras esto, en 2022, se alcanza un máximo histórico, al agravarse la situación provocada por la crisis sanitaria con la guerra entre Ucrania y Rusia, que produjo que los precios energéticos (particularmente, del gas natural), y de materias primas encareciesen. A partir de este instante, puede verse una moderación gradual en la inflación, aún manteniéndose en niveles muy elevados durante todo el año.

Finalmente, en 2023 empieza a desacelerarse, cerrando el año con la inflación más baja desde 2022. Esta sigue creciendo, pero a ritmos más moderados, manteniéndose aún muy por encima de los niveles de 2020. Haciendo un análisis más detallado por componentes, se puede observar que esta bajada se debe al descenso de los precios de la energía, pero otros sectores como son la alimentación y los servicios siguen en niveles muy elevados.

Ante esta situación de inestabilidad en los precios, parece importante contar con alguna herramienta que permita adelantarse a *shocks* económicos como han sido la crisis sanitaria o la guerra de Ucrania y llevar un seguimiento continuo de la evolución de la inflación.

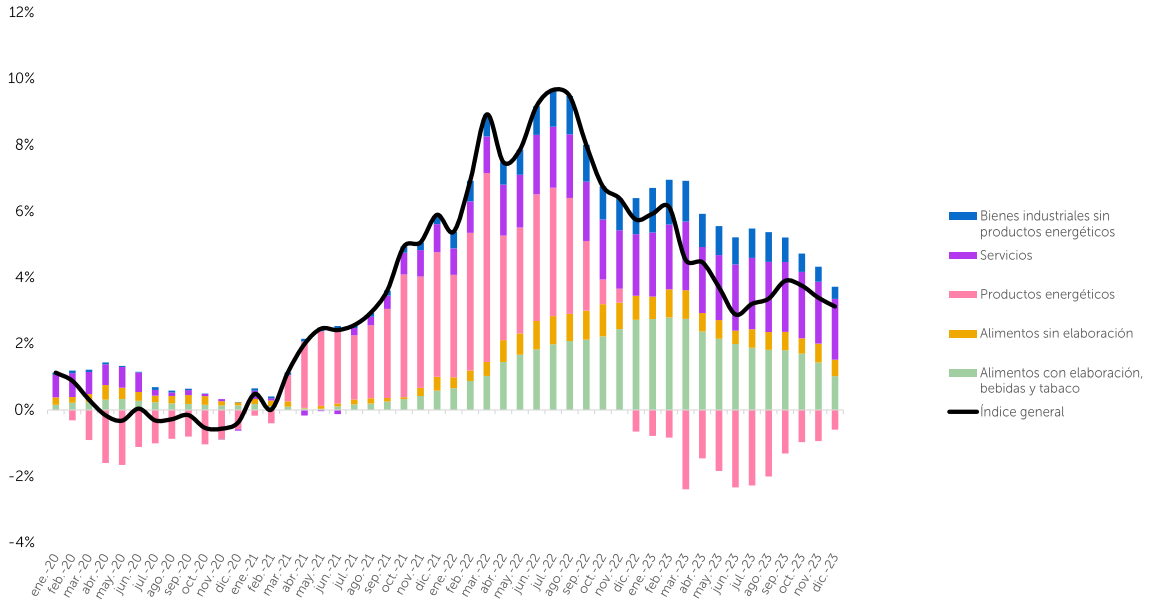

Figura 1: Porcentajes de variación interanual del IPC y su descomposición de grupos especiales desde 2020 hasta 2023.

Así, el objetivo del presente proyecto será el análisis de la inflación en España, medida a partir del Índice de Precios de Consumo (IPC) que elabora el Instituto Nacional de Estadística (INE). Se estudiarán técnicas estadísticas en el marco de las series temporales que sirvan no solo para modelizar el IPC y construir predicciones, sino también para analizar el efecto de posibles impactos (como puede ser el encarecimiento de materias primas) sobre el IPC.

Existen múltiples posibilidades de modelos para series temporales que pueden ser aplicables en el contexto económico y son numerosos los organismos que cuentan con modelos de proyección macroeconómica y en particular, del IPC. Así pues, como trabajo previo al análisis y modelización de la inflación, se realizó una labor de investigación acerca de las metodologías utilizadas con el mismo fin.

A modo de ejemplo, se puede citar a Durán et al. (2012) en el contexto univariante, donde se ajustan modelos de efectos fijos dinámicos, corrección del equilibrio y factores dinámicos por componentes para la inflación mexicana. En esta misma línea, en el Banco Central Europeo (Benalal et al., 2004) se desarrollan modelos ARIMA, VAR y BVAR (VAR bayesiano), también sobre las componentes de la inflación.

Continuando con los modelos VAR, Zubieta Huaygua (2016) trata modelos VAR con el IPC y el PIB en la economía boliviana, Ceasar (2006) ajusta una combinación de múltiples modelos VAR sobre distintas variables económicas de Suiza y el Banco de España (González Mínguez et al., 2022) recurre a modelos VAR por componentes. Asimismo, desde el Banco de España (López et al., 2022) también explican la inflación a partir de un modelo BVAR con el precio del gas y el precio del petróleo.

Adicionalmente, desde BBVA *Research* (Dong, 2020) predicen la inflación china utilizando tanto modelos univariantes como multivariantes, desarrollando modelos ARIMA, VAR y VECM que incluyen indicadores económicos y monetarios como el PIB, el desempleo, el precio del petróleo o el número de préstamos otorgados.

En cuanto a modelos en el marco del *machine learning*, el Banco Central Europeo (Lenza et al., 2023) propone un *Quantile Regression Forest* (QRF), incluyendo todo tipo de variables económicas;

Barkan et al. (2023) sugieren un modelo de redes neuronales recurrentes jerárquicas (HRNN) para la inflación por componentes de Israel; y Paranhos (2021) escoge un modelo de redes neuronales LSTM para la inflación en Inglaterra, empleando una amplía batería de variables económicas.

En este proyecto se recurre a cuatro de los modelos mencionados: los modelos de regresión dinámica en el contexto univariante y los modelos VAR, VECM y las redes LSTM, en el multivariante. El desarrollo de la memoria se estructura en dos partes muy diferenciadas. Una primera parte teórica con el desarrollo metodológico de los modelos empleados, seguida de la aplicación práctica de dichos modelos al caso del IPC en España. Los detalles de los capítulos desarrollados en cada una de estas partes se incluyen a continuación.

Metodología

En el Capítulo 1 se define formalmente el IPC y se desarrolla la metodología que se sigue desde el INE para su cálculo, con el fin de contextualizar la variable objetivo del proyecto e introducir una de las desagregaciones más comunes de este índice, la de grupos especiales, que se empleará a lo largo del trabajo. Además, se reflexiona, desde un punto de vista económico, acerca de qué puede afectar en los precios y así generar una batería de posibles variables influyentes en el IPC.

Los Capítulos 2, 3 y 4 reúnen el marco teórico estadístico de los modelos empleados. En el Capítulo 2 se comienza revisando los principales conceptos de series temporales, necesarios para el resto del estudio, para después centrarse en la metodología Box-Jenkins y su extensión a los modelos de regresión dinámica, una generalización de los modelos de regresión lineal para el caso de las series temporales.

A continuación, en el Capítulo 3 se amplía el estudio de las series temporales al caso multivariante, estudiándose los procesos autorregresivos vectoriales (VAR), como generalización de los procesos autorregresivos que se introducen en el anterior capítulo. Además, se trata el problema de cointegración, habitual al trabajar con series económicas, y se introducen los modelos de corrección de errores vectoriales (VECM), una transformación de los modelos VAR que resuelve este problema. Adicionalmente, se investigan dos herramientas muy útiles para el análisis de impactos con estos modelos: las funciones impulso-respuesta (IRF) y la descomposición de la varianza del error de predicción (FEVD).

Por último, en el Capítulo 4, se revisa una metodología menos tradicional como son las redes neuronales, centrándose en las redes neuronales de memoria a corto y largo plazo (*Long Short-Term Memory*, LSTM), una extensión de las redes neuronales recurrentes (*Recurrent Neural Networks*, RNN), diseñadas específicamente para tratar con datos secuenciales.

Aplicación práctica

En la segunda parte de esta memoria se incluyen los aspectos relacionados con la aplicación práctica de la metodología expuesta en la primera parte, haciendo uso del *software* estadístico `R` (R Core Team, 2023).

Para empezar, en el Capítulo 5 se describen las variables empleadas, detallándose su fuente de obtención y un método de automatización en la descarga de los datos de las mismas. Asimismo, se realiza un análisis exploratorio de las variables empleadas, aplicando en algún caso transformaciones sobre los datos para facilitar su posterior modelización. A su vez, se detallan ciertas hipótesis de alto nivel realizadas sobre las variables que jugarán el papel de regresoras en los modelos.

En los Capítulos 6, 7 y 8 se ponen en práctica las metodologías desarrolladas en la primera parte, seleccionando en cada caso los modelos más adecuados y detallando la validez de los ajustes realizados.

Las predicciones de los modelos escogidos se incluyen en el Capítulo 9, en el que se comparan además con las previsiones publicadas por otros organismos e incluso con los datos reales disponibles en el momento previo a la entrega de esta memoria. Esto permite tener una visión de la bondad de los modelos ajustados en términos predictivos, apoyada con un ejercicio de *backtesting* para evaluar la precisión de los mismos. Por otro lado, se analiza la sensibilidad de los modelos ante impactos en la economía, una característica requerida en el marco de los modelos desarrollados en la entidad al permitir anticiparse al comportamiento de la economía ante situaciones de crisis o estrés.

Finalmente, en el Capítulo 10, se presenta una aplicación web desarrollada para recoger los resultados de este trabajo, que permite al usuario interactuar con los modelos propuestos a través de una interfaz gráfica, mediante elementos visuales como botones, menús e iconos. Esto hace que sea fácil de manejar por el usuario, independientemente de sus conocimientos en el lenguaje de programación empleado en el proyecto.

Se concluye esta memoria con el Capítulo 11, analizando los propósitos iniciales del trabajo y comparando los modelos escogidos, con el fin de concluir si alguna de las metodologías desarrolladas es más adecuada para el objetivo planteado. Por último, se comentan algunas líneas futuras de investigación que han quedado abiertas con la realización de este proyecto.

Parte I

Metodología

Capítulo 1

La inflación

Este capítulo servirá de introducción para contextualizar la inflación, una variable macroeconómica muy relevante, que en el año 2023 ha tomado protagonismo y que será objeto de estudio de este trabajo.

En cualquier economía, los precios de los bienes y servicios están sujetos a cambios y la inflación se puede definir como el crecimiento de dicho nivel de precios. En el caso de España, esta se mide a partir del Índice de Precios de Consumo, de ahora en adelante IPC.

El incremento de los niveles generales de precios en una economía hace que el dinero pierda valor, puesto que si hay un gran aumento de precios, podremos comprar menos productos con la misma cantidad de dinero. Es por este motivo por el que resulta de interés analizar la inflación y generar una herramienta que nos permita predecir cómo puede evolucionar en el futuro.

En este caso nos centraremos en la economía española, con lo que en la Sección 1.1 se comenzará introduciendo el IPC. A continuación, en la Sección 1.2 se explicará la metodología que se sigue en el Instituto Nacional de Estadística (INE), encargado de su publicación, para su cálculo y, por último, en la Sección 1.3 se analizarán variables que podrían resultar influyentes a la hora de modelizar la inflación, desde un punto de vista económico.

1.1. El IPC

El Índice de Precios de Consumo (IPC) es un índice económico cuyo objetivo es medir la evolución del nivel de precios de los bienes y servicios de consumo adquiridos por los hogares. Tiene numerosas aplicaciones y de gran importancia, entre las que destaca su uso como medida de la inflación. En el caso de España, es publicado mensualmente por el Instituto Nacional de Estadística (INE) y se calcula sobre una cesta de consumo que incluye todas las categorías de bienes y servicios consumidos por las familias residentes en España, acogiéndose a la metodología recogida en INE (2016).

El IPC se basa en dos cualidades de las que depende su precisión: la representatividad y la comparabilidad temporal.

Por un lado, el grado de representatividad viene determinado por la proximidad de este indicador a la realidad económica del momento. Para ello, los artículos que conforman la cesta deben ser los más consumidos por la mayoría de la población y la importancia de cada uno de los artículos en la cesta debe ser representativa de las tendencias de consumo.

Por otro lado, la comparabilidad temporal se refiere a la necesidad de que los elementos que definen el IPC se mantengan constantes a lo largo del tiempo. El índice en sí es un indicador que carece de significado si no se establece una comparación entre índices de diferentes períodos y, de este modo, se consigue que las variaciones en el IPC solo se deban a cambios en los precios de los artículos incluidos y no a cualquier cambio metodológico, que es lo realmente interesante para medir la evolución del nivel de precios.

Para medir el IPC se considera un período en el que este índice se hace igual a 100, que se denomina período base o período de referencia del índice. Habitualmente se trata de un período anual y lo que se hace es considerar la media aritmética de los doce índices mensuales del año como 100 y calcular los índices de cada uno de los períodos con respecto a dicho año.

Actualmente, se emplea el IPC con base 2021 y la cesta de la compra contiene 955 artículos, cuyos precios son obtenidos atendiendo a dos metodologías distintas. La mayoría de los precios que se utilizan son recogidos mediante visita personal a los establecimientos, no obstante, el INE trabaja constantemente en el desarrollo de nuevos métodos basados en la explotación de registros administrativos y el uso de dispositivos electrónicos de recogida, con el fin de producir sus estadísticas de forma más eficiente. En este contexto, en los últimos años se ha implementado la metodología del *scanner data*, que utiliza las bases de datos de las empresas como sustituto de la recogida de los precios en cada establecimiento. Para más detalles sobre el IPC base 2021 puede recurrirse a INE (2022) y en particular, las especificaciones de esta nueva metodología pueden consultarse en INE (2020).

En lo que sigue, en la Sección 1.1.1 se explicarán las principales desagregaciones del IPC general, que resultarán de utilidad a la hora de analizar su evolución y ajustar modelos estadísticos para este índice. Posteriormente, en la Sección 1.2 se indicará la fórmula del cálculo del índice general, explicando brevemente cómo se construyen también los índices desagregados de cada uno de los productos o grupos de productos y cómo estos repercuten sobre el índice general.

1.1.1. Desagregaciones del IPC general

Mensualmente se publica el índice para el IPC general, así como los índices correspondientes con diferentes desagregaciones tanto geográficas como funcionales. Para cada una de las diferentes desagregaciones se cuenta con el dato de la ponderación sobre el índice general, siendo estas ponderaciones revisadas cada año atendiendo a la fracción del gasto total que se realiza en dicho agregado.

La desagregación geográfica permite conocer la variación de los precios por comunidad autónoma y por provincias. Asimismo, los artículos de la cesta de la compra se agregan en subclases, que a su vez se agregan en clases, posteriormente en subgrupos y, por último, en grupos.

Una desagregación común del índice general es la de grupos especiales. En particular, el presente trabajo se basará en las componentes del IPC para España de:

- Alimentos con elaboración, bebidas y tabaco.
- Alimentos sin elaboración.
- Productos energéticos.
- Servicios.[1]
- Bienes industriales sin productos energéticos.

[1] Debido al alto peso de los servicios sobre el IPC general, en algunos casos se desagregará esta componente en dos más, separando por un lado el IPC de los restaurantes y hoteles del resto de servicios.

Estas 5 componentes, de acuerdo con sus ponderaciones, que se recogen en la Tabla 1.1, agregan el total del IPC.

	Ponderación 2023	**Ponderación 2024**
IPC Alimentos con elaboración, bebidas y tabaco	16.82 %	16.67 %
IPC Alimentos sin elaboración	6.76 %	6.34 %
IPC Productos energéticos	9.72 %	9.36 %
IPC Servicios	45.63 %	46.86 %
IPC Bienes industriales sin productos energéticos	21.06 %	20.77 %
IPC general	100.00 %	100.00 %

Tabla 1.1: Ponderaciones de las componentes de grupos especiales del IPC en 2023 y 2024 publicadas por el INE.

Por un lado, entre los alimentos sin elaboración se encuentran por ejemplo las frutas, verduras, huevos, pescado o carne fresca, mientras que entre los alimentos elaborados podemos encontrar los cereales, harinas, productos lácteos, etc. En cuanto a los productos energéticos, incluyen el petróleo, los carburantes y el gas natural, entre otros. Asimismo, los bienes industriales sin productos energéticos engloban la ropa, el calzado, los muebles, los electrodomésticos, los materiales de construcción, los automóviles,...

Por otra parte, los servicios agrupan tanto los restaurantes y hoteles, como los servicios de transporte, enseñanza, telefonía, la sanidad, los seguros, las tasas administrativas y otro tipo de tasas y servicios.

Teniendo en cuenta la descomposición anterior, si se excluyen los precios de alimentos no elaborados y de la energía, se obtiene el IPC subyacente. Dada la alta volatilidad de los precios de los alimentos no elaborados y de los productos energéticos, en comparación con el resto de los artículos de la cesta de consumo, el IPC subyacente es ampliamente utilizado para obtener una mejor aproximación de cómo se comportarán los precios a largo plazo.

Además, a partir del IPC se obtiene el indicador conocido como Índice de Precios de Consumo Armonizado (IPCA), cuyo objetivo es proporcionar una medida común de la inflación que permita realizar comparaciones entre países de la Unión Europea. Este se obtiene de forma muy similar al IPC, salvo algunos aspectos en los que la metodología propuesta por la reglamentación europea no resulta adecuada para cumplir con el objetivo del IPC en España. Dichas diferencias pueden encontrarse en INE (2016).

1.2. Cálculo del índice

La fórmula de cálculo que emplea el INE para el IPC es la fórmula de Laspeyres encadenado (LE) (Allen et al., 1963), consistente en una media ponderada de cantidades por los precios del período base. Más concretamente, el índice general en el mes m del año t se expresa como

$$ {}_0I^t_{LE} = \prod_{k=1}^{t} \frac{\sum_i p_i^k q_i^{k-1}}{\sum_i p_i^{k-1} q_i^{k-1}}, $$

donde

$_0I^t_{LE}$ es el índice general, con base 0, del mes m del año t,
p^k_i es el precio del artículo i en el período k,
q^{k-1}_i es la ponderación del artículo i en el período $k-1$.

Consiste en comparar el período corriente, t, con el período base, 0, considerando las situaciones intermedias $k = 0, \ldots, t$, que se corresponden con los meses de diciembre de los años intermedios.

Así, resulta que el índice base 2021 para el mes m del año t se obtiene como producto de índices de la forma

$$\begin{aligned} {}_{21}I^{mt}_G &= {}_{21}\, I^{dic(t-1)}_G \cdot \left(\frac{{}_{dic(t-1)}I^{mt}_G}{100}\right) \\ &= {}_{21}\, I^{dic21}_G \cdot \left(\frac{{}_{dic21}I^{dic22}_G}{100}\right) \cdot \ldots \cdot \left(\frac{{}_{dic(t-2)}I^{dic(t-1)}_G}{100}\right) \cdot \left(\frac{{}_{dic(t-1)}I^{mt}_G}{100}\right), \end{aligned}$$

donde

$_{21}I^{mt}_G$ es el índice general, en base 2021, del mes m del año t,
$_{dic(t-1)}I^{mt}_G$ es el índice general, referido a diciembre del año $(t-1)$ en el mes m del año t.

Debido a la estructura de la fórmula del cálculo del IPC, no es posible obtener el índice de cualquier agregado como media ponderada de los índices agregados que lo componen, ya que no se tiene la propiedad de aditividad. De este modo, para agregar el índice general a partir de los índices desagregados por componentes, es necesario calcular las repercusiones de estos en el índice general.

1.2.1. Índices elementales y agregados

Antes de estudiar cómo calcular dichas repercusiones, es interesante ver cómo se calculan los índices elementales y agregados.

Por un lado, los índices elementales se refieren a las componentes de consumo de menor nivel de agregación, es decir, los artículos de la cesta de la compra, en cuyo cálculo no intervienen las ponderaciones. Este índice se calcula únicamente a partir del cociente del precio medio de dicho artículo en el período actual y el precio medio en el mes de diciembre del año anterior.

En cuanto a los índices agregados, se obtienen como la suma ponderada de los índices elementales de los artículos pertenecientes a dicha agregación, considerando las ponderaciones vigentes en el correspondiente año. Una vez hecho ese cálculo, es necesario encadenar estos índices agregados, sin más que multiplicar por el índice de dicho agregado en la base correspondiente para el mes de diciembre del año anterior, dividido por 100.

Al llevar a cabo este proceso de encadenado de los índices, la suma ponderada de los índices de todos los grupos no coincide con el índice general.

1.2.2. Repercusiones

La fórmula de la repercusión mensual (denotada por $mt/(m-1)t$) de un artículo o agregado i en el mes m del año t viene dada por

$$R_i^{mt/(m-1)t} = \frac{{}_{dic(t-1)}I_i^{mt} - {}_{dic(t-1)}\, I_i^{(m-1)t}}{{}_{dic(t-1)}I_G^{(m-1)t}} \cdot {}_{dic(t-1)}\, W_i \cdot 100,$$

donde

${}_{dic(t-1)}I_i^{mt}$ es el índice del artículo i, referido a diciembre del año $(t-1)$ en el mes m del año t,
${}_{dic(t-1)}I_G^{(m-1)t}$ es el índice general, referido a diciembre del año $(t-1)$ en el mes $(m-1)$ del año t,
${}_{dic(t-1)}W_i$ es la ponderación del artículo i, referida a diciembre del año $(t-1)$ en tanto por uno;

y representa la variación que el artículo o agregado i habría experimentado si los precios de todos los demás artículos o agregados que componen el índice general hubieran permanecido estables ese mes.

La suma de todas las repercusiones mensuales es igual a la variación mensual del índice general. Para más detalles puede consultarse INE (2016).

1.3. Variables influyentes

La inflación, al medir los precios finales de los bienes y servicios consumidos por los hogares, está influenciada por un conjunto muy amplio de variables económicas, financieras y estructurales.

Por un lado, se deben considerar las variables relacionadas con la oferta, es decir, todos los factores que influyen en el coste de producción o de prestación de los servicios. Entre estas variables, cabe destacar los precios de las materias primas, con un papel protagonista para los productos energéticos.

Por otro lado, la evolución de los precios también se ve condicionada por los factores relacionados con la demanda, es decir, por la capacidad de gasto de los consumidores. Entre estos factores, se pueden señalar elementos como el crecimiento del conjunto de la economía, las condiciones del mercado laboral, la renta disponible de los hogares, así como las expectativas y los niveles de confianza de los consumidores.

Finalmente, la inflación también se ve condicionada por las políticas monetarias de los bancos centrales y las políticas fiscales de los gobiernos. Factores estructurales como la competencia en los mercados, el marco regulatorio y la innovación tecnológica también pueden influir en la inflación.

El IPC, así como el resto de variables mencionadas, son variables con cierta frecuencia temporal, donde el dato correspondiente con determinado instante de tiempo depende de los datos de los instantes anteriores. Así, este tipo de variables se pueden encuadrar en el marco de las conocidas series temporales.

De este modo, en los siguientes capítulos se tratarán diversos conceptos relacionados con las series de tiempo y el análisis de las mismas, tanto de forma univariante como multivariante; pues el objetivo de este trabajo es analizar la serie de tiempo del IPC y de sus componentes y ajustar modelos adecuados en el marco de este tipo de variables; con el fin de predecir el comportamiento a corto y largo plazo del IPC en España y llevar a cabo diversos análisis de impactos en los precios.

Capítulo 2

Análisis univariante de series temporales

Como ya se adelantaba en el Capítulo 1, el IPC general y los demás índices de precios son series temporales. Existen numerosos conjuntos de datos en múltiples campos que se representan en forma de serie de tiempo y en particular, estas son muy importantes en economía, puesto que lo realmente interesante es conocer la evolución de las variables a lo largo del tiempo, más que su valor en cierto instante concreto.

Antes de estudiar modelos para ajustar y predecir las series de tiempo, será necesario revisar las principales características de este tipo de datos. Así, en el presente capítulo, se hará una revisión de los principales conceptos sobre series temporales, que se tratarán en la Sección 2.1, para después pasar a estudiar en la Sección 2.2 los modelos AR, MA y $ARMA$ para series estacionarias y los modelos $ARIMA$, una extensión de los anteriores en el caso de que no haya estacionariedad. Además, en la Sección 2.3 se ampliarán estos modelos con la inclusión de variables regresoras en los mismos. Por último, en la Sección 2.4 se revisarán metodologías para seleccionar el modelo más adecuado, estimar sus parámetros, validar las hipótesis del mismo y llevar a cabo tareas de predicción de valores futuros.

Para ello, se emplearán Aneiros (2022), Box et al. (2015), Chan and Cryer (2008), Cowpertwait and Metcalfe (2009), el Capítulo 2 de Lütkepohl and Krätzig (2004) y la Parte I de Pfaff (2008a) como principales puntos de referencia.

2.1. Conceptos previos

Es importante comenzar describiendo qué es un proceso estocástico.

Definición 2.1. *Un **proceso estocástico** se puede definir como el conjunto de variables aleatorias*

$$\{y(s,t) : s \in \mathcal{S}, t \in \mathcal{I}\}, \tag{2.1}$$

donde $\mathcal{S}$ es el espacio muestral e $\mathcal{I}$ el intervalo de tiempo.

De este modo, dado un proceso estocástico definido por (2.1), $y(\cdot, t)$ es una variable aleatoria en el espacio muestral $\mathcal{S}$ y, para cada $s \in \mathcal{S}$, $y(s, \cdot)$ es una realización del proceso estocástico respecto del intervalo de tiempo $\mathcal{I}$ (Pfaff, 2008a).

En este contexto, una **serie temporal** no es más que una realización de un proceso estocástico,

$$\{y_t\}_{t=1}^{T} = \{y_1, y_2, \ldots, y_T\}$$

con $t = 1, \ldots, T \in \mathcal{I}$, o lo que es lo mismo, una sucesión de observaciones dispuestas de forma ordenada y uniformemente espaciadas a lo largo del tiempo.

No obstante, en la práctica no siempre será posible determinar el proceso estocástico del que proviene cierta serie temporal.

La base del análisis de las series de tiempo parte del concepto de estacionariedad. Este se basa en la idea de que las leyes de probabilidad que gobiernan el comportamiento del proceso no cambien con el tiempo. Así pues, a continuación se introducen formalmente las definiciones de proceso estocástico estacionario de forma débil y de forma estricta, que se pueden consultar en el Capítulo 2 de Chan and Cryer (2008).

Definición 2.2. *Un proceso estocástico $\{Y_t\}_t$ se dice que es **estrictamente estacionario** si la distribución conjunta de $\{Y_{t_1}, \ldots, Y_{t_n}\}$ coincide con la distribución conjunta de $\{Y_{t_1-k}, \ldots, Y_{t_n-k}\}$, para cualquier elección de $t_1, \ldots, t_n$ y cualquier retardo k.*

Al ser esta condición difícil de verificar empíricamente, surge la necesidad de asumir una versión menos estricta.

Definición 2.3. *Se dice que un proceso estocástico $\{Y_t\}_t$ es **débilmente estacionario**, o **estacionario de segundo orden** si*

1. $\mathbb{E}(Y_t) = \mu < \infty$ *para todo* $t \in \mathcal{I}$, *y*
2. $\mathbb{E}((Y_t - \mu)(Y_{t-k} - \mu)) = \gamma_k$ *para todo* $t \in \mathcal{I}$ *y para cualquier retardo* k.

Por tanto, un proceso estocástico es débilmente estacionario si sus dos primeros momentos son finitos e invariantes en el tiempo. Esto se traduce en que la media y varianza del mismo han de ser constantes y que la relación lineal entre observaciones tomadas en distintos instantes de tiempo tan solo dependerá de la distancia entre dichos instantes.

De estas definiciones se puede deducir que si un proceso es estrictamente estacionario y los momentos de primer y segundo orden son finitos, entonces también es débilmente estacionario; pero el recíproco no es cierto. Ambas condiciones sí son equivalentes en el caso de que la serie siga una distribución normal (Tsay, 2005).

Mencionar que, en lo que sigue, a lo largo de este trabajo se utilizará el término *serie estacionaria* o *proceso estacionario* para referirse a esta definición más débil de estacionariedad.

A partir de estas definiciones, cabe introducir un ejemplo muy importante de proceso estacionario, el ruido blanco, el cual se utilizará en la construcción del resto de procesos y modelos que se estudien.

Definición 2.4. *Un proceso de **ruido blanco** es una secuencia $\{a_t\}_t$ de variables aleatorias incorreladas, con media nula y varianza finita,*

$$\mathbb{E}(a_t) = 0, \quad \mathbb{E}\left(a_t^2\right) = \sigma^2 \quad y \quad \mathbb{E}(a_t \varepsilon_\tau) = 0, \quad \text{para } t \neq \tau.$$

En el caso de que además el proceso siga una distribución normal, se dirá que el proceso es de ruido blanco gaussiano y las variables alatorias que lo conforman serán independientes e idénticamente distribuidas.

Por último, cabe explicar distintas formas de representar un proceso estocástico.

Definición 2.5. *Se dirá que un proceso estocástico $\{X_t\}_t$ es **lineal** si admite una representación de la forma*

$$X_t = \mu + \sum_{i=-\infty}^{+\infty} \psi_i a_{t-i}, \ \ con \ \sum_{i=-\infty}^{+\infty} |\psi_i| < \infty$$

y donde $\{a_t\}_t$ es un proceso de ruido blanco.

Es importante notar que todo proceso lineal será estacionario, ya que no es más que una "combinación lineal" de un proceso de ruido blanco, el cual es estacionario.

Definición 2.6. *Un proceso estocástico $\{X_t\}_t$ será **causal** (o $MA(\infty)$) si se puede escribir como*

$$X_t = c + \psi_0 a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \ldots, \ \ con \ \sum_{i=0}^{\infty} |\psi_i| < \infty,$$

siendo c una constante y $\{a_t\}_t$ un proceso de ruido blanco.

Definición 2.7. *Un proceso estocástico $\{X_t\}_t$ será **invertible** (o $AR(\infty)$) si admite una representación de la forma*

$$X_t = c + a_t + \pi_1 X_{t-1} + \pi_2 X_{t-2} + \ldots, \ \ con \ \sum_{i=1}^{\infty} |\pi_i| < \infty,$$

donde c es una constante y $\{a_t\}_t$ un proceso de ruido blanco.

En este contexto, el Teorema de descomposición de Wold (Wold, 1938) garantiza que cualquier proceso estocástico estacionario sin componentes deterministas puede escribirse como un proceso lineal. Así, cualquier proceso estacionario o bien es lineal o bien puede ser transformado para serlo, luego la clase de procesos lineales constituye un marco general para el estudio de este tipo de procesos. En efecto, la mayoría de modelos que se estudiarán en este capítulo no serán más que casos particulares de esta representación.

2.2. Modelos Box-Jenkins

Una vez introducidas las series de tiempo, así como algunas posibles representaciones de las mismas, se pretenden estudiar modelos estocásticos que hayan podido generar cierta serie de tiempo, con el fin de entender la dinámica de la misma y predecir valores futuros. Para ello, se recurre a la conocida como metodología Box-Jenkins (Box et al., 2015), consistente en un proceso en tres etapas: identificación y selección del modelo generador de la serie, estimación del modelo seleccionado y diagnosis o validación de las hipótesis del mismo.

Comenzamos estudiando los principales posibles modelos para series estacionarias.

2.2.1. Series estacionarias

Definición 2.8. *Un **modelo autorregresivo** de orden p, $AR(p)$, viene dado por*

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + a_t, \tag{2.2}$$

siendo $\{X_t\}_t$ un proceso estocástico, $c, \phi_1, \phi_2, \ldots, \phi_p$ constantes con $\phi_p \neq 0$ y $\{a_t\}_t$ un proceso de ruido blanco.

Observación 2.9. *Este modelo se puede reescribir de una manera más compacta sin más que utilizar el operador retardo B tal que*

$$BX_t = X_{t-1},$$

como

$$\phi_p(B)X_t = \left(1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p\right) X_t = a_t.$$

Es importante tener en cuenta que la representación dada por (2.2) da lugar a un proceso estacionario si y solo si las raíces del polinomio característico $\phi(z) = 1 - \phi_1 z - \phi_2 z^2 - \ldots - \phi_p z^p$ tienen módulo distinto de uno. Además, atendiendo a las definiciones 2.6 y 2.7, un proceso de este tipo siempre será invertible y en el caso de que todas las raíces del polinomio $\phi(z)$ tengan módulo mayor que uno, el proceso será causal. Para más detalles acerca de esto se puede consultar el Capítulo 3 de Shumway and Stoffer (2000).

Definición 2.10. *Un **modelo de medias móviles** de orden q, $MA(q)$, viene dado por*

$$X_t = c + a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \ldots + \theta_q a_{t-q}, \tag{2.3}$$

siendo $\{X_t\}_t$ un proceso estocástico, $c, \theta_1, \theta_2, \ldots, \theta_q$ constantes con $\theta_q \neq 0$ y $\{a_t\}_t$ un proceso de ruido blanco.

Observación 2.11. *De nuevo, la ecuación (2.3) se puede reescribir recurriendo al operador retardo B como*

$$X_t = c + \theta_q(B)a_t = c + (1 + \theta_1 B + \theta_2 B^2 + \ldots + \theta_q B^q)a_t.$$

En este caso, un proceso $MA(q)$ siempre será estacionario y causal, mientras que será invertible si y solo si el polinomio característico $\theta_q(z) = 1 + \theta_1 z + \theta_2 z^2 + \ldots + \theta_q z^q$ no tiene raíces cuyo módulo sea superior a uno.

También existe la posibilidad de que en un mismo proceso haya una estructura autorregresiva (AR) y otra de medias móviles (MA), surgiendo así los modelos $ARMA$, que se describen a continuación.

Definición 2.12. *Un **modelo ARMA** de órdenes p y q, $ARMA(p, q)$, viene dado por*

$$\begin{aligned} X_t = c &+ \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} \\ &+ a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \ldots + \theta_q a_{t-q}, \end{aligned} \tag{2.4}$$

siendo $\{X_t\}_t$ un proceso estocástico, $c, \phi_1, \phi_2, \ldots, \phi_p, \theta_1, \theta_2, \ldots, \theta_q$ constantes con $\phi_p \neq 0$, $\theta_q \neq 0$ y $\{a_t\}_t$ un proceso de ruido blanco.

Observación 2.13. *Igual que antes, se puede reescribir la ecuación (2.4) del modelo de manera más compacta,*

$$\phi_p(B)X_t = c + \theta_q(B)a_t.$$

Ahora bien, para que la representación anterior de lugar a un proceso estacionario, el polinomio característico asociado a la parte autorregresiva, $\phi_p(z)$ no puede tener ninguna raíz con módulo uno. De igual manera, un proceso $ARMA(p, q)$ será causal siempre y cuando las raíces de dicho polinomio tengan módulo estrictamente mayor que uno y será invertible si el polinomio $\theta_q(z)$ no tiene ninguna raíz con módulo mayor que uno.

Los modelos vistos hasta ahora sirven para modelizar la **dependencia regular** de una serie, esto es, la dependencia de una observación X_t con las p (o q) observaciones (o innovaciones) inmediatamente anteriores. Sin embargo, puede haber casos en los que dicha dependencia ocurra entre observaciones o innovaciones separadas por múltiplos de cierto período estacional s. Esto es lo que se conoce como **dependencia estacional**. En este contexto surge un nuevo tipo de modelos.

Definición 2.14. *Un **modelo ARMA estacional** de órdenes P y Q y con período estacional s, $ARMA(P,Q)_s$, viene dado por*

$$\begin{aligned} X_t = c &+ \Phi_1 X_{t-s} + \Phi_2 X_{t-2s} + \ldots + \Phi_P X_{t-Ps} \\ &+ a_t + \Theta_1 a_{t-s} + \Theta_2 a_{t-2s} + \ldots + \Theta_Q a_{t-Qs}, \end{aligned} \tag{2.5}$$

siendo $\{X_t\}_t$ un proceso estocástico, $c, \Phi_1, \Phi_2, \ldots, \Phi_P, \Theta_1, \Theta_2, \ldots, \Theta_Q$ constantes con $\Phi_p \neq 0$, $\Theta_q \neq 0$ y $\{a_t\}_t$ un proceso de ruido blanco.

Observación 2.15. *Es importante notar los siguientes aspectos acerca de estos modelos:*

- *La ecuación (2.5) se puede reescribir como*

$$\Phi_P(B^s) X_t = c + \Theta_Q(B^s) a_t,$$

siendo B^s es el operador retardo estacional tal que

$$B^s X_t = X_{t-s}$$

y donde

$$\begin{aligned} \Phi_P(B^s) &= 1 - \Phi_1 B - \Phi_2 B^{2s} - \ldots - \Phi_P B^{Ps}, \\ \Theta_Q(B^s) &= 1 + \Theta_1 a_{t-s} + \Theta_2 a_{t-2s} + \ldots + \Theta_Q a_{t-Qs}. \end{aligned}$$

- *Un modelo $ARMA(P,Q)_s$ no es más que un modelo $ARMA(sP, sQ)$ con muchos de sus coeficientes nulos, luego las propiedades de estos no diferirán de las de los procesos $ARMA$.*

Además, es posible que en una serie haya dependencia regular y estacional al mismo tiempo. Para modelizar esta última clase de procesos, surgen los $ARMA$ estacionales multiplicativos.

Definición 2.16. *Un **modelo ARMA estacional multiplicativo** de órdenes p, q, P y Q y período estacional s, $ARMA(p,q) \times (P,Q)_s$, se puede escribir como*

$$\Phi_P(B^s)\phi_p(B) X_t = c + \Theta_Q(B^s)\theta_q(B) a_t,$$

siendo $\{X_t\}_t$ un proceso estocástico, c una constante, $\Phi_P(B^s)$ y $\phi_p(B)$ los correspondientes polinomios autorregresivos, $\Theta_Q(B^s)$ y $\theta_q(B)$ los de medias móviles y $\{a_t\}_t$ un proceso de ruido blanco.

2.2.2. Series no estacionarias

Hasta ahora se han presentado modelos partiendo del supuesto de que la serie temporal es estacionaria, sin embargo, son muchos los casos, y en especial en el ámbito económico, en los que las series de tiempo no son constantes en media o en varianza. Ante esta situación, antes de poder ajustar un modelo $ARMA$ será necesario transformar la serie con el fin de alcanzar esa estacionariedad.

Un posible motivo para que una serie no sea estacionaria puede ser la presencia de cierta **tendencia** en el nivel de la serie, de forma que exista una relación entre el nivel medio de la misma y el tiempo. Esto se puede resolver diferenciando la serie de forma regular.

Definición 2.17. *Un proceso estocástico $\{X_t\}_t$ no estacionario y con tendencia se dice que es **integrado** de orden $d > 0$, $\{X_t\}_t \sim I(d)$, si $(1-B)^d X_t$ es estacionario pero $(1-B)^{d-1} X_t$ no lo es.*

A partir de este concepto surge un nuevo tipo de modelos, los $ARIMA$, que consisten en aplicar d diferencias regulares a la serie de tiempo y después ajustar un modelo $ARMA$.

Definición 2.18. *Un **modelo ARIMA** de órdenes p, d y q, $ARIMA(p,d,q)$, se puede escribir como*

$$\phi_p(B)(1-B)^d X_t = c + \theta_q(B)a_t,$$

donde $\{X_t\}_t$ es un proceso estocástico, c es una constante, $\phi_p(B)$ es el correspondiente polinomio autorregresivo, $\theta_q(B)$ el polinomio de medias móviles y $\{a_t\}_t$ un proceso de ruido blanco.

Observación 2.19. *Dada una serie temporal no estacionaria con tendencia, usualmente es suficiente con $d \leq 3$, es decir, tres o menos diferencias regulares, para corregir la ausencia de estacionariedad debida a tendencia en la serie.*

Además de tendencia, es posible que una serie no sea estacionaria debido a la presencia de una **componente estacional**. Cuando en una serie hay dependencia estacional y el nivel medio de la serie depende de dicho período estacional, de forma paralela al caso de la tendencia, será necesario diferenciar la serie, pero en este caso de forma estacional.

Definición 2.20. *Un proceso estocástico $\{X_t\}_t$ no estacionario y con dependencia estacional de período s se dice que es **integrado** de orden $D > 0$, $\{X_t\}_t \sim I(D)$, si $(1-B^s)^D X_t$ es estacionario pero $(1-B^s)^{D-1}X_t$ no lo es.*

Es así como surgen los modelos $ARIMA$ estacionales, resultantes de aplicar D diferencias estacionales y ajustar un modelo $ARMA$ estacional.

Definición 2.21. *Un **modelo ARIMA estacional** de órdenes P, D y Q y con período estacional s, $ARMA(P,D,Q)_s$, se puede escribir como*

$$\Phi_P(B^s)(1-B^s)^D X_t = c + \Theta_Q(B^s)a_t,$$

siendo $\{X_t\}_t$ un proceso estocástico, c una constante, $\Phi_P(B^s)$ el correspondiente polinomio autorregresivo, $\Theta_Q(B^s)$ el polinomio de medias móviles y $\{a_t\}_t$ un proceso de ruido blanco.

Juntando las ideas presentadas hasta ahora se llega a los modelos $ARIMA$ estacionales multiplicativos, a partir de los cuáles podremos ajustar la mayoría de las series de tiempo y en los que se recogen todos los modelos anteriormente expuestos como casos particulares, sin más que fijar alguno de los parámetros a cero.

Definición 2.22. *Un **modelo ARIMA estacional multiplicativo** de órdenes p, d, q, P, D y Q y período estacional s, $ARMA(p,d,q)\times(P,D,Q)_s$, se puede escribir como*

$$\Phi_P(B^s)\phi_p(B)(1-B^s)^D(1-B)^d X_t = c + \Theta_Q(B^s)\theta_q(B)a_t,$$

siendo $\{X_t\}_t$ un proceso estocástico, c una constante, $\Phi_P(B^s)$ y $\phi_p(B)$ los correspondientes polinomios autorregresivos, $\Theta_Q(B^s)$ y $\theta_q(B)$ los de medias móviles y $\{a_t\}_t$ un proceso de ruido blanco.

Por último, otro motivo por el que una serie puede no ser estacionaria es la **heterocedasticidad** de la misma, que se da cuando la variabilidad de la serie cambia con el tiempo. En la mayoría de casos, la heterocedasticidad puede solucionarse aplicando una transformación de Box-Cox[1] (Shumway and Stoffer, 2000) sobre los datos

$$Y_t = \begin{cases} \frac{X_t^\lambda - 1}{\lambda}, & \text{para } \lambda \neq 0, \\ \log X_t, & \text{para } \lambda = 0, \end{cases} \tag{2.6}$$

que conseguirá estabilizar la varianza de la serie.

[1] En el caso de las serie económicas es habitual emplear el logaritmo.

2.3. Modelos de regresión con series temporales

Hay muchas ocasiones en las que resultará interesante no solo explicar el comportamiento de una serie a partir del histórico de la misma, sino también a partir de otra serie de tiempo con la que esté relacionada. En este punto juegan un papel importante los conocidos como **modelos de regresión dinámica**.

Sea $\{Y_t\}_t$ el proceso generador de la serie de tiempo que queremos modelizar, es decir, de nuestra variable respuesta, y sea $\{X_t\}_t$ el proceso generador de otra serie de tiempo que tomará el papel de variable explicativa. Entonces, se puede considerar el modelo de regresión lineal

$$Y_t = \beta_0 + \beta_1 X_{t-r} + \varepsilon_t;$$

de forma que la primera de las series en un instante t se relaciona con la segunda en el instante $t-r$, para cierto entero r, y donde β_0 y β_1 son constantes y $\{\varepsilon_t\}_t$ el proceso estocástico correspondiente con la serie de los errores.

En este contexto, habrá que tener en cuenta:

1. Los procesos $\{X_t\}_t$ e $\{Y_t\}_t$ deben ser conjuntamente estacionarias.
2. Será necesario identificar el retardo r con el que se relacionan las series.
3. Los errores del modelo $\{\varepsilon_t\}_t$ puede que no sean independientes.

2.3.1. Relación lineal entre series de tiempo

Para medir la relación lineal entre dos series de tiempo y seleccionar el retardo r con el que estas se relacionan, es necesario introducir algunas definiciones previas.

Definición 2.23. *Sean $\{X_t\}_t$ e $\{Y_t\}_t$ dos procesos estocásticos:*

- *Se define la función de **covarianzas cruzadas** como*

$$\gamma_{s,t}(X,Y) = Cov(X_s, Y_t).$$

- *La función de **correlaciones cruzadas** viene dada por*

$$\rho_{s,t} = \frac{\gamma_{s,t}(X,Y)}{\sigma_{X_s}\sigma_{Y_t}},$$

donde σ_{X_s} es la desviación típica de la variable X_s y σ_{Y_t}, la de Y_t.

Definición 2.24. *Dos procesos estocásticos $\{X_t\}_t$ e $\{Y_t\}_t$ se dice que son **conjuntamente estacionarios** cuando:*

- *Ambos procesos son estacionarios.*
- *Las covarianzas cruzadas dependen únicamente del retardo entre las variables,*

$$\gamma_k(X,Y) = \gamma_{t,t-k}(X,Y) = \gamma_{s,s-k}(X,Y), \text{ para todo } t,s,k.$$

Para analizar entonces si hay correlación entre dos series de tiempo y determinar el retardo que las relaciona se analizará la función de correlaciones cruzadas, la cuál se desconoce. En su lugar, se recurrirá a la función de correlaciones cruzadas muestral,

$$\widehat{\rho}_k(X,Y) = \frac{\widehat{\gamma}_k(X,Y)}{\widehat{\sigma}_{X_t}\widehat{\sigma}_{Y_{t-k}}} = \frac{\sum \left(X_t - \bar{X}\right)\left(Y_{t-k} - \bar{Y}\right)}{\sqrt{\sum \left(X_t - \bar{X}\right)^2}\sqrt{\sum \left(Y_{t-k} - \bar{Y}\right)^2}},$$

que bajo la condición de que ambos procesos sean ruido blanco o uno sea ruido blanco y el otro estacionario y sean independientes, sigue una distribución $N(0, T^{-1})$, donde T es el tamaño muestral (Shumway and Stoffer, 2000, p. 491).

Así, bajo dichas condiciones, se considerará que hay relación lineal entre las series si existe algún retardo k tal que

$$|\widehat{\rho}_k(X,Y)| \geq 1.96/\sqrt{T};$$

y en ese caso dichos valores de k serían los candidatos para el retardo r que relaciona las series.

Sin embargo, en pocos casos nos encontramos con tal situación. En su defecto, normalmente se trabaja con series de tiempo que no son de ruido blanco e incluso ni siquiera suelen ser estacionarias, en cuyo caso para determinar si están relacionadas y con qué retardo, será necesario aplicar una transformación sobre las series, denominada **preblanqueado**.

El proceso sería el siguiente:

1. Transformar $\{X_t\}_t$ mediante el operador lineal $\pi(B) = 1 - \pi_1 B - \pi_2 B^2 - \ldots$ a un proceso de ruido blanco $\{\tilde{X}_t\}_t = \{\pi(B)X_t\}_t$.
2. Aplicar la misma transformación sobre el proceso $\{Y_t\}_t$ obteniendo $\{\tilde{Y}_t\}_t = \{\pi(B)Y_t\}_t$.
3. Asumir que $\{\tilde{Y}_t\}_t$ es estacionario y estudiar si existe algún valor de k tal que

   $$|\widehat{\rho}_k(\tilde{X},\tilde{Y})| \geq 1.96/\sqrt{T},$$

   y en caso de que sí, proponerlo como el retardo del modelo.

Observación 2.25. *Nótese que:*

- *En el caso de que los procesos $\{X_t\}_t$ o $\{Y_t\}_t$ no sean estacionarios, deben diferenciarse primero y sobre los procesos transformados estacionarios aplicar el preblanqueado.*
- *Como $\pi(B)$ es un operador lineal, la relación lineal entre los procesos transformados $\{\tilde{X}_t\}_t$ e $\{\tilde{Y}_t\}_t$ se mantendrá para los originales.*

2.3.2. Modelo para los errores

Una vez detectada la existencia de relación lineal entre las series y el correspondiente retardo, como se está trabajando con series de tiempo, cabe la posibilidad de que los errores del modelo $\{\varepsilon_t\}_t$ estén correlados. Es por esto que será necesario proponer un modelo para los errores.

Como estos son desconocidos, se analizarán los residuos del modelo y se propondrá un modelo adecuado para los mismos, en el marco de los modelos estudiados en la Sección 2.2.

En cuanto a la selección, estimación, diagnosis y predicción de estos modelos de regresión, será análoga a la de los modelos antes expuestos, con la salvedad de que será necesario estimar también los coeficientes β_0 y β_1 del modelo lineal. Estos temas se tratarán en la siguiente sección.

2.4. Selección, estimación, diagnosis y predicción del modelo

A lo largo de este capítulo se han introducido diferentes modelos para series temporales. Ahora bien, llegados a este punto, será importante estudiar cómo identificar el proceso generador de una serie de tiempo dada.

2.4.1. Selección del modelo generador de la serie

En relación con lo que se ha comentado anteriormente, antes de nada será necesario analizar la serie de tiempo de manera exploratoria. Para ello se puede recurrir, por ejemplo, al gráfico secuencial de la serie o a los gráficos de autoccorrelaciones simples (ACF) y parciales (PACF). Para más detalles sobre las funciones de autocorrelación simple y parcial se puede consultar Chan and Cryer (2008).

Así, en primer lugar, cabe prestar atención a la estacionalidad. La estacionalidad en series de tiempo se refiere a una variación periódica y predecible en un período igual o inferior a un año, que puede deberse a las estaciones del año, los períodos vacacionales o de rebajas, los factores climáticos, etc. Si una serie presenta estacionalidad, se observará un patrón repetitivo dentro del ciclo estacional, que puede complicar la modelización de la misma.

La estacionalidad puede detectarse a partir de diferentes tests, como son el QS-Test (Maravall, 2011), el test de Friedman (1937), el de Kruskal and Wallis (1952) o el de Welch (1951). Y, en el caso de que haya, es recomendable corregirla mediante alguna transformación. Existen distintas posibilidades para ello, pero no se desarrollarán con profundidad al no ser el principal objetivo de este proyecto. Pueden consultarse algunas de dichas metodologías en Dagum and Bianconcini (2016).

A continuación, si la variabilidad de la serie no es constante, se aplicará una transformación de la forma (2.6) con el fin de estabilizar la varianza. Una vez hecho esto, si la serie presenta tendencia (el gráfico de autocorrelaciones simples presenta valores positivos altos que tardan en caer) se aplican d diferencias regulares. Y, eliminada la tendencia, si la serie contiene componente estacional (se observan repuntes positivos y que tardan en bajar en los retardos múltiplos del período estacional s del gráfico de autocorrelaciones simples), esta se elimina mediante D diferencias estacionales de período s.

En el caso de que se tengan dudas acerca de la estacionariedad de la serie, se podría realizar un test de estacionariedad. Se propone el contraste de Dickey-Fuller, por ser el más habitual. En Said and Dickey (1984) se pueden consultar todos los detalles acerca de este contraste.

Con las series ya transformadas, hay que analizar métodos para seleccionar los órdenes p, q, P y Q. Hay dos posibilidades, por un lado un método gráfico consistente en observar las funciones de autocorrelación simples y parciales muestrales y, por otro, la elección de dichos órdenes de acuerdo a la minimización de algún criterio de información.

En el primero de los casos, no siempre es posible identificar el proceso generador de la serie. Además, también es fácil que este método nos lleve a identificar varios procesos como generadores de la serie o bien, a seleccionar uno que no es el mejor en términos de error. No se incluyen más detalles puesto que este no será utilizado en la práctica, pero para conocer las características que dichas funciones de autocorrelación muestral deben cumplir para seleccionar cierto proceso puede recurrise al Capítulo 6 de Box et al. (2015).

En su defecto, se seleccionará el modelo que minimice el valor de algún criterio de información como el AIC propuesto por Akaike (1974), el BIC de Schwarz (1978) o el $AICc$ desarrollado por Hurvich and Tsai (1989). Estos criterios consisten en alcanzar un equilibrio entre un buen ajuste y el número de parámetros del modelo, que interesa que sea lo menor posible.

2.4.2. Estimación del modelo seleccionado

El siguiente paso una vez identificado el proceso generador de la serie será estimar los parámetros implicados en el mismo.

Por simplicidad supondremos el caso de un modelo $ARMA$[2] y estudiaremos dos métodos: la estimación por mínimos cuadrados y por máxima verosimilitud.

Los parámetros a estimar serán los coeficientes del modelo $ARMA(p,q)$, c, ϕ_1, ..., ϕ_p, θ_1, ..., θ_q, así como la varianza de las innovaciones, σ_a^2. En lo que sigue, se denotarán los parámetros estimados como

$$\hat{\beta} = \left(\hat{c}, \hat{\phi}_1, \ldots, \hat{\phi}_p, \hat{\theta}_1, \ldots, \hat{\theta}_q, \hat{\sigma}_a^2\right);$$

y será importante tener presentes los residuos del modelo asociados a esta estimación,

$$\hat{a}_t = X_t - \left(\hat{c} + \hat{\phi}_1 X_{t-1} + \ldots + \hat{\phi}_p X_{t-p} + \hat{\theta}_1 \hat{a}_{t-1} + \ldots + \hat{\theta}_q \hat{a}_{t-q}\right), \text{ con } t = 1, \ldots, T.$$

Estimación por mínimos cuadrados y mínimos cuadrados condicionados

La estimación de los parámetros por el método de mínimos cuadrados se obtiene a partir de los valores de $\hat{\beta}$ que minimizan la suma residual de cuadrados,

$$\hat{\beta} = \arg\min_{\hat{\beta}} \sum_{t=1}^{T} \hat{a}_t^2.$$

En este punto pueden surgir dos complicaciones:

- Si $p > 0$, los valores de los residuos $\hat{a}_1, \ldots, \hat{a}_p$ dependen de los valores de $X_0, \ldots, X_{1-p}$, que se desconocen.
- Si además $q > 0$, $\hat{a}_{p+1}$ depende de los valores de $\hat{a}_p$, $\hat{a}_{p-1}$, ..., $\hat{a}_{p+1-q}$, que dependen de valores no observados de la serie. En este caso se fijan dichos valores y se construyen $\hat{a}_{p+1}, \ldots, \hat{a}_T$ iterativamente.

Ambos problemas se pueden resolver recurriendo al método de mínimos cuadrados condicionados,

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

Estimación por máxima verosimilitud

Para estimar los parámetros por máxima verosimilitud se seleccionan aquellos valores de $\hat{\beta}$ que maximizan la función de verosimilitud,

$$\hat{\beta} = \arg\max_{\hat{\beta}} f_{\hat{\beta}}\left(X_1, \ldots, X_T\right),$$

[2] Nótese que dada una serie de tiempo no estacionaria, si se diferencia regularmente d veces y estacionalmente (con período s) D veces, se transforma en una serie estacionaria que podrá ser modelizada a través de un proceso $ARMA$, luego el supuesto no será restrictivo.

siendo $f_{\hat{\beta}}$ la función de densidad conjunta asociada a un vector aleatorio $\left(\hat{X}_1, \ldots, \hat{X}_T\right)'$ procedente de un proceso $ARMA(p,q)$ con coeficientes $\hat{\beta}$.

Es importante tener en cuenta que, como ya se adelantaba en la Sección 2.3, en el caso de los modelos de regresión dinámicos, además de la estimación de los parámetros del modelo de los errores, será necesario estimar los parámetros del modelo lineal, β_0 y β_1. Para ello, se pueden utilizar las técnicas de mínimos cuadrados generalizados (GLS) o máxima verosimilitud (ML). Como referencia, se puede consultar Hamilton (2020) para ampliar esta información.

2.4.3. Diagnosis

Una vez ajustado el modelo, la siguiente etapa es validar que las hipótesis realizadas sobre el mismo se cumplen. Si al llegar a este punto alguna de las hipótesis falla, el modelo seleccionado no será válido y habrá que seleccionar otro modelo y repetir el proceso hasta conseguir un modelo que verifique todas las hipótesis.

La hipótesis principal es que las innovaciones deben ser ruido blanco, es decir:

1. $\mathbb{E}(a_t) = 0$ (media cero).
2. $Var(a_t) = \sigma_a^2$ (varianza constante).
3. $Cov(a_s, a_t) = 0$ para todo $s \neq t$ (incorreladas).

Adicionalmente, se comprueba que las innovaciones sean gaussianas, no siendo esta condición excluyente pero sí conveniente.

Nótese que, como no se pueden observar directamente las innovaciones del modelo, las tareas de validación se realizarán sobre los residuos del mismo.

Contraste de incorrelación

En primer lugar, para comprobar que los residuos están incorrelados se puede o bien analizar la función de autocorrelación muestral o bien utilizar el contraste de independencia de Ljung-Box propuesto por Ljung and Box (1978). No obstante, la segunda de las técnicas propuestas es sin duda más potente: la primera es conservadora, no tiende a rechazar la independencia, ya que analiza cada correlación de manera individual, mientras que el contraste de Ljung-Box permite comprobar si las primeras h correlaciones se anulan (en conjunto).

Contraste de media nula

Una vez comprobado que los residuos son independientes, se contrasta si la media de los mismos es cero. Para ello, se recurre al test *t de Student*, que parte de la hipótesis de que los residuos provienen de variables aleatorias independientes e idéndicamente distribuídas.

Contraste de normalidad

Por último, para contrastar la normalidad de los residuos, se proponen dos posibilidades: el test de Shapiro-Wilk (Shapiro and Wilk, 1965) y el de Jarque-Bera (Jarque and Bera, 1987).

2.4.4. Predicción

Una vez identificado el modelo y verificado que se cumplen las hipótesis básicas, ya se está en condiciones de predecir el comportamiento futuro de la serie temporal, lo cual era el principal objetivo de este capítulo.

Para ello, igual que en el caso anterior nos limitaremos a un proceso $ARMA(p,q)$ y supondremos también que las innovaciones son gaussianas.

Dado un serie de tiempo generada por el proceso $\{X_t\}_t$, observada hasta el instante T y un horizonte de predicción h, el objetivo será predecir el valor de X_{T+h}. Se considerará el predictor que minimice el error cuadrático medio de predicción,

$$\hat{X}_{T+h} = \mathbb{E}\left(X_{T+h}|X_1,\ldots,X_T\right).$$

De acuerdo con la ecuación de un modelo $ARMA(p,q)$, se tendrá

$$\hat{X}_{T+h} = \hat{c}+\hat{\phi}_1\hat{X}_{T+h-1}+\ldots+\hat{\phi}_p\hat{X}_{T+h-p}+\hat{\theta}_1\mathbb{E}\left(a_{T+h-1}|X_1,\ldots,X_T\right)+\ldots+\hat{\theta}_q\mathbb{E}\left(a_{T+h-q}|X_1,\ldots,X_T\right),$$

donde

$$\mathbb{E}\left(a_{T+j}|X_1,\ldots,X_T\right) = \begin{cases} 0, & \text{si } j>0, \\ a_{T+j}, & \text{si } j\leq 0. \end{cases}$$

Nótese que $\hat{X}_{T+j} = X_{T+j}$ en el caso de que $j \leq 0$. Además, es importante darse cuenta que para dicha predicción hacen falta los valores de $a_{T+h},\ldots,a_{T+h-q}$. Por un lado, los valores de $a_{T+1},\ldots,a_{T+h}$ pueden estimarse por la media del proceso $\{a_t\}_t$, esto es, por cero. Y, por otro, los valores de $a_1,\ldots,a_T$ habrá que calcularlos. Por ejemplo, en el caso de un proceso invertible $X_t = c+a_t+\pi_1X_{t-1}+\pi_2X_{t-2}+\ldots$, se pueden calcular como una combinación lineal

$$a_t = -c + X_t - \pi_1X_{t-1} - \pi_2X_{t-2} - \ldots$$

Para un desarrollo más detallado de los métodos de predicción puntual, así como del cálculo de intervalos de predicción, que no se han incluido en este capítulo, se puede consultar el Capítulo 9 de Chan and Cryer (2008).

Capítulo 3

Análisis multivariante de series temporales

Una vez vistos modelos univariantes para explicar y predecir series de tiempo, en este capítulo se verá un enfoque multivariante o vectorial. La idea será introducir modelos estadísticos adecuados que permitan ajustar de forma simultánea las distintas series de tiempo que se consideren para el estudio, de forma que cada serie dependa tanto de sus propios retardos anteriores como de los valores rezagados de las demás variables.

La ventaja de este enfoque frente a los modelos univariantes vistos en el capítulo anterior es que se tienen en cuenta las posibles relaciones entre las series de tiempo a la hora de ajustar los correspondientes modelos, facilitando la detección de la dinámica de las series al retroalimentarse entre ellas. No obstante, precisan de la estimación de muchos más parámetros que los modelos univariantes, lo que puede resultar en una desventaja frente a estos.

En este contexto, se considerará un conjunto $\boldsymbol{y}$ de K series temporales, $\boldsymbol{y} = \{\boldsymbol{y}_t\}_t = \{y_{1t}, \ldots, y_{Kt}\}_t$, denominado serie de tiempo múltiple, y se analizarán las K variables conjuntamente, con la posibilidad de introducir al mismo tiempo variables exógenas, al igual que en el caso de los modelos univariantes. Para ello, se estudiarán los modelos vectoriales, entre los que destacan los procesos autorregresivos vectoriales, VAR, que se revisarán en la Sección 3.1; y los modelos de corrección de errores vectoriales, VECM, que se consideran en la Sección 3.2. En ambos casos se revisará la definición y especificación del modelo y la estimación de los respectivos coeficientes, para después pasar a estudiar algunos contrastes que se deben realizar para validar cada modelo, así como las correspondientes técnicas de predicción.

Como principales puntos de referencia se consideran la primera y segunda parte del trabajo de Lütkepohl (2005), el Capítulo 3 de Lütkepohl and Krätzig (2004) y Pfaff (2008a).

3.1. Modelo VAR

Se comenzará esta sección introduciendo los modelos VAR, para posteriormente examinar en detalle sus especificaciones, hipótesis y el procedimiento de estimación asociado.

Los procesos autorregresivos vectoriales (VAR) son una generalización de los procesos autorregresivos univariantes (AR) vistos en el Capítulo 2, diseñados para abordar múltiples series temporales. Esta adaptación permite que el modelo capture las interacciones dinámicas entre las diferentes series

temporales que constituyen la denominada serie de tiempo múltiple. A continuación, se define formalmente un modelo VAR con un cierto orden p, para después pasar a revisar las especificaciones y estimaciones del modelo resultante.

Definición 3.1. *Dado un conjunto de K variables endógenas $\boldsymbol{y}_t = (y_{1t}, \ldots, y_{Kt})'$, se define el modelo autorregresivo vectorial de orden p, $VAR(p)$, como*

$$\boldsymbol{y}_t = \boldsymbol{A}_1 \boldsymbol{y}_{t-1} + \ldots + \boldsymbol{A}_p \boldsymbol{y}_{t-p} + \boldsymbol{C}\boldsymbol{D}_t + \boldsymbol{u}_t, \tag{3.1}$$

donde $\boldsymbol{A}_i \in \mathcal{M}_{K\times K}$, con $i = 1, \ldots, p$, son las matrices de coeficientes del modelo de dimensión $K \times K$, de la forma

$$\boldsymbol{A}_i = \begin{pmatrix} \alpha_{11,i} & \ldots & \alpha_{1K,i} \\ \vdots & \ddots & \vdots \\ \alpha_{K1,i} & \ldots & \alpha_{KK,i} \end{pmatrix}$$

y donde $\boldsymbol{u}_t = (u_{1t}, \ldots, u_{Kt})' \in \mathbb{R}^K$ es un proceso de ruido blanco $K-$dimensional con vector de medias cero y matriz de varianzas covarianzas invariante definida positiva $\mathbb{E}(\boldsymbol{u}_t \boldsymbol{u}_t') = \boldsymbol{\Sigma}_{\boldsymbol{u}}$. La matriz $\boldsymbol{C} \in \mathcal{M}_{K\times M}$ es la matriz de coeficientes asociada a las potenciales variables regresoras deterministas, $\boldsymbol{D}_t \in \mathcal{M}_{M\times 1}$, entre las que se pueden incluir una constante, la tendencia o la estacionalidad de la serie.

Es fácil ver que la expresión (3.1) se puede reescribir en función del polinomio autorregresivo matricial $\boldsymbol{A}_p(B) = \boldsymbol{I}_K - \boldsymbol{A}_1 B - \ldots - \boldsymbol{A}_p B^p$ como

$$\boldsymbol{A}_p(B)\boldsymbol{y}_t = C\boldsymbol{D}_t + \boldsymbol{u}_t,$$

siendo B el operador retardo tal que $B^i \boldsymbol{y}_t = \boldsymbol{y}_{t-i}$ para cierto $i \in \mathbb{Z}$.

En base a dicho polinomio, se puede estudiar una característica importante de los modelos $VAR(p)$, que es la **estabilidad**. Esta se refiere a la propiedad de que estos modelos generan series de tiempo estacionarias con medias y matriz de varianzas covarianzas invariantes en el tiempo, dados valores iniciales suficientes. Para comprobar si un modelo $VAR(p)$ verifica la condición de estabilidad, una vez ajustado el modelo, se puede evaluar el polinomio autorregresivo. Así, se tendrá que el proceso es estable sí y solo sí el polinomio $\boldsymbol{A}_p(\boldsymbol{z})$ no tienen ninguna raíz en el círculo unidad complejo, esto es, si

$$\det\left(\boldsymbol{A}_p(\boldsymbol{z})\right) = \det\left(\boldsymbol{I}_K - \boldsymbol{A}_1 \boldsymbol{z} - \ldots - \boldsymbol{A}_{\boldsymbol{p}} \boldsymbol{z}^p\right) \neq 0, \text{ para todo } |\boldsymbol{z}| < 1. \tag{3.2}$$

En caso contrario, se podría decir que alguna de las variables es integrada.

Además, por otra parte, si la solución a la ecuación (3.2) tiene una raíz unitaria, entonces o bien alguna o todas las variables son integradas de orden 1, es decir, $I(1)$. Esto ocurre cuando alguna variable es resultado de una combinación lineal de otras variables implicadas en el modelo, en cuyo caso se dirá que dichas variables están **cointegradas**, problema que se estudiará más adelante en la Sección 3.2 con la introducción de los modelos VECM.

En la práctica, para evaluar la estabilidad de un modelo $VAR(p)$ puede ser útil reescribirlo de una forma más sencilla. Un modelo $VAR(p)$ puede ser visto como un modelo $VAR(1)$ dado por la expresión

$$\boldsymbol{\xi}_t = \boldsymbol{A}\boldsymbol{\xi}_{t-1} + \boldsymbol{v}_t, \tag{3.3}$$

donde

$$\boldsymbol{\xi}_t = \begin{pmatrix} \boldsymbol{y}_t \\ \vdots \\ \boldsymbol{y}_{t-p+1} \end{pmatrix} \in \mathbb{R}^{K_p}, \quad \boldsymbol{A} = \begin{pmatrix} \boldsymbol{A}_1 & \boldsymbol{A}_2 & \dots & \boldsymbol{A}_{p-1} & \boldsymbol{A}_p \\ \boldsymbol{I} & 0 & \dots & 0 & 0 \\ 0 & \boldsymbol{I} & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & \boldsymbol{I} & 0 \end{pmatrix} \in \mathcal{M}_{K_p \times K_p}, \quad \boldsymbol{v}_t = \begin{pmatrix} \boldsymbol{u}_t \\ 0 \\ \vdots \\ 0 \end{pmatrix} \in \mathbb{R}^{K_p}.$$

A partir de esta nueva expresión del modelo, se pueden calcular los autovalores asociados a la matriz $\boldsymbol{A}$, λ_i, con $i = 1, \dots, K_p$. Si el módulo de los autovalores es menor que uno, es decir, si $|\lambda_i| < 1$, para todo $i = 1, \dots, K_p$, entonces el modelo (3.3) es estable y, en consecuencia, el proceso $VAR(p)$ también lo es.

Es importante tener en cuenta que los procesos $VAR(p)$ estables son estacionarios, pues la estabilidad implica estacionariedad. Sin embargo, que un proceso $VAR(p)$ no sea estable no implica necesariamente que no sea estacionario (Lütkepohl, 2005).

3.1.1. Estimación de los coeficientes

Una vez definido un modelo VAR de cierto orden p, estudiaremos cómo se estiman los parámetros implicados en el mismo.

Por simplicidad, se considera un modelo VAR de orden p sin la componente determinista, es decir,

$$\boldsymbol{y}_t = \boldsymbol{A}_1 \boldsymbol{y}_{t-1} + \dots + \boldsymbol{A}_p \boldsymbol{y}_{t-p} + \boldsymbol{u}_t. \tag{3.4}$$

Las K ecuaciones de este se estiman individualmente por *mínimos cuadrados* (OLS), pues el estimador resultante tiene la misma eficiencia que el estimador por *mínimos cuadrados generalizados* (GLS) (Zellner, 1962).

Utilizando la notación dada por $\boldsymbol{Y} = (\boldsymbol{y}_1, \dots, \boldsymbol{y}_T) \in \mathcal{M}_{K \times T}$, $\boldsymbol{A} = (\boldsymbol{A}_1 : \dots : \boldsymbol{A}_p) \in \mathcal{M}_{K \times K_p}$, $\boldsymbol{U} = (\boldsymbol{u}_1, \dots, \boldsymbol{u}_T) \in \mathcal{M}_{K \times T}$ y $\boldsymbol{Z} = (\boldsymbol{Z}_0, \dots, \boldsymbol{Z}_{T-1}) \in \mathcal{M}_{K \times T-1}$, donde $\boldsymbol{Z}_{t-1} = (\boldsymbol{y}_{t-1}, \dots, \boldsymbol{y}_{t-p})'$, se puede reescribir el modelo (3.4) matricialmente,

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{Z} + \boldsymbol{U}.$$

De este modo, el estimador de $\boldsymbol{A}$ por mínimos cuadrados resulta

$$\hat{\boldsymbol{A}} = \left(\hat{\boldsymbol{A}}_1 : \dots : \hat{\boldsymbol{A}}_p \right) = \boldsymbol{Y}\boldsymbol{Z}'(\boldsymbol{Z}\boldsymbol{Z}')^{-1}; \tag{3.5}$$

y es consistente y tiene distribución asintótica normal (Lütkepohl, 2005),

$$\hat{\boldsymbol{A}} \overset{a}{\sim} N\left(\boldsymbol{A}, T^{-1}\boldsymbol{\Sigma}_{\hat{\boldsymbol{A}}}\right).$$

La matriz de varianzas-covarianzas de la distribución asintótica verifica que

$$\boldsymbol{\Sigma}_{\hat{\boldsymbol{A}}} \xrightarrow{p} \left(\boldsymbol{Z}\boldsymbol{Z}'/T\right)^{-1} \otimes \boldsymbol{\Sigma}_u,$$

donde $\xrightarrow{p}$ denota convergencia en probabilidad, luego resulta que el estimador por *OLS* verifica

$$\hat{\boldsymbol{A}} \overset{a}{\sim} N\left(\boldsymbol{A}, (\boldsymbol{Z}\boldsymbol{Z}')^{-1} \otimes \boldsymbol{\Sigma}_u\right).$$

Observación 3.2. *Es interesante tener en cuenta los siguientes aspectos sobre el estimador (3.5), que se pueden encontrar en Lütkepohl and Krätzig (2004):*

- *Este coincide con el estimador de máxima verosimilitud si cada serie implicada en el modelo, $\boldsymbol{y}_t$, es $I(0)$ y se tiene normalidad. Además, en dicho caso la matriz $\boldsymbol{\Sigma}_{\hat{\boldsymbol{A}}}$ es no singular.*
- *En caso contrario, es decir, si existe alguna variable integrada, $I(1)$, la matriz $\boldsymbol{\Sigma}_{\hat{\boldsymbol{A}}}$ es singular. A pesar de ello, el estimador sigue siendo válido y se sigue verificando la normalidad asintótica. No obstante, en este caso algunos estimadores convergen con mayor rapidez que $T^{1/2}$, lo que provoca que los usuales $t-$, χ^2- y $F-$test sobre los parámetros del modelo podrían no ser válidos. Esto se relaciona con el concepto de cointegración que se mencionaba antes y que se tratará en la siguiente sección, puesto que es el fenómemo que motiva los modelos VECM.*

Por último, una vez estimados los coeficientes del modelo, cabe atender a la distribución de los residuos del mismo. De acuerdo a la definición del modelo VAR, los residuos $\boldsymbol{u}_t$ tienen media 0 y matriz de varianzas-covarianzas asociada $\boldsymbol{\Sigma}_u$. Para ajustar dicha matriz, se proponen estimadores usuales como son

$$\hat{\boldsymbol{\Sigma}}_u = \tfrac{1}{T-K_p}\textstyle\sum_{t=1}^{T} \hat{\boldsymbol{u}}_t\hat{\boldsymbol{u}}_t' \quad \text{y} \quad \tilde{\boldsymbol{\Sigma}}_u = \tfrac{1}{T}\textstyle\sum_{t=1}^{T} \hat{\boldsymbol{u}}_t\hat{\boldsymbol{u}}_t',$$

siendo $\hat{\boldsymbol{u}}_t = \boldsymbol{y}_t - \hat{\boldsymbol{A}}\boldsymbol{Z}_{t-1}$ los residuos de la estimación por mínimos cuadrados. Ambos estimadores propuestos son consistentes y asintóticamente normal distribuidos, así como independientes de $\hat{\boldsymbol{A}}$.

3.1.2. Especificación del modelo

Ahora que ya se ha definido el modelo VAR y se han estimado sus coeficientes, se debe estudiar cómo determinar el orden p más adecuado para cada caso, esto es, especificar el modelo.

Por un lado, una posible aproximación sería comenzar con un modelo con algún orden máximo preespecificado y aplicar tests para determinar el orden adecuado. En particular, podría contrastarse sucesivamente la hipótesis nula $H_0 : \boldsymbol{A}_{p_{max}-i} = 0$ para $i = 0, 1, \ldots, p_{max}$ hasta que se rechace el test. En este caso, es importante escoger un valor adecuado para dicho valor máximo de p, ya que si se considera un valor demasiado pequeño puede dar lugar a problemas en la diagnosis del modelo, mientras que si se considera uno demasiado grande puede aumentar el error de Tipo I.

Por otro lado, se puede seleccionar el orden p de forma que minimice algún criterio de información. En general, se basará en la expresión

$$Cr(m) = \log\left(\det\left(\tilde{\boldsymbol{\Sigma}}_u(m)\right)\right) + c_T\phi(m),$$

donde $m = 0, \ldots, p_{max}$ denota el orden del modelo VAR, $\tilde{\boldsymbol{\Sigma}}_u$ es el estimador de la matriz de varianzas-covarianzas de los residuos para un modelo $VAR(m)$, c_T es una constante que depende del tamaño muestral T y $\Phi(m)$ es una función que penaliza valores muy grandes de m y dependerá de cada criterio.

Los criterios más habituales son el Criterio de Información de Akaike propuesto por Akaike (1974, 1998), el de Hamilton-Quinn recogido en Hannan and Quinn (1979) y el de Schwarz formulado por Schwarz (1978), que vienen dados por

$$AIC(m) = \log\left(\det\left(\tilde{\boldsymbol{\Sigma}}_u(m)\right)\right) + \frac{2}{T}mK^2,$$

$$HQ(m) = \log\left(\det\left(\tilde{\boldsymbol{\Sigma}}_u(m)\right)\right) + \frac{2\log\left(\log(T)\right)}{T} mK^2$$

y

$$SC(m) = \log\left(\det\left(\tilde{\boldsymbol{\Sigma}}_u(m)\right)\right) + \frac{\log(T)}{T} mK^2,$$

respectivamente.

El AIC sobreestima el orden asintóticamente, mientras que los otros dos criterios estiman el orden de forma consistente bajo condiciones bastante generales si el verdadero orden del modelo VAR es finito y menor que el orden p_{max} fijado (Lütkepohl and Krätzig, 2004). Así, denotando por $\hat{p}(AIC)$, $\hat{p}(HQ)$ y $\hat{p}(SC)$ los órdenes óptimos según cada uno de los criterios, para muestras de tamaño superior a 16 ($T \geq 16$), se verifica que

$$\hat{p}(SC) \leq \hat{p}(HQ) \leq \hat{p}(AIC).$$

Nótese que estos resultados son válidos no solo para procesos $I(0)$, sino también para procesos que incluyan variables cointegradas.

3.1.3. Diagnosis

La estimación y especificación del modelo VAR se ha hecho en el supuesto de que ciertas hipótesis sean ciertas. Es por esto que, como siguiente paso, surge la necesidad de estudiar cuáles son dichas hipótesis y cómo comprobar que efectivamente se verifican.

Dichas hipótesis se basarán en los residuos del modelo final, que deberán ser heterocedásticos, estar incorrelados e idealmente, normalmente distribuidos. Además, se comprobará la hipótesis de estabilidad estructural del ajuste y se analizará la causalidad de cada una de las variables implicadas en el modelo.

Incorrelación de los residuos

Para comprobar la ausencia de correlación en los residuos $\hat{\boldsymbol{u}} = (\hat{\boldsymbol{u}}_{1t}, \ldots, \hat{\boldsymbol{u}}_{Kt})$ de un modelo $VAR(p)$ se plantearán dos posibles tests.

En primer lugar, el test de *Portmanteau* (Hosking, 1980), que sirve para contrastar la hipótesis nula

$$H_0 : \mathbb{E}\left(\boldsymbol{u}_t \boldsymbol{u}'_{t-i}\right) = 0, \text{ con } i = 1, \ldots, h > p,$$

frente a la hipótesis alternativa de que al menos una autocorrelación sea no nula. El correspondiente estadístico de contraste es de la forma

$$Q_h = T \sum_{j=1}^{h} tr\left(\hat{\boldsymbol{C}}'_j \hat{\boldsymbol{C}}_0^{-1} \hat{\boldsymbol{C}}_j \hat{\boldsymbol{C}}_0^{-1}\right), \tag{3.6}$$

siendo $\hat{\boldsymbol{C}}_i = \frac{1}{T}\sum_{t=i+1}^{T} \hat{\boldsymbol{u}}_t \hat{\boldsymbol{u}}'_{t-i}$.

Este estadístico de distribuye aproximadamente de acuerdo con una distribución $\chi^2_{K^2h-n^*}$, estando así los grados de libertad determinados por la diferencia entre las correlaciones incluidas, K^2h, y el número de coeficientes estimados en el modelo $VAR(p)$ excluyendo las componentes deterministas, n^*.

Esta distribución límite es válida cuando $h \to \infty$ y se tiene un tamaño muestral suficientemente grande. No obstante, se puede considerar el estadístico de *Pormanteau* ajustado, que se asemeja al estadístico de Ljung-Box para series univariantes y consiste en una modificación de (3.6) con propiedades

potencialmente mejores para muestras más pequeñas,

$$Q_h^* = T^2 \sum_{j=1}^{h} \frac{1}{T-j} tr\left(\hat{\boldsymbol{C}}_j' \hat{\boldsymbol{C}}_0^{-1} \hat{\boldsymbol{C}}_j \hat{\boldsymbol{C}}_0^{-1}\right).$$

Es importante notar que en el test de *Portmanteau* está implicado un parámetro $h > p$. La elección de este será crucial para el resultado del mismo, luego será necesario aplicar el test para diferentes valores de este parámetro. Si se escoge un valor de h demasiado pequeño, la aproximación a la distribución χ^2 puede ser débil, así mismo, un valor muy grande puede resultar en una pérdida de poder en el contraste.

Por otra parte, el test de *Breusch-Godfrey* (Godfrey, 1988) se basa en las regresiones auxiliares

$$\hat{\boldsymbol{u}}_t = \boldsymbol{A}_1 \boldsymbol{y}_{t-1} + \ldots + \boldsymbol{A}_p \boldsymbol{y}_{t-p} + \boldsymbol{C}\boldsymbol{D}_t + \boldsymbol{B}_1 \hat{\boldsymbol{u}}_{t-1} + \ldots + \boldsymbol{B}_h \hat{\boldsymbol{u}}_{t-h} + \boldsymbol{\varepsilon}_t \tag{3.7}$$

y contrasta la hipótesis nula

$$H_0 : \ \boldsymbol{B}_1 = \ldots = \boldsymbol{B}_h = 0$$

frente a la hipótesis alternativa de que algún $\boldsymbol{B}_i$ sea distinto de 0.

El estadístico, que se basa en multiplicadores de Lagrange, viene dado por

$$LM_h = T\left(K - tr(\tilde{\boldsymbol{\Sigma}}_R^{-1} \tilde{\boldsymbol{\Sigma}}_\varepsilon)\right),$$

siendo $\tilde{\boldsymbol{\Sigma}}_\varepsilon = \frac{1}{T}\sum_{t=1}^{T} \hat{\boldsymbol{\varepsilon}}_t \hat{\boldsymbol{\varepsilon}}_t'$ y $\tilde{\boldsymbol{\Sigma}}_R = \frac{1}{T}\sum_{t=1}^{T} \hat{\boldsymbol{\varepsilon}}_t^R \hat{\boldsymbol{\varepsilon}}_t^{R\prime}$, y donde $\hat{\boldsymbol{\varepsilon}}_t^R$ hacen referencia a los residuos del modelo (3.7) bajo la restricción de que se cumpla la hipótesis nula. Este estadístico tiene distribución asintótica $\chi^2_{hK^2}$ y, en analogía con el test de *Portmanteau*, se verifica para tamaños muestrales grandes. En este caso se tiene la siguiente corrección, válida para muestras pequeñas,

$$LMF_h = \frac{1 - \left(1 - R_r^2\right)^{1/r}}{\left(1 - R_r^2\right)^{1/r}} \frac{N_r - q}{Km},$$

con

$$R_r^2 = 1 - |\tilde{\boldsymbol{\Sigma}}_\varepsilon| / |\tilde{\boldsymbol{\Sigma}}_R|, \quad r = \sqrt{\frac{K^2 m^2 - 1}{K^2 + m^2 - 5}}, \quad q = \tfrac{1}{2} Km - 1, \quad N = T - K - m - \tfrac{1}{2}(K - m + 1),$$

siendo n el número de regresoras que conforman el modelo y $m = Kh$. Este nuevo estadístico se distribuye atendiendo a una $F_{hK^2,[Nr-q]}$.

En cuanto a la elección entre cada uno de los estadísticos propuestos, de acuerdo con Lütkepohl and Krätzig (2004) el test de *Breusch-Godfrey* es preferible ante órdenes bajos de autocorrelaciones de los residuos (h pequeño), mientras que el test de *Portmanteau* funciona mejor para valores de h más grandes.

Heterocedasticidad condicional

La heterocedasticidad condicional de los residuos puede contrastarse mediante tests *ARCH* (Modelos autorregresivos con heterocedasticidad condicional) univariantes y multivariantes. En este caso, se recurrirá al test *LM-ARCH* multivariante (Doornik and Hendry, 1997) basado en multiplicadores de Lagrange, que se basa en la regresión

$$vech\left(\hat{\boldsymbol{u}}_t \hat{\boldsymbol{u}}_t'\right) = \boldsymbol{\beta}_0 + \boldsymbol{B}_1 vech\left(\hat{\boldsymbol{u}}_{t-1} \hat{\boldsymbol{u}}_{t-1}'\right) + \ldots + \boldsymbol{B}_q vech\left(\hat{\boldsymbol{u}}_{t-q} \hat{\boldsymbol{u}}_{t-q}'\right) + \boldsymbol{v}_t \tag{3.8}$$

donde $\boldsymbol{\beta}_0 \in \mathbb{R}^{\frac{1}{2}K(K+1)}$ y $\boldsymbol{B}_i \in \mathcal{M}_{\frac{1}{2}K(K+1)\times\frac{1}{2}K(K+1)}$ con $i = 1, \ldots, q$ son matrices de coeficientes, $\boldsymbol{v}_t$ es un vector cuyos elementos son procesos de ruido blanco y $vech()$ es un operador conocido como *half-vectorization*, que consiste en

$$vech(\boldsymbol{A}) = (A_{11}, \ldots, A_{n1}, A_{22}, \ldots, A_{n2}, \ldots, A_{n-1\ n-1}, A_{n\ n-1}, A_{nn}),$$

para cualquier matriz $\boldsymbol{A}$ simétrica.

El test consiste en contrastar la hipótesis nula $H_0 : \boldsymbol{B}_i = 0$ para todo $i = 1, \ldots, q$, frente a $H_a : \exists i = 1, \ldots, q$ tal que $\boldsymbol{B}_i \neq 0$, mediante el estadístico de contraste

$$VARCH_{LM}(q) = \frac{1}{2}TK(K+1)R_m^2 \sim \chi^2_{qK^2(K+1)^2/4},$$

siendo

$$R_m^2 = 1 - \frac{2}{K(K+1)}tr(\hat{\boldsymbol{\Omega}}\hat{\boldsymbol{\Omega}}_0^{-1})$$

y donde $\hat{\boldsymbol{\Omega}} \in \mathcal{M}_{\frac{1}{2}K(K+1)}$ es la matriz de covarianzas de la regresión (3.8) y $\hat{\boldsymbol{\Omega}}_0$ la correspondiente matriz con $q = 0$.

Normalidad

La normalidad de los residuos del modelo se puede contrastar tanto de forma univariante como multivariante, recurriendo en ambos casos al test de *Jarque-Bera* (Jarque and Bera, 1987).

Por un lado, se pueden emplear test univariantes sobre los residuos de las expresiones de cada serie para contrastar la normalidad de forma individual; y por otro, se puede aplicar una versión multivariante de este test sobre los residuos estandarizados $\hat{\boldsymbol{u}}_t^s = \tilde{\boldsymbol{P}}^{-1}\left(\hat{\boldsymbol{u}}_t - \bar{\hat{\boldsymbol{u}}}_t\right)$, resultantes de aplicar una descomposición de Cholesky a la matriz de varianzas-covarianzas $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}} = \tilde{\boldsymbol{P}}\tilde{\boldsymbol{P}}'$ de los residuos centrados, siendo $\tilde{\boldsymbol{P}}$ una matriz diagonal inferior con diagonal positiva.

Siguiendo esta idea, el test de *Jarque-Bera* multivariante sería

$$JB_{mv} = s_3^2 + s_4^2 \sim \chi^2_{2K},$$

donde

$$s_3^2 = T\boldsymbol{b}_1'\boldsymbol{b}_1/6 \sim \chi^2_K$$
$$s_4^2 = T(\boldsymbol{b}_2 - \boldsymbol{3}_K)'(\boldsymbol{b}_2 - \boldsymbol{3}_K)/24 \sim \chi^2_K$$

y $\boldsymbol{b}_1$ y $\boldsymbol{b}_2$ son los momentos de tercer y cuarto orden de los residuos $\hat{\boldsymbol{u}}_t^s = (\hat{\boldsymbol{u}}_{1t}^s, \ldots, \hat{\boldsymbol{u}}_{Kt}^s)'$, es decir,

$$\boldsymbol{b}_1 = (b_{11}, \ldots, b_{1K})', \text{ con } b_{1K} = \frac{1}{T}\sum_{t=1}^{T}(\hat{\boldsymbol{u}}_{kt}^s)^3,$$
$$\boldsymbol{b}_2 = (b_{21}, \ldots, b_{2K})', \text{ con } b_{2K} = \frac{1}{T}\sum_{t=1}^{T}(\hat{\boldsymbol{u}}_{kt}^s)^4.$$

Nótese que s_3^2 y s_4^2 son estadísticos de la asimetría y la curtosis, respectivamente.

Estabilidad estructural

De forma adicional a las hipótesis anteriores sobre los residuos, puede resultar de interés analizar la estabilidad estructural del modelo. Para ello, se puede recurrir a los test CUSUM (*cumulative sum*) o CUSUM-sq (*CUSUM squared*), así como a los test de Chow. Estos primeros analizan las sumas acumulativas de residuos con el fin de detectar cambios en los coeficientes estimados del modelo, mientras que los test de Chow comprueban si se ha producido un cambio en los parámetros en algún momento, comparando los parámetros estimados antes y después de una posible fecha de interrupción dada.

En la práctica, para estudiar la estabilidad estructural recurriremos a los test CUSUM, luego los detalles de los test de Chow no se incluyen. No obstante, para un desarrollo detallado de los mismos puede consultarse Chow (1960).

CUSUM es la suma acumulativa de residuos recursivos $\hat{\boldsymbol{u}}_t^{(r)}$ (Lütkepohl and Krätzig, 2004, p. 52),

$$CUSUM_\tau = \sum_{t=K+1}^{\tau} \hat{\boldsymbol{u}}_t^{(r)} / \hat{\sigma}_u, \tag{3.9}$$

y fue propuesta por Brown et al. (1975) con el objetivo de detectar cambios estructurales. En la práctica, se suele representar (3.9) para valores de $\tau = K+1, \ldots, T$ y si se aleja mucho de la línea del cero, se dice que hay evidencias en contra de la estabilidad estructural del modelo. En particular, se rechazará la estabilidad con un nivel de significación del 5 % si $CUSUM_\tau$ cruza las líneas $\pm 0.948 \left[\sqrt{T-K} + 2(\tau - K)/\sqrt{T-K}\right]$.

Nótese que este test está diseñado para detectar cambios estructurales debidos a cambios en un parámetro del modelo. No obstante, puede reducirse su poder si existen varios cambios que puedan compensar sus impactos en la media residual. En este caso, puede ser más apropiado el CUSUM-sq,

$$CUSUM - sq_\tau = \sum_{t=K+1}^{\tau} \left(\hat{\boldsymbol{u}}_t^{(r)}\right)^2 \Big/ \sum_{t=K+1}^{T} \left(\hat{\boldsymbol{u}}_t^{(r)}\right)^2, \tag{3.10}$$

que determina que un modelo no es estable si (3.10) supera los límites $\pm c + (\tau - K)\,/\,(T-K)$, siendo c una constante que depende del nivel de significación, el tamaño muestral T y el número de variables regresoras implicadas en el modelo.

Causalidad

La detección de causalidades entre variables es otro tema interesante a abordar en el ajuste de modelos $VAR(p)$. En este contexto, la definición de causalidad de *Granger* (Granger, 1969) ha tomado bastante importancia entre los modelos econométricos.

Si se considera la división del vector de variables endógenas $\boldsymbol{y}_t$ en dos subvectores, $\boldsymbol{y}_{1t} \sim \mathcal{M}_{K_1 \times 1}$ e $\boldsymbol{y}_{2t} \sim \mathcal{M}_{K_2 \times 1}$, con $K = K_1 + K_2$, se dice que la variable $\boldsymbol{y}_{2t}$ es causal de *Granger* de $\boldsymbol{y}_{1t}$, si la primera ayuda a predecir la segunda. Más formalmente:

Definición 3.3. *Sea $\boldsymbol{y}_{1,1+h|\Omega_t}$ la predicción a horizonte h de $\boldsymbol{y}_{1t}$ con origen t basado en el conjunto de información relevante del espacio Ω_t, entonces se dice que $\boldsymbol{y}_{2t}$ es no causal de Granger si y solo si*

$$\boldsymbol{y}_{1,1+h|\Omega_t} = \boldsymbol{y}_{1,1+h|\Omega_t \setminus \{\boldsymbol{y}_{2,s} | s \leq t\}}, \;\; h = 1, 2, \ldots$$

Esta definición se traduce en que $\boldsymbol{y}_{2t}$ no es causal de $\boldsymbol{y}_{1t}$ si eliminar del conjunto de información relevante el pasado de $\boldsymbol{y}_{2t}$ no cambia la predicción de $\boldsymbol{y}_{1t}$ a ningún horizonte h.

Para contrastar la causalidad, se considera el proceso autorregresivo de orden p reescrito como

$$\begin{pmatrix} \boldsymbol{y}_{1t} \\ \boldsymbol{y}_{2t} \end{pmatrix} = \sum_{i=1}^{p} \begin{pmatrix} \alpha_{11,i} & \alpha_{12,i} \\ \alpha_{21,i} & \alpha_{22,i} \end{pmatrix} \begin{pmatrix} \boldsymbol{y}_{1,t-i} \\ \boldsymbol{y}_{2,t-i} \end{pmatrix} + \boldsymbol{C}\boldsymbol{D}_t + \begin{pmatrix} \boldsymbol{u}_{1t} \\ \boldsymbol{u}_{2t} \end{pmatrix},$$

de forma que la Definición 3.3 equivale a $\alpha_{12,i} = 0$ para $i = 1, 2, \ldots, p$. Así, surge el contraste

$$\begin{aligned} H_0 &: \alpha_{12,i} = 0, \text{ para todo } i = 1, \ldots, p, \\ H_a &: \exists \alpha_{12,i} \neq 0 \text{ para } i = 1, \ldots, p, \end{aligned}$$

cuyo estadístico asociado sigue una distribución F de Snédecor con pK_1K_2 y $KT - n^*$ grados de libertad, siendo n^* el número total de parámetros en el proceso $VAR(p)$, incluyendo las componentes deterministas.

3.1.4. Predicción

En analogía al desarrollo del capítulo anterior, una vez especificado el modelo $VAR(p)$, estimados sus coeficientes y pasada la validación, se pasan a estudiar las correspondientes técnicas de predicción.

Dado cierto horizonte h, la predicción de la serie multivariante $\boldsymbol{y}_t$ modelizada mediante un $VAR(p)$ de la forma (3.1) se puede calcular recursivamente como

$$\boldsymbol{y}_{T+h} = \boldsymbol{A}_1\boldsymbol{y}_{T+h-1} + \ldots + \boldsymbol{A}_p\boldsymbol{y}_{T+h-p} + \boldsymbol{C}\boldsymbol{D}_{T+h},$$

para $h = 1, \ldots, n$.

Ahora bien, si los verdaderos coeficientes se sustituyen por sus estimaciones se tiene que

$$\hat{\boldsymbol{y}}_{T+h} = \hat{\boldsymbol{A}}_1\hat{\boldsymbol{y}}_{T+h-1} + \ldots + \hat{\boldsymbol{A}}_p\hat{\boldsymbol{y}}_{T+h-p} + \hat{\boldsymbol{C}}\hat{\boldsymbol{D}}_{T+h}.$$

Así, la matriz de covarianzas asociada al error de predicción resulta,

$$Cov\left(\begin{pmatrix} \boldsymbol{y}_{T+1} - \hat{\boldsymbol{y}}_{T+1} \\ \vdots \\ \boldsymbol{y}_{T+h} - \hat{\boldsymbol{y}}_{T+h} \end{pmatrix}\right) = \begin{pmatrix} \boldsymbol{I} & 0 & \ldots & 0 \\ \boldsymbol{\Phi}_1 & \boldsymbol{I} & \ldots & 0 \\ \vdots & \vdots & \ddots & 0 \\ \boldsymbol{\Phi}_{h-1} & \boldsymbol{\Phi}_{h-2} & \ldots & \boldsymbol{I} \end{pmatrix} \cdot \left(\boldsymbol{\Sigma}_{\boldsymbol{u}} \otimes \boldsymbol{I}_h\right) \cdot \begin{pmatrix} \boldsymbol{I} & 0 & \ldots & 0 \\ \boldsymbol{\Phi}_1 & \boldsymbol{I} & \ldots & 0 \\ \vdots & \vdots & \ddots & 0 \\ \boldsymbol{\Phi}_{h-1} & \boldsymbol{\Phi}_{h-2} & \ldots & \boldsymbol{I} \end{pmatrix}',$$

donde las matrices $\boldsymbol{\Phi}_i$ con $i = 1, \ldots, h-1$ son las matrices de coeficientes asociadas a la representación MA de un proceso $VAR(P)$ estable[1].

Análogamente al caso univariante, además de la predicción puntual, se pueden construir intervalos de predicción. Para más detalles, pueden consultarse la Sección 3.5 de Lütkepohl (2005) o la Sección 2.2 de Pfaff (2008a).

[1] Igual que un proceso estable $AR(p)$ se puede representar como un proceso $MA(\infty)$, tal y como se vio en el Capítulo 2, un proceso estable $VAR(p)$ se puede representar como un proceso de medias móviles,

$$\boldsymbol{y}_t = \boldsymbol{\Phi}_0\boldsymbol{u}_t + \boldsymbol{\Phi}_1\boldsymbol{u}_{t-1} + \boldsymbol{\Phi}_2\boldsymbol{u}_{t-2} + \ldots, \tag{3.11}$$

con $\boldsymbol{\Phi}_0 = \boldsymbol{I}_k$ y las matrices $\boldsymbol{\Phi}_s = \sum_{j=1}^{s} \boldsymbol{\Phi}_{s-j}\boldsymbol{A}_j$ para $s = 1, 2, \ldots$, siendo $\boldsymbol{A}_j = 0$ para $j > p$.

3.2. Modelo VECM

Como se adelantaba en la sección anterior, cuando las series de partida no son estacionarias (integradas de orden uno, $I(1)$), el modelo VAR resultante no es estable. Este es el caso en el que nos encontraremos en múltiples ocasiones al intentar modelizar series de tiempo económicas, cuyo comportamiento no puede ser capturado por procesos estacionarios.

En este punto, atendiendo a todos los conceptos introducidos anteriormente, se podría pensar en transformar las series de partida, diferenciándolas para que sean estacionarias como una posible solución al problema. No obstante, procediendo de esta forma podríamos estar perdiendo información relevante para la modelización, relacionada con las posibles dependencias a largo plazo entre las distintas series de tiempo. Es así como surge la necesidad de modelos que permitan ajustar series de tiempo no estacionarias.

En este contexto, surge el concepto de cointegración, introducido por Granger (1981) y Engle and Granger (1987).

Definición 3.4. *Dada una serie de tiempo multivariante $\boldsymbol{y}_t$ se dice que es **cointegrada** si*

1. *Alguna de sus componentes es integrada.*
2. *Existe al menos un vector $\boldsymbol{\beta} \neq \mathbf{0}$ tal que la serie $\boldsymbol{\beta}'\boldsymbol{y}_t$ sea estacionaria.*

En dicho caso, el vector $\boldsymbol{\beta}$ se denomina vector de cointegración.

*Además, si existen r vectores $\boldsymbol{\beta}_i$, con $i = 1, \ldots, r$ linealmente independientes, se dice que $\boldsymbol{y}_t$ es **cointegrada de rango r** y $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_r)$ es la matriz de vectores de cointegración.*

En otras palabras, si una serie de tiempo multivariante no es estacionaria (alguna de sus componentes o ninguna lo es), pero existe una combinación lineal de sus componentes que sí lo es, entonces se dice que la serie es cointegrada.

Así, para integrar en la modelización las relaciones de cointegración (o relaciones de largo plazo estacionarias) entre las series de tiempo, surgen los *Modelos de Corrección de Errores Vectoriales*, VECM. Estos establecen un marco único en el que se integran las dependencias temporales dinámicas capturadas por los modelos VAR, junto con las relaciones de cointegración.

Definición 3.5. *Dado un conjunto de K variables endógenas $\boldsymbol{y}_t = (y_{1t}, \ldots, y_{Kt})'$, se define el modelo de corrección de errores vectorial de orden $p-1$, $VECM(p-1)$, como*

$$\Delta \boldsymbol{y}_t = \boldsymbol{\Pi} \boldsymbol{y}_{t-1} + \boldsymbol{\Gamma}_1 \Delta \boldsymbol{y}_{t-1} + \ldots + \boldsymbol{\Gamma}_{p-1} \Delta \boldsymbol{y}_{t-p+1} + \boldsymbol{u}_t, \tag{3.12}$$

donde

$$\begin{aligned} \boldsymbol{\Pi} &= -\left(\boldsymbol{I}_K - \boldsymbol{A}_1 - \ldots - \boldsymbol{A}_p\right) \\ \boldsymbol{\Gamma}_i &= -\left(\boldsymbol{A}_{i+1} + \ldots + \boldsymbol{A}_p\right), \text{ para } i = 1, \ldots, p-1, \end{aligned}$$

y siendo $\Delta \boldsymbol{y}_t = \boldsymbol{y}_t - \boldsymbol{y}_{t-1}$ la primera diferencia regular de $\boldsymbol{y}_t$ y $\boldsymbol{u}_t$ un proceso de ruido blanco K-dimensional con vector de medias cero y matriz de varianzas covarianzas $\boldsymbol{\Sigma}_{\boldsymbol{u}}$.

El término $\boldsymbol{\Pi} \boldsymbol{y}_{t-1}$ del modelo incluye las relaciones de cointegración entre las variables y recoge los parámetros denominados *long-run* (relaciones de largo plazo). Asimismo, los parámetros $\boldsymbol{\Gamma}_i$, con $i = 1, \ldots, p-1$ se conocen como *short-run* (relaciones de corto plazo).

Observación 3.6. *La definición del modelo VECM se ha hecho para el orden $p-1$ debido a la relación directa entre un $VECM(p-1)$ y un $VAR(p)$. En efecto, las matrices implicadas en el modelo VECM se han definido a partir de las matrices de coeficientes del modelo VAR. Y, del mismo modo, se pueden determinar los coeficientes del VAR a partir de las matrices del modelo VECM. En particular,*

$$\boldsymbol{A}_i = \begin{cases} \boldsymbol{\Gamma}_1 + \boldsymbol{\Pi} + \boldsymbol{I}_k, & si\ i = 1, \\ \boldsymbol{\Gamma}_i - \boldsymbol{\Gamma}_{i-1}, & si\ i = 2, \dots, p-1, \\ -\boldsymbol{\Gamma}_{p-1}, & si\ i = p. \end{cases}$$

Es importante notar que el modelo VECM se obtiene a partir de los niveles del modelo VAR, al restar $\boldsymbol{y}_{t-1}$ en ambos lados de la ecuación del mismo. Así, todos los términos del modelo VECM darán lugar a procesos estacionarios, por ser la mayoría las diferencias de primer orden de las series originales y $\boldsymbol{\Pi}\boldsymbol{y}_{t-1}$ una combinación lineal estacionaria de las mismas. Ahora bien, cabe distinguir qué condiciones sobre la matriz $\boldsymbol{\Pi}$ son necesarias para que dicho término sea estacionario.

Como ya se adelantaba, el término $\boldsymbol{\Pi}\boldsymbol{y}_{t-1}$ recoge las relaciones de cointegración y el rango de la matriz $\boldsymbol{\Pi}$ coincidirá con el rango de cointegración. En este punto, se pueden diferenciar tres casos:

1. $r = rk(\boldsymbol{\Pi}) = 0$.

   Este primer caso se da cuando no existe ningún vector de cointegración, o lo que es lo mismo, no existe ninguna combinación lineal estacionaria distinta de $\boldsymbol{\Pi}\boldsymbol{y}_{t-1} = 0$. Así, este caso se correspondería con un modelo VAR estacionario en primeras diferencias.

2. $r = rk(\boldsymbol{\Pi}) = K$.

   Si el orden de cointegración es igual al número de variables endógenas, entonces quiere decir que todas las componentes de la serie $\boldsymbol{y}_t$ son estacionarias, en cuyo caso se tendría un modelo VAR en niveles de $\boldsymbol{y}_t$.

3. $0 < r = rk(\boldsymbol{\Pi}) < K$.

   Por último, si el rango de la matriz $\boldsymbol{\Pi}$ es inferior al número de variables endógenas y distinto de 0, entonces existirán dos matrices, $\boldsymbol{\alpha} \in \mathcal{M}_{K\times r}$ (matriz de cargas) y $\boldsymbol{\beta} \in \mathcal{M}_{K\times r}$ (matriz de cointegración), tales que $\boldsymbol{\Pi} = \boldsymbol{\alpha}\boldsymbol{\beta}'$. En este caso, se dirá que existen r relaciones de cointegración sobre el vector $\boldsymbol{y}_t$ linealmente independientes.

3.2.1. Estimación y especificación del modelo

En la literatura se han propuesto diversos métodos para la estimación de un modelo VECM, como pueden ser el de mínimos cuadrados ordinarios (OLS), el de mínimos cuadrados generalizados estimados ($EGLS$) o el de máxima verosimilitud (ML), entre otros. Una revisión de los mismos puede verse, por ejemplo, en el Capítulo 7 de Lütkepohl (2005). Sin embargo, en lo que sigue, por brevedad, se tratará únicamente uno de estos.

Para una mayor sencillez, consideremos el modelo VECM dado por (3.12) reescrito en forma matricial,

$$\Delta \boldsymbol{Y} = \boldsymbol{\Pi}\boldsymbol{Y}_{-1} + \boldsymbol{\Gamma}\Delta\boldsymbol{X} + \boldsymbol{U}, \tag{3.13}$$

donde

$$\begin{aligned}
&\Delta \boldsymbol{Y} = (\Delta \boldsymbol{y}_1, \ldots, \Delta \boldsymbol{y}_T)\,, \\
&\boldsymbol{Y}_{-1} = (\boldsymbol{y}_0, \ldots, \boldsymbol{y}_{T-1}) \\
&\boldsymbol{\Gamma} = (\boldsymbol{\Gamma}_1 : \ldots : \boldsymbol{\Gamma}_{p-1}) \\
&\Delta \boldsymbol{X} = (\Delta \boldsymbol{X}_0 : \ldots : \Delta \boldsymbol{X}_{T-1})\,, \text{ con } \Delta \boldsymbol{X}_{t-1} = (\Delta \boldsymbol{y}_{t-1}, \ldots, \Delta \boldsymbol{y}_{t-p+1})'\,, \\
&\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_t)\,.
\end{aligned}$$

Siguiendo Lütkepohl and Krätzig (2004), a partir de la expresión (3.13) y del método OLS obtenemos la estimación

$$\hat{\boldsymbol{\Gamma}} = (\Delta \boldsymbol{Y} - \boldsymbol{\Pi} \boldsymbol{Y}_{-1})\, \boldsymbol{X}' \left(\boldsymbol{X}\boldsymbol{X}'\right)^{-1}.$$

Y si la sustituimos en la ecuación (3.13), se tiene

$$\Delta \boldsymbol{Y}\boldsymbol{M} = \boldsymbol{\Pi}\boldsymbol{Y}_{-1}\boldsymbol{M} + \hat{\boldsymbol{U}}, \text{ con } \boldsymbol{M} = \boldsymbol{I} - \boldsymbol{X}' \left(\boldsymbol{X}\boldsymbol{X}'\right)^{-1} \boldsymbol{X}.$$

Pero aún falta estimar $\boldsymbol{\Pi} = \boldsymbol{\alpha}\boldsymbol{\beta}'$, para lo que se necesitará especificar el rango de cointegración, y la matriz de varianzas-covarianzas de los residuos del modelo, $\boldsymbol{\Sigma}_{\boldsymbol{u}}$.

En primer lugar, se definen las matrices

$$\boldsymbol{S}_{00} = \tfrac{1}{T}\Delta \boldsymbol{Y}\boldsymbol{M}\Delta \boldsymbol{Y}', \quad \boldsymbol{S}_{01} = \tfrac{1}{T}\Delta \boldsymbol{Y}\boldsymbol{M}\boldsymbol{Y}'_{-1}, \quad \boldsymbol{S}_{11} = \tfrac{1}{T}\boldsymbol{Y}_{-1}\boldsymbol{M}\boldsymbol{Y}'_{-1}.$$

Ahora, Johansen (1995) mostró que el estadístico de máxima verosimilitud para el contraste con hipótesis nula $H_0 : rk(\boldsymbol{\Pi}) \leq r$, es decir, que haya al menos r vectores de cointegración, viene dado por

$$-2ln\,(Q) = -T \sum_{i=r+1}^{K} ln(1 - \hat{\lambda}_i), \tag{3.14}$$

donde $\hat{\lambda}_i$ son los $K - r$ autovalores más pequeños de la ecuación

$$|\lambda \boldsymbol{S}_{11} - \boldsymbol{S}'_{01}\boldsymbol{S}_{00}^{-1}\boldsymbol{S}_{01}| = 0.$$

Luego, para hallar el orden de cointegración, habrá que repetir el contraste para todos los valores de $r \leq K$ hasta el primer rechazo de la hipótesis nula.

Una vez determinado el rango, los vectores de cointegración se estiman como

$$\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{v}}_1, \ldots, \hat{\boldsymbol{v}}_r)\,,$$

donde $\hat{\boldsymbol{v}}_i$ son los autovectores asociados a los r autovalores más grandes (los no empleados en el cálculo del estadístico (3.14)). Asimismo, la estimación de la matriz de cargas depende de la elección de la matriz de cointegración, y será

$$\hat{\boldsymbol{\alpha}} = \boldsymbol{S}_{01}\hat{\boldsymbol{\beta}},$$

resultando así que

$$\hat{\boldsymbol{\Pi}} = \boldsymbol{S}_{01}\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}'.$$

Por último, la estimación de la matriz de varianzas covarianzas de los errores vendrá dada por

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}} = \boldsymbol{S}_{00} - \hat{\boldsymbol{\alpha}}\hat{\boldsymbol{\alpha}}'.$$

3.2.2. Diagnosis y predicción

Ya especificado el modelo y estimados sus parámetros, como es habitual, cabe validar las hipótesis que deben cumplir los residuos del mismo y estudiar técnicas para obtener predicciones dado un cierto horizonte h.

En este caso, como un modelo VECM no es más que una transformación de un modelo VAR, las técnicas de validación de las hipótesis sobre los residuos de un modelo VAR expuestas en la Sección 3.1.3 serán generalizables al caso de los modelos VECM. Además, para predecir, bastará escribir el modelo VECM en términos de los coeficientes del modelo VAR y aplicar las técnicas de predicción correspondientes con dicho modelo. Es por esto que, conocidas las metodologías correspondientes con los modelos VAR, no será necesaria una revisión de las mismas para los VECM. No obstante, para un mayor detalle acerca de estos aspectos puede consultarse la Parte II de Lütkepohl (2005).

3.3. Funciones impulso-respuesta (IRF)

Al ajustar modelos multivariantes, puede resultar de interés cuantificar el impacto del impulso que una variable endógena produce sobre otra, entendiendo como impulso el efecto esperado de cierta variable con el cambio de otra. Para investigar este tipo de interaciones dinámicas entre variables endógenas surgen las **funciones de impulso respuesta**.

Estas se basan en la representación de medias móviles (3.11) del proceso. Los coeficientes de esta representación se pueden interpretar como las respuestas a impulsos: en particular, el elemento (i, j) de la matriz $\mathbf{\Phi}_s$ representa la respuesta esperada de la variable $y_{i,t+s}$ a un cambio de una unidad en la variable $y_{j,t}$.

Por un lado, en el caso de que el proceso $\boldsymbol{y}_t$ sea estacionario, se verifica que $\mathbf{\Phi}_s \to 0$ a medida que $s \to \infty$, entonces el efecto será transitorio, ya que se desvanece con el tiempo. En consecuencia, también pueden resultar de interés los efectos acumulados del impulso, que se obtienen sumando las matrices $\mathbf{\Phi}_s$. Así, por ejemplo,

$$\mathbf{\Phi} = \sum_{s=0}^{\infty} \mathbf{\Phi}_s, \tag{3.15}$$

representa el efecto acumulado de todos los períodos.

Observación 3.7. *Esta matriz dada por* (3.15) *existe si el proceso es estable (Lütkepohl and Krätzig, 2004).*

Por otro lado, en el caso de el proceso $\boldsymbol{y}_t$ sea no estacionario, las matrices $\mathbf{\Phi}_s$ pueden no converger a cero a medida que $s \to \infty$. Por ende, algunos impactos pueden tener efectos permanentes.

Una carencia de estas funciones es que no se pueden utilizar para evaluar las reacciones contemporáneas entre variables, ya que las matrices de coeficientes $\boldsymbol{A}_j$ que intervienen en el cáclulo de las $\mathbf{\Phi}_s$ no contienen información sobre las relaciones contemporáneas. Así, como alternativa a estas funciones, se emplean las **funciones de impulso respuesta ortogonales**. La idea consiste en descomponer la matriz de varianzas covarianzas del error mediante una descomposición de Cholesky, de forma que $\mathbf{\Sigma}_{\boldsymbol{u}} = \boldsymbol{P}\boldsymbol{P}'$, siendo $\boldsymbol{P}$ una matriz triangular inferior positiva. De este modo, la representación MA se transforma a

$$\boldsymbol{y}_t = \mathbf{\Psi}_0 \boldsymbol{\varepsilon}_t + \mathbf{\Psi}_1 \boldsymbol{\varepsilon}_{t-1} + \dots,$$

donde $\boldsymbol{\varepsilon} = \boldsymbol{P}^{-1}\boldsymbol{u}_t$ y $\mathbf{\Psi}_i = \mathbf{\Phi}_i \boldsymbol{P}$, para $i = 1, 2, \dots$ y $\mathbf{\Psi}_0 = \boldsymbol{P}$.

Observación 3.8. *Nótese que como el resultado de la descomposición de Cholesky es una matriz triangular inferior, la primera variable nunca será sensible a un impacto contemporáneo de ninguna otra variable, mientras que la última será sensible a los impactos de todas las demás variables. De este modo, los resultados pueden ser sensibles a la ordenación de las variables en el proceso* $\boldsymbol{y}_t$.

3.4. Descomposición de la varianza del error de predicción (FEVD)

Otra herramienta útil en el contexto de los modelos VAR y VECM es la **descomposición de la varianza del error de predicción** (FEVD), que permite analizar la contribución de una variable a la varianza del error de la predicción a horizonte h de otra variable del modelo.

La FEVD se basa en los coeficientes de las matrices de impulso-respuesta ortogonales, $\boldsymbol{\Psi}_i$. De acuerdo con Pfaff (2008a), la varianza del error de predicción en términos de dichas matrices viene dada por

$$\sigma_k(h)^2 = \sum_{n=0}^{h-1} \left(\boldsymbol{\Psi}_{k1,n}^2 + \ldots + \boldsymbol{\Psi}_{kK,n}^2 \right),$$

o equivalentemente por,

$$\sigma_k(h)^2 = \sum_{j=1}^{K} \left(\boldsymbol{\Psi}_{kj,0}^2 + \ldots + \boldsymbol{\Psi}_{kj,h-1}^2 \right).$$

Y, en consecuencia, se puede obtener la descomposición de la varianza del error de predicción en términos porcentuales como

$$\omega_{kj}(h) = \left(\boldsymbol{\Psi}_{kj,0}^2 + \ldots + \boldsymbol{\Psi}_{kj,h-1}^2 \right) / \sigma_k(h)^2.$$

Así, $\omega_{kj}(h)$ representa el porcentaje de la varianza del error de predicción a horizonte h de la variable y_k que se debe a y_j.

Capítulo 4

Redes neuronales *LSTM*

Finalmente, una vez estudiados los modelos "clásicos" en el contexto de las series temporales, en este capítulo se recurrirá a otra metodología diferente, como es la de las redes neuronales, que se podrían encuadrar en el marco de modelos de *deep learning*.

En particular, se abordarán las redes neuronales de memoria a corto y largo plazo, *LSTM* (*Long Short-Term Memory*), un tipo de redes neuronales recurrentes muy utilizadas en el ámbito de los datos secuenciales como las series de tiempo. Así, primero se introducirán las redes neuronales artificiales y las redes neuronales recurrentes en la Sección 4.1, para más tarde centrarse en este tipo particular en la Sección 4.2. Por último, en la Sección 4.3 se abordará el problema de la selección de hiperparámetros implicados en las redes.

Las principales referencias consideradas para el desarrollo de este capítulo son los apuntes de Aprendizaje Estadístico de Fernández-Casal et al. (2021), el Capítulo 11 de Hastie et al. (2009), el Capítulo 7 de Torres (2020) y los artículos de Graves (2013) y de Pascanu et al. (2013).

4.1. Conceptos previos sobre redes neuronales

Las redes neuronales o redes neuronales artificiales, inicialmente estudiadas por McCulloch and Pitts (1990), son un modelo matemático computacional inspirado en el cerebro humano que se ha vuelto muy popular en los últimos tiempos por su utilidad para abordar problemas con estructuras subyacentes muy complejas.

Consisten en métodos de aprendizaje, que entrenan los datos y aprenden de si mismos. Se componen por capas de nodos, también conocidas como **neuronas**, cuya estructura se acoge a la representación recogida en la Figura 4.1. Generalmente, están formadas por varias capas: una capa de entrada (o *input layer*) que consiste en las variables originales que entran en el modelo, una o varias capas ocultas (o *hidden layers*) y una capa de salida (o *output layer*), que contiene las predicciones finales.

Las neuronas que forman la red neuronal están conectadas entre sí a través de enlaces, en los que el valor de salida de la neurona anterior se multiplica por un peso. Además, a la salida de cada neurona puede haber una función umbral, denominada **función de activación**, que impone un límite que se debe sobrepasar para que los datos se envíen a la siguiente capa.

Para construir una red neuronal y realizar ese aprendizaje automático, se van actualizando los pesos de las neuronas, con el fin de minimizar una **función de pérdida**.

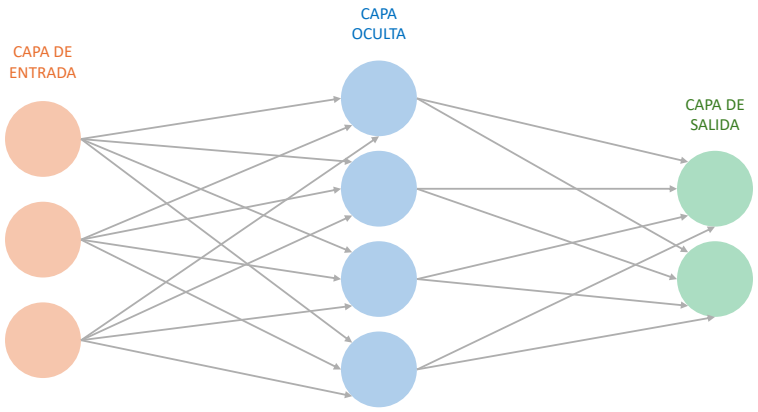

Figura 4.1: Esquema representativo de una red neuronal con una única capa oculta, en el que cada nodo circular representa una neurona y cada flecha representa el enlace desde la salida de una neurona a la entrada de otra.

Para que el rendimiento de una red neuronal sea aceptable se necesita que el tamaño de la muestra sea grande, ya que estas dependen de numerosos hiperparámetros, lo que las convierte en problemas de optimización complicados. Además, las redes son sensibles a la escala de las variables predictoras, luego será necesario preprocesar los datos, aplicando alguna transformación sobre los mismos.

Como ventajas y desventajas de este tipo de modelos, se podría decir que si bien suelen requerir de más tiempo de computación que otros algoritmos, se trata de modelos muy robustos, especialmente ante datos de grandes dimensiones. No obstante, aunque su rendimiento predictivo puede ser muy bueno, carecen de interpretabilidad, complicándose cada vez más, a medida que se aumenta el número de capas ocultas.

Tras introducir de manera genérica las redes y antes de pasar a estudiar las redes $LSTM$, cabe introducir de manera más formal las redes neuronales artificiales y las redes neuronales recurrentes, un tipo de las anteriores en el que se encuadran las $LSTM$.

4.1.1. Redes neuronales artificiales ($\boldsymbol{ANN}$)

Como ya se adelantaba, las redes neuronales artificiales (*Artificial Neural Networks*, ANN) son modelos de aprendizaje estadístico que constan de una estructura formada por nodos o neuronas, organizados en capas. Por simplicidad, para explicar las ANN nos centraremos en una red con una única capa oculta.

Dicha capa oculta estará formada por M nodos, resultantes de una combinación lineal de las variables de entrada del modelo $\boldsymbol{X} = (X_1, \ldots, X_N)$,

$$Z_m = \sigma\left(\alpha_{0m} + \boldsymbol{\alpha}_m^T \boldsymbol{X}\right),\ m = 1, \ldots, M,$$

donde σ es una función no lineal conocida como función de activación y α_{0m}, $\boldsymbol{\alpha}_m = (\alpha_{1m}, \alpha_{Nm})'$ parámetros a estimar.

Observación 4.1. *Es habitual utilizar la función logística o sigmoidal como* σ*,*

$$\sigma(u) = \frac{1}{1+e^{-u}},$$

especialmente en problemas de clasificación. En el caso de problemas de regresión es más habitual la función tangente hiperbólica,

$$\sigma(u) = \frac{sinh(u)}{cosh(u)} = \frac{e^{u}-e^{-u}}{e^{u}+e^{-u}}.$$

Se pueden consultar otras funciones de activación en Sharma et al. (2017).

El modelo final consiste en una transformación de una combinación lineal de los nodos de la capa oculta, es decir,

$$f_k(\boldsymbol{X}) = g_k(T_k) = g_k(\beta_{0k} + \boldsymbol{\beta}_k^T \boldsymbol{Z}),\ k = 1, \ldots, K,$$

donde g_k son funciones de activación que permiten adaptar la predicción a cualquier tipo de respuestas y β_0, $\boldsymbol{\beta}_k = (\beta_{1k}, \ldots, \beta_{Mk})$ son de nuevo parámetros a estimar. En problemas de regresión habitualmente se tiene una única respuesta ($K = 1$) y la función g_k es la identidad, mientras que en problemas de clasificación hay tantas respuestas como clases y es habitual emplear la función *sotfmax*,

$$g_k(T_k) = \frac{e^{T_k}}{\sum_{l=1}^{K} e^{T_l}}.$$

Así, en una red neuronal con una única capa oculta hay $M(N+1) + K(M+1)$ parámetros desconocidos, usualmente llamados pesos. La estimación de los mismos se hace minimizando una función de pérdidas, habitualmente la suma residual de cuadrados (RSS) en el caso de regresión y la función de entropía cruzada (*cross-entropy-function*), en el de clasificación; problemas cuya solución exacta puede ser imposible de obtener. Por ello, se recurre al algoritmo heurístico de descenso de gradientes, *backpropagation*, que convergerá a un óptimo local, pero difícilmente al óptimo global. En el algoritmo se van tomando los datos de entrenamiento por lotes (*batchs*) y calculando el ratio de aprendizaje (*learning rate*), un hiperparámetro que se encarga de controlar cuánto cambia el modelo cuando se actualizan los pesos. Este proceso se repite de manera iterativa, siendo el número total de iteraciones conocido como *epochs*. No se dan más detalles al no ser este tipo de redes el objetivo primordial del trabajo, sin embargo, un desarrollo más detallado acerca del mismo puede consultarse en la Sección 11.4 de Hastie et al. (2009) o en Werbos (1974).

Notar que algunas de las principales desventajas de dicho algoritmo son su lentitud e inestabilidad. El modelo resultante es muy sensible a la solución inicial, que típicamente se toma de manera aleatoria con valores próximos a cero (pero distintos de cero porque sino el algoritmo no se movería). Además, se ve afectado negativamente por la correlación entre las variables predictoras, problema que suele abordarse preprocesando los datos.

4.1.2. Redes neuronales recurrentes (*RNN*)

Las redes neuronales recurrentes (*Recurrent Neural Networks*, RNN) son un tipo de redes neuronales artificiales especialmente diseñadas para tratar datos de series temporales. La diferencia principal frente a las redes neuronales artificiales es que puede haber interacción entre los nodos de las capas ocultas del modelo, modelizando así la dependencia de la serie temporal con respecto a los instantes de tiempo anteriores.

Las neuronas de una capa oculta de una RNN, además de recibir la entrada de la capa anterior, reciben en cada instante de tiempo su propia salida del instante de tiempo anterior. Para hacerse una idea, en la Figura 4.2 se recoge un esquema representativo de este comportamiento.

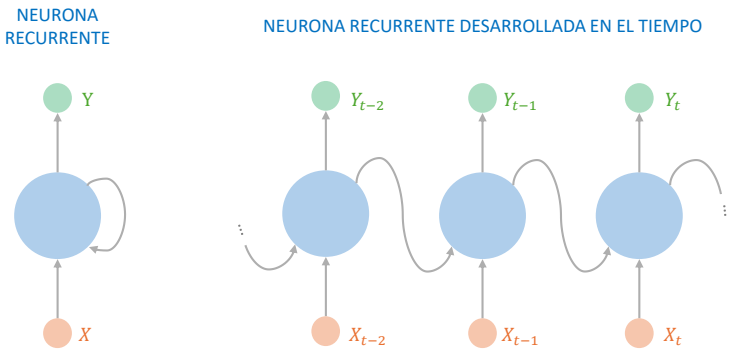

Figura 4.2: Esquema representativo del proceso que ocurre en un nodo de la capa oculta de una red neuronal recurrente atendiendo a Torres (2020).

Por tanto, la salida de una neurona recurrente en cierto instante de tiempo es una función que depende de los instantes de tiempo anteriores, por lo que podría decirse que tienen "memoria". A esta parte de las neuronas se le denomina célula de memoria (*memory cell*).

Así, en este caso cada neurona tiene asociados dos conjuntos de parámetros, los pesos de los datos x_t, que recibe de la capa anterior y, a mayores, otro conjunto de pesos asociados a la entrada de datos que proviene del vector de salida del instante anterior, y_{t-1}. Siguiendo Pascanu et al. (2013), este proceso podría formularse como

$$y_t = \sigma\left(W_{in}x_t + W_{rec}y_{t-1} + b\right),$$

donde y_t representa el resultado de la neurona, W_{in} la matriz de pesos asociada a las variables de entrada, W_{rec} la matriz de pesos asociada a la salida del instante anterior, b el sesgo del modelo y σ la función de activación. Nótese que el valor de salida inicial, y_0, es fijado por el usuario, habitualmente a cero.

Para estimar dichas matrices de pesos, en similitud con las redes neuronales artificiales, se sigue el algoritmo de *backpropagation*, denominado en este caso *backpropagation through time* (BPTT), al ser una generalización del anterior que tiene en cuenta el proceso de recurrencia con respecto al tiempo.

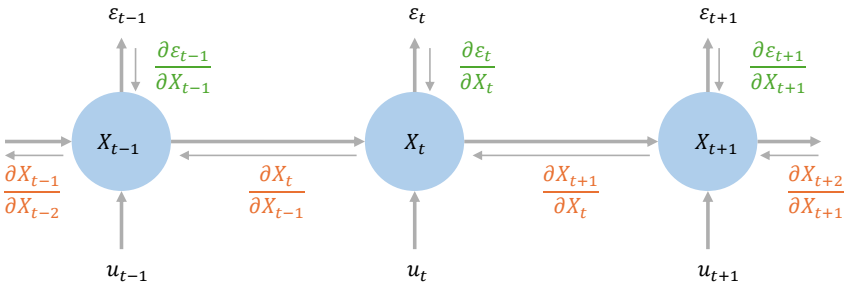

Figura 4.3: Esquema representativo del algoritmo de *backpropagation through time* (BPTT) en una red neuronal recurrente de acuerdo con Pascanu et al. (2013).

En la Figura 4.3 se recoge un esquema representativo del funcionamiento del algoritmo, cuyos elementos implicados se detallan a continuación. Sea ε_t, $t = 1, \dots, T$, el error obtenido en el instante t, se define $\varepsilon = \sum_{1 \leq t \leq T} \varepsilon_t$ como el coste que mide el rendimiento de la red.

De nuevo siguiendo Pascanu et al. (2013), el algoritmo de BPTT se puede formular como,

$$\begin{aligned}
\frac{\partial \varepsilon}{\partial \theta} &= \sum_{1 \leq t \leq T} \frac{\delta_t \varepsilon}{\delta \theta}, \\
\frac{\partial \varepsilon_t}{\partial \theta} &= \sum_{1 \leq k \leq t} \left(\frac{\partial \varepsilon_t}{\partial y_t} \frac{\partial y_t}{\partial y_k} \frac{\partial^+ y_k}{\partial \theta} \right), \\
\frac{\partial y_t}{\partial y_k} &= \prod_{k < i \leq t} \frac{\partial y_i}{\partial y_{i-1}} = \prod_{k < i \leq t} W_{rec}^T diag(\sigma'(y_{i-1})),
\end{aligned}$$

donde θ representa los parámetros del modelo.

Observación 4.2. *Acerca de las ecuaciones del algoritmo BPTT:*

- $\frac{\partial^+ y_k}{\partial \theta}$ *se refiere a la derivada parcial "inmediata" de* y_k *con respecto a* θ*, siendo* y_{k-1} *constante con respecto a* θ*.*
- $\frac{\partial \varepsilon_t}{\partial y_t} \frac{\partial y_t}{\partial y_k} \frac{\partial^+ y_k}{\partial \theta}$ *mide cómo* θ *en el instante* k *afecta al error* ε_t *en los instantes posteriores* $t > k$*. Estos factores se denominan contribuciones o componentes temporales a los gradientes* $\frac{\partial \varepsilon_t}{\partial \theta}$*.*
- *Se dice que los factores* $\frac{\partial y_i}{\partial y_{i-1}}$ *transportan el error del instante* t *hacia el instante* k*, diferenciando entre:*
  - *Contribuciones a largo plazo (long term contributions), referidas a componentes en las que el instante* k *es muy anterior a* t*,* $k \ll t$*.*
  - *Contribuciones a corto plazo (short term contributions), cuando los valores de* k *son menores que* t*, pero próximos.*
- *La notación diag se refiere al operador que convierte un vector en una matriz diagonal y* σ' *es la derivada de la función de activación.*

Dos problemas que afectan a las redes neuronales recurrentes son los gradientes explosivos (*exploding gradients*) y los gradientes que se desvanecen (*vanishing gradients*). Por un lado, el primero de estos ocurre cuando el algoritmo asigna un valor exageradamente alto a los pesos, lo que genera un problema de entrenamiento. En particular, se refiere a un incremento muy significativo en la norma de los gradientes en la fase de entrenamiento de la red neuronal, debido a un gran aumento en las contribuciones a largo plazo en comparación con las componentes a corto plazo. Este problema podría solucionarse fácilmente reduciendo el número de gradientes a calcular

Mientras tanto, el problema de *vanishing gradients* ocurre en el caso opuesto, cuando hay un decrecimiento exponencial en las contribuciones a largo plazo, lo que imposibilita el entrenamiento del modelo para aprender correlaciones entre instantes de tiempo muy distantes. Este problema depende en gran medida de la función de activación, pero es mucho más difícil de resolver que el anterior, al pasar su solución por modificar la célula de memoria de la red neuronal. No obstante, este problema fue resuelto con las redes $LSTM$, que se introducen en la siguiente sección.

Para un desarrollo más detallado y formal de estos problemas, puede consultarse Pascanu et al. (2013).

4.2. *LSTM*

Las redes neuronales de memoria a corto y largo plazo, $LSTM$ (*Long Short-Term Memory*), propuestas por Hochreiter and Schmidhuber (1997), son una extensión de las redes neuronales recurrentes, que amplían la memoria de las mismas a largo plazo.

Una desventaja de las RNN modelizando series de tiempo es que no retienen información histórica a largo plazo, volviéndose altamente dependientes del pasado reciente. Este problema se soluciona con las $LSTM$, cuyas células de memoria deciden si almacenar o eliminar información, dependiendo de su importancia.

A diferencia del resto de redes neuronales, en una célula de memoria de una red $LSTM$ se distinguen tres puertas de control de la información: la puerta de entrada (*input gate*), la puerta de salida (*output gate*) y la puerta de "olvidar" (*forget gate*). Estas puertas sirven para determinar si entra la información, si afecta a la salida en el paso de tiempo actual y si se elimina porque no es importante, respectivamente. La asignación de dicha importancia se realiza a partir de los pesos del modelo.

De acuerdo con esto y siguiendo Graves (2013), la formulación matemática del modelo $LSTM$ vendría dada por:

$$\begin{aligned} f_t &= \sigma\left(W_f x_t + U_f y_{t-1} + b_f\right), && (\textit{forget gate}) \\ i_t &= \sigma\left(W_i x_t + U_i y_{t-1} + b_i\right), && (\textit{input gate}) \\ o_t &= \sigma\left(W_o x_t + U_o y_{t-1} + b_o\right), && (\textit{output gate}) \\ \tilde{C}_t &= \tanh\left(W_C x_t + U_C y_{t-1} + b_C\right), && (\textit{cell memory state}) \end{aligned} \tag{4.1}$$

donde x_t representa las variables de entrada; W_f, W_i, W_o y W_C las matrices de pesos que relacionan las variables de entrada en la capa oculta con las puertas y con el estado de la célula; U_f, U_i, U_o y U_C las matrices de pesos que relacionan la salida de la célula en el instante anterior con las tres puertas y el estado de la célula; y b_f, b_i, b_o y b_C los vectores de sesgo en cada una de las puertas.

A su vez, σ representa la función de activación, que habitualmente es la función sigmoidal y $tanh$ es la función tangente hiperbólica, cuyo rol es filtrar la información de cara al siguiente nodo.

De acuerdo con las ecuaciones recogidas en (4.1), en cada tiempo t, el estado de la célula, C_t, y la salida, y_t, se pueden calcular como

$$\begin{aligned} C_t &= f_t C_{t-1} + i_t \tilde{C}_t, \\ y_t &= o_t \tanh(C_t), \end{aligned}$$

siendo C_{t-1} el estado de la célula en el instante anterior.

De este modo, el resultado final de un modelo $LSTM$ será un vector con todas las salidas

$$\boldsymbol{Y}_T = \left(y_{T-n}, \ldots, y_{T-1}\right).$$

Se puede ver en la Figura 4.4 un esquema representativo del funcionamiento de la célula de memoria de una red de este tipo.

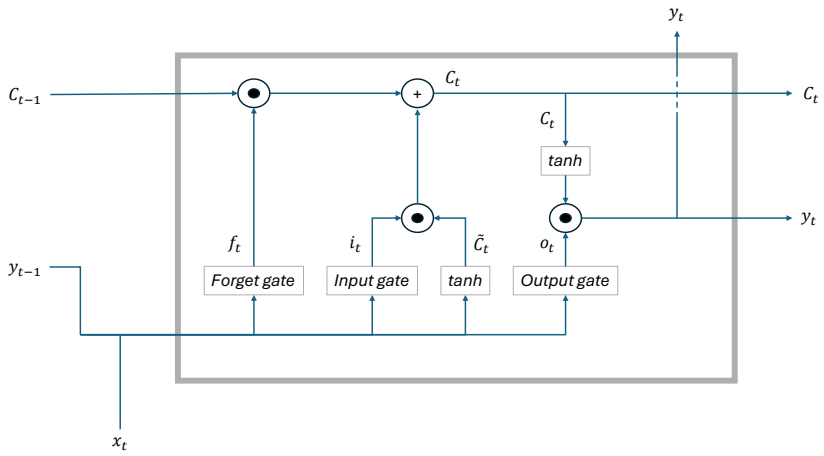

Figura 4.4: Esquema representativo de la célula de memoria de una red neuronal $LSTM$.

4.3. Selección de hiperparámetros

En la estimación de una red neuronal están implicados diversos hiperparámetros. Así, para estimar un modelo de este tipo será necesario seleccionar el valor óptimo de los mismos por algún método adecuado.

En el campo del aprendizaje estadístico, es habitual emplear la **validación cruzada** (*Cross Validation*, CV), que permite cuantificar el error de predicción utilizando una muestra de datos. Existen diversas variantes, siendo la más sencilla la validación cruzada dejando uno fuera (*Leave-one out cross-validation*), que consiste en realizar el ajuste empleando todas las observaciones salvo una y medir el error de predicción en dicha observación, de forma que la combinación de todos esos errores individuales resulte en una medida global del error de predicción.

Alternativamente, se puede recurrir a la **búsqueda en cuadrícula** (*grid search*) o a la **búsqueda aleatoria** (*random search*). El primero de los procesos consiste en crear una malla de todas las posibles combinaciones de hiperparámetros y seleccionar la mejor combinación en términos de error. A su vez, el segundo consiste en una búsqueda de tipo aleatorio en el espacio de posibles hiperparámetros. En algunos casos, la búsqueda aleatoria puede ser más eficiente, proporcionando mejores resultados y en menor tiempo (Bergstra and Bengio, 2012).

Otras metodologías para la selección de hiperparámetros son la **optimización bayesiana** (Snoek et al., 2012) o la **búsqueda mediante hiperbanda** (Li et al., 2018).

Parte II

Aplicación práctica

Capítulo 5

Análisis exploratorio de los datos

Para comenzar con la aplicación práctica, en este capítulo se desarrolla un análisis exploratorio de los datos de interés en el marco del proyecto. Más concretamente, en la Sección 5.1 se describen las variables con las que se va a trabajar y se explica el proceso de obtención de las mismas. A continuación, en la Sección 5.2 se realiza un análisis exploratorio de dichas variables, con el fin de conocer su comportamiento y preparar los datos antes de ajustar los correspondientes modelos. Y, por último, en la Sección 5.3 se exponen algunas hipótesis consideradas sobre las variables predictoras.

5.1. Variables empleadas en la modelización

Como ya se adelantaba en el Capítulo 1, el objetivo principal del trabajo es estudiar modelos para predecir el IPC y sus componentes, índices económicos mensuales que son publicados a mediados del mes siguiente por el INE. Para ello, se consideran además otras variables macroeconómicas que resultan de utilidad a lo largo de este estudio.

En la Sección 1.3 ya se adelantaba el tipo de variables que tendría sentido considerar. No obstante, para seleccionar las que se emplearán en los modelos, también hay que tener en cuenta otros aspectos. La fecha de publicación no es igual para todas ellas y algunas son de frecuencia trimestral (por ejemplo el PIB). En consecuencia, hay variables que se conocen con un mayor retraso que el IPC, lo cual no es muy interesante a la hora de estimar los modelos, pues tendríamos datos faltantes.

Las variables de interés son proporcionadas por organismos públicos a través de formatos abiertos, públicamente accesibles. Estos se recogen en la Tabla 5.1, junto con la abreviatura que se empleará a partir de este momento y la página web de referencia a ellos.

Organismos públicos proveedores de datos	**Abreviatura**	**Referencia**
Instituto Nacional de Estadística	INE	https://www.ine.es/
Ministerio de Economía, Comercio y Empresa	MINECO	https://portal.mineco.gob.es/
Ministerio de Industria y Turismo	MINTUR	https://www.mintur.gob.es/
Organización de las Naciones Unidas para la Alimentación y la Agricultura	FAO	https://www.fao.org/

Tabla 5.1: Organismos proveedores de los datos, abreviaturas empleadas y referencia de los mismos.

En la Tabla 5.2 se recogen todas las variables que se van a emplear en los próximos capítulos, acompañadas por la abreviatura que se utilizará para referirse a cada una de ellas, la fuente de la que provienen, su frecuencia (M mensual y D diaria) y sus unidades de medida.

Variable	Abreviatura	Proveedor	Frecuencia	Unidades de medida
IPC GENERAL	IPC	INE	M	Índice base 2021
IPC ALIMENTOS CON ELABORACIÓN, BEBIDAS Y TABACO	ALIMENTOS_SIN	INE	M	Índice base 2021
IPC ALIMENTOS SIN ELABORACIÓN	ALIMENTOS_CON	INE	M	Índice base 2021
IPC PRODUCTOS ENERGÉTICOS	PROD_ENERG	INE	M	Índice base 2021
IPC SERVICIOS	SERVICIOS	INE	M	Índice base 2021
IPC RESTAURANTES Y HOTELES	SERVICIOS_REST	INE	M	Índice base 2021
IPC SERVICIOS SIN RESTAURANTES Y HOTELES	SERVICIOS_OTROS	A partir del INE	M	Índice base 2021
IPC BIENES INDUSTRIALES SIN PRODUCTOS ENERGÉTICOS	BIENES_INDUS	INE	M	Índice base 2021
ÍNDICE FAO	FAO	FAO	M	Índice base 2014-16
PRECIO FUTURO A 1 MES. PETRÓLEO. CRUDO. BRENT	PREC_PETRO_FUT	MINECO	D	Dólares
TIPO DE CAMBIO DÓLAR USA/EURO - DATOS DIARIOS	TIPO_CAMBIO	MINECO	D	-
PRECIO PETRÓLEO EN EUROS MENSUALIZADO	PREC_PETRO	A partir de MINECO	M	Euros
PRECIO CARBURANTES	-	MINTUR	D	Euros
PRECIO CARBURANTES MENSUALIZADO	PREC_GASOLINA	A partir de MINTUR	M	Euros
PRECIO CARBURANTES MENSUALIZADO CON DESCUENTO	PREC_GASOLINA_DESC	A partir de MINTUR	M	Euros
ENCUESTA DE OCUPACIÓN HOTELERA. PERNOCTACIONES	PERNOCTACIONES	INE	M	Nº de pernoctaciones
PRECIO MAYORISTA DE LA ELECTRICIDAD. SPOT	-	MINECO	D	Euros
PRECIO MAYORISTA DE LA ELECTRICIDAD MENSUALIZADO	ELECTRICIDAD	A partir de MINTUR	M	Euros

Tabla 5.2: Información principal sobre las variables implicadas en los diferentes modelos.

Nótese que hay variables en la tabla que no se han obtenido directamente de dichos organismos, sino a partir de otras variables de la misma tabla. En particular:

- SERVICIOS_OTROS. A la serie del IPC de servicios se le sustrae la parte correspondiente al IPC de servicios de restaurantes y hoteles, teniendo en cuenta su peso.
- PREC_PETRO. Con las series diarias PREC_PETRO_FUT y TIPO_CAMBIO se obtiene el precio diario del petróleo en euros [1] y se calculan las medias mensuales para obtener esta serie.

[1] El motivo por el cuál se cambia el precio del petróleo a euros es que la inflación se mide en euros. Si considerásemos el petróleo en dólares, podríamos estar registrando crecimientos que en realidad se deben al tipo de cambio.

- `PREC_GASOLINA`. Se calcula la media de cada mes sobre los datos diarios del precio de los carburantes.
- `PREC_GASOLINA_DESC`. Se construye a partir de la variable `PREC_GASOLINA` teniendo en cuenta el descuento del Gobierno de 20 céntimos por litro de combustible que estuvo en vigor desde abril de 2022 hasta diciembre de 2022, con el objetivo de minimizar el impacto de la constante subida de precios de los carburantes que estaba habiendo.
- `ELECTRICIDAD`. Se obtiene a partir de la serie diaria extraída del Ministerio de Economía, Comercio y Empresa (MINECO) calculando las medias mensuales.

5.1.1. Automatización en la descarga de los datos

Con el fin de facilitar el acceso a estas variables, sin tener la necesidad de acceder a las correspondientes páginas web cada vez que se quieran actualizar los datos, se ha automatizado la descarga de esta información con `R`.

INE

Las series correspondientes con el IPC general y sus desagregaciones, la ponderación de cada una de las componentes sobre el IPC total y el número de pernoctaciones se obtienen del Instituto Nacional de Estadística (INE).

El INE cuenta con el servicio API JSON INE, que permite acceder a toda la información disponible en su base de datos mediante peticiones URL. La estructura de las URLs viene dada por el siguiente formato:

```
https://servicios.ine.es/wstempus/js/ES/DATOS_TABLA/{id_tabla}[nult=
                    n_ult_datos||date=AAAAMMDD:AAAAMMDD],
```

donde las llaves indican parámetros obligatorios y los corchetes parámetros opcionales. El "id_tabla" hace referencia a un código identificativo de cada tabla de la base de datos del INE, "nult" se refiere a la cantidad de períodos de datos y "date" al rango de fechas deseado. En el caso de la descarga de los datos del IPC, la URL correspondiente sería

`https://servicios.ine.es/wstempus/js/ES/DATOS_TABLA/50907?nult=10000000`.

El formato de salida de los datos a partir de las URLs es de tipo JSON (*Java Script Object Notation*), luego para automatizar la descarga de las series de índices requeridas basta leer los datos recogidos en la URL anterior desde `R` con la función `fromJSON` de la librería `jsonlite` (Ooms, 2014) y acceder a cada una de las series almacenadas en dicho conjunto de datos.

De igual modo, es posible obtener mediante este mismo método las correspondientes ponderaciones de cada una de las componentes del IPC consideradas para el análisis. Para obtener las ponderaciones de cada uno de los grupos resultantes de la desagregación del índice en los grupos especiales: alimentos con elaboración, bebidas y tabaco; alimentos sin elaboración; productos energéticos; servicios y bienes industriales sin productos energéticos en el último año disponible, basta acceder a los correspondientes elementos del conjunto de datos almacenado en

`https://servicios.ine.es/wstempus/js/ES/DATOS_TABLA/50951?nult=1`.

Una vez descargados estos datos, se pueden guardar en algún archivo en formato Excel, o bien utilizarlos directamente en R para los modelos.

Se muestra a continuación un ejemplo de este proceso, con la descarga del IPC de alimentos elaborados:

```
datos<-jsonlite::fromJSON("https://servicios.ine.es/wstempus/js/ES/DATOS_TABLA/
  50907?nult=10000000")
alimentos_con<-na.omit(ts(as.numeric(rev(datos$Data[[5]]$Valor)), start=rev(datos$
  Data[[5]]$Anyo)[1], freq=max(datos$Data[[5]]$FK_Periodo)))
```

Asimismo, este sistema de acceso desde R a la base de datos del INE a partir de peticiones URL es generalizable a cualquier serie estadística publicada por el INE. Consiguiéndose así una metodología útil para la explotación automática de la información estadística publicada, ya sea para utilizarse para posteriores análisis estadísticos y/o posibles modelizaciones mediante modelos estadísticos desde R; o bien con el fin de agilizar la descarga de diferentes series económicas publicadas por el INE que se necesitan consultar en el banco.

Ministerio de Industria y Turismo (MINTUR)

La descarga de los datos correspondientes con el precio de los carburantes se hace a partir del Ministerio de Industria y Turismo y también se automatiza.

La página web de MINTUR cuenta con un archivo de los precios diarios de todos los productos en todas las estaciones de servicio, que se actualiza cada media hora con los precios en vigor de ese momento. Las consultas a dicho archivo se hacen mediante peticiones URL diarias de la forma siguiente:

`https://sedeaplicaciones.minetur.gob.es/ServiciosRESTCarburantes/PreciosCarburantes/EstacionesTerrestresHist/dd-mm-aaaa`,

que igual que en el caso del INE, su formato de salida es de tipo JSON. Así pues, de manera análoga al caso anterior, se automatiza su descarga desde R.

Los datos históricos están disponibles desde el 1 de enero de 2007, y como se decía, cada media hora se van actualizando con los nuevos datos. Así pues, en primer lugar se construyó una rutina de descarga de los datos históricos, que después se completó con una rutina de actualización de los mismos.

Los datos de dichas peticiones URL vienen almacenados por estación de servicio y tipo de carburante. En este caso se seleccionan los carburantes "Gasolina 95" y "Gasoleo A" y se obtiene la media del precio de cada uno de ellos en todas las estaciones de servicio por día. Asimismo, también se obtiene la media de estos dos carburantes por día, y los precios aplicando el descuento de 20 céntimos por litro del Gobierno entre abril y diciembre de 2022.

El código correspondiente con la rutina de descarga sería el siguiente:

```
fechas<-format(seq(as.Date("2007-01-01"), today()-1, "days"), format="%d-%m-%Y")
datos<-data.frame(Fecha = fechas, matrix(0,nrow=length(fechas),ncol=2))
for (i in 1:length(fechas)){
  url<-paste("https://sedeaplicaciones.minetur.gob.es/ServiciosRESTCarburantes/
  PreciosCarburantes/EstacionesTerrestresHist/",as.character(fechas[i]),sep = "")
  dato<-fromJSON(url)
  if (is.null(dim(dato$ListaEESSPrecio))) {datos[i,2:3]<-c(NA,NA)} else {
```

```
    datos[i,2:3]<-c(mean(na.omit(as.numeric(gsub(",",".",na.omit(dato$
  ListaEESSPrecio$`Precio Gasolina 95 E5`)))), mean(na.omit(as.numeric(gsub(",",".",
  na.omit(dato$ListaEESSPrecio$`Precio Gasoleo A`))))))
}}
colnames(datos)[2:3]<-c("Gasolina95","Gasoleo_A")
datos[,2:3]<-as.numeric(datos[,2:3])
datos <- datos %>%  rowwise() %>% mutate(
"Promedio" = mean(c(Gasolina95, Gasoleo_A)),
"Gasolina95 con descuento" = if_else(as.Date(Fecha, format="%d-%m-%Y")<as.Date("
  2022-04-01") | as.Date(Fecha, format="%d-%m-%Y")> as.Date("2022-12-31"), Gasolina95
  , Gasolina95 - 0.2),
"Gasoleo_A con descuento" = if_else(as.Date(Fecha, format="%d-%m-%Y")<as.Date("
  2022-04-01") | as.Date(Fecha, format="%d-%m-%Y")> as.Date("2022-12-31"), Gasoleo_A,
   Gasoleo_A - 0.2),
"Promedio con descuento" = mean(c(`Gasolina95 con descuento`, `Gasoleo_A con
  descuento`)))
```

Todos estos datos se almacenan en un archivo excel que se irá actualizando. Además, se guarda la fecha del último día de descarga, con el fin de que no haya solapamiento en los datos.

Ministerio de Economía, Comercio y Empresa (MINECO)

En cuanto a los datos del Ministerio de Economía, Comercio y Empresa (MINECO), existen dos vías disponibles. Por un lado, se puede explorar su base de datos a través de un árbol,

https://portal.mineco.gob.es/es-es/economiayempresa/EconomiaInformesMacro/Paginas/bdsice.aspx,

que permite seleccionar las series deseadas y descargarlas manualmente. Y, por otra parte, los datos pueden descargarse en un archivo `.zip` que contiene ficheros `.csv` con todas las series.

Para automatizar la descarga, recurrimos a la segunda de las vías. Basta recurrir a la función `download.file` de la librería `utils` (R Core Team, 2023) para obtener dicho archivo comprimido y extraer las series requeridas con la función `unzip` de esta misma librería.

En este caso, nos interesan las series del tipo de cambio, el precio del petróleo y el precio de la electricidad, que son de frecuencia diaria. A modo ilustrativo se incluye a continuación el código para la descarga de las tres series y la lectura de la primera:

```
download.file(url="https://portal.mineco.gob.es/economiayempresa/
  EconomiaInformesMacro/Documents/bdsicecsv.zip",  destfile = "mineco.zip",mode="wb")
unzip(zipfile="mineco.zip", files=c("410100.CSV", "634824q.CSV", "880001q.CSV"),
  overwrite = TRUE, exdir="./datos mineco")

tipo_cambio<-fread("datos mineco/410100.CSV", header=T, sep=";", dec=",", fill=TRUE,
   encoding = "UTF-8")[,1:4]
date_tipo_cambio<-c()
for (i in 1:dim(tipo_cambio)[1]){
  date_tipo_cambio<-c(date_tipo_cambio,paste0(tipo_cambio[i,1], "-", tipo_cambio[i
  ,2], "-", tipo_cambio[i,3]))}

tipo_cambio<-data.frame(Fechas=as.Date(date_tipo_cambio),Tipos=as.numeric(gsub(",",
  ".",tipo_cambio$observaciones)))
tipo_cambio$Fechas<-floor_date(tipo_cambio$Fechas, "month")
tipo_cambio_mens<-tipo_cambio %>%
group_by(Fechas) %>%
summarize(media = mean(na.omit(Tipos)))
tipo_cambio_ts<-ts(tipo_cambio_mens$media, start=2006, freq=12)
```

Índice FAO

Por último, para automatizar la descarga de los datos del índice FAO, fue necesario investigar en la propia página web cómo obtener el enlace de descarga de los datos. Lo que se hizo fue analizar el código fuente de la página para encontrar el enlace asociado al "botón de descarga".

Una vez hecho esto, desde `R` es fácil leer las series recogidas en el mismo con ayuda de la función `read.table` de la librería `utils`:

```
url<-"https://www.fao.org/docs/worldfoodsituationlibraries/default-document-library/
  food_price_indices_data.csv?sfvrsn=e57cb8b0_24&download=true"
FAO<-as.data.frame(read.table(url,header=FALSE,sep=",",na.strings="_",
  stringsAsFactors=FALSE,fill=TRUE,fileEncoding ="latin1",encoding = "latin1"))[-c
  (1,2,3,4),1:2]
for (i in 1:(dim(FAO)[1])){
  FAO[i,1]<-paste0(FAO[i,1], "-01")
}
FAO[,1]<-as.Date(FAO[,1])
fao<-na.omit(ts(as.numeric(FAO[,2]), start = c(year(FAO[1,1]),month(FAO[1,1])), freq
  =12))
```

Programación de la automatización

Por último, si además se quiere automatizar la ejecución de los códigos mencionados en cada caso, se puede utilizar la librería `taskscheduleR` (Wijffels and Belmans, 2023), que permite automatizar la ejecución de un *script* con una frecuencia determinada.

En la Figura 5.1 se muestra la ventana de mandos de dicha librería. Basta seleccionar el archivo que se quiere automatizar, la frecuencia con la que se va a hacer y la fecha y hora de inicio.

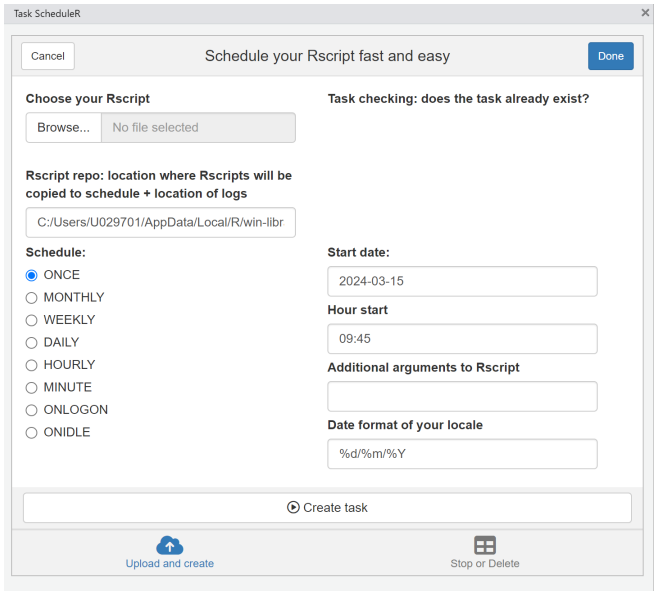

Figura 5.1: Ventana de automatización con la librería `taskscheduleR`.

Nótese que esto es aplicable al caso de los datos del INE, de los distintos ministerios y del índice FAO, así como a cualquier *script* de código que se quiera programar para su ejecución. Así, es una

herramienta muy útil para automatizar procesos, que resulta de interés para diversas operaciones realizadas en el banco.

5.2. Análisis de las variables

Ya descargados los datos con los que se va a trabajar, cabe realizar un análisis exploratorio de los mismos.

En primer lugar se analiza la serie del IPC y de sus componentes. Por un lado, en la Figura 5.2 puede verse la representación gráfica de las series de índices, que nos da una idea del comportamiento de las mismas.

A simple vista se aprecia que son series que presentan tendencia. Además, se puede observar que la variabilidad no es constante en todas ellas, por ejemplo, la serie del IPC de productos energéticos presenta una heterocedasticidad bastante marcada. Asimismo, aquí podemos apreciar un comportamiento heterogéneo en las componentes, lo que motiva la modelización de cada una de ellas por separado y agregarlas para obtener el IPC.

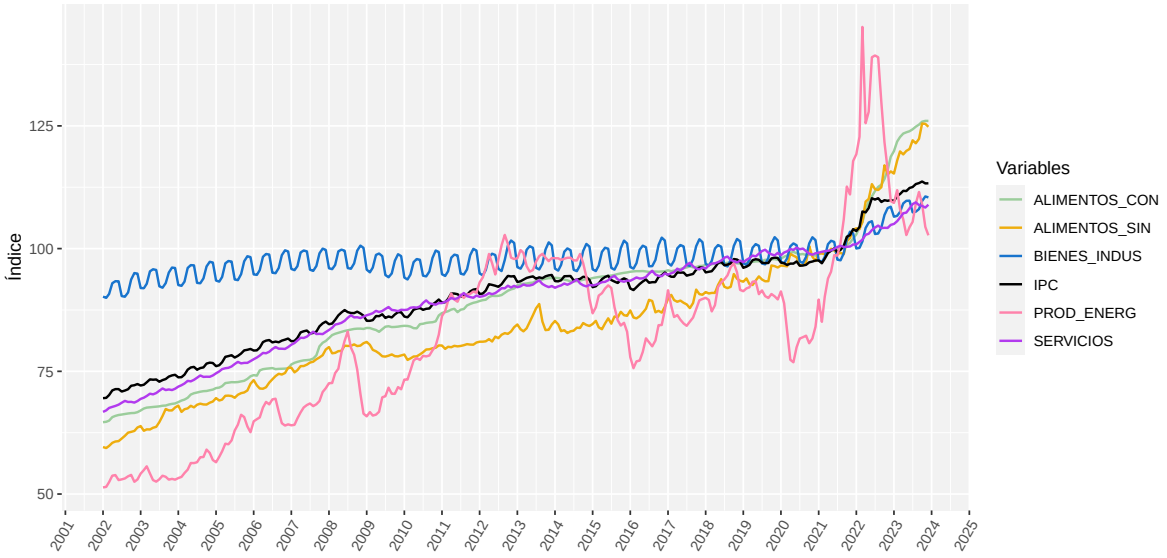

Figura 5.2: Series del IPC general y de sus componentes de grupos especiales.

Por otro lado, en la Figura 5.3 se representan las variaciones interanuales de dichos índices de precios, con el fin de estudiar cómo ha sido su evolución, especialmente en los últimos años. En particular, en los últimos dos o tres años se puede ver que todas estas series han presentado una variación mucho mayor que la que venían teniendo con respecto a los años anteriores. Es por esto que surge la necesidad de adelantarse al comportamiento de estas series y predecir como van a ir evolucionando para construir escenarios macroeconómicos para los próximos años.

Además, como ya se adelantaba observando la gráfica anterior, se puede apreciar un comportamiento distinto en cada una de las componentes del IPC, destacando sobre todo el IPC de productos energéticos.

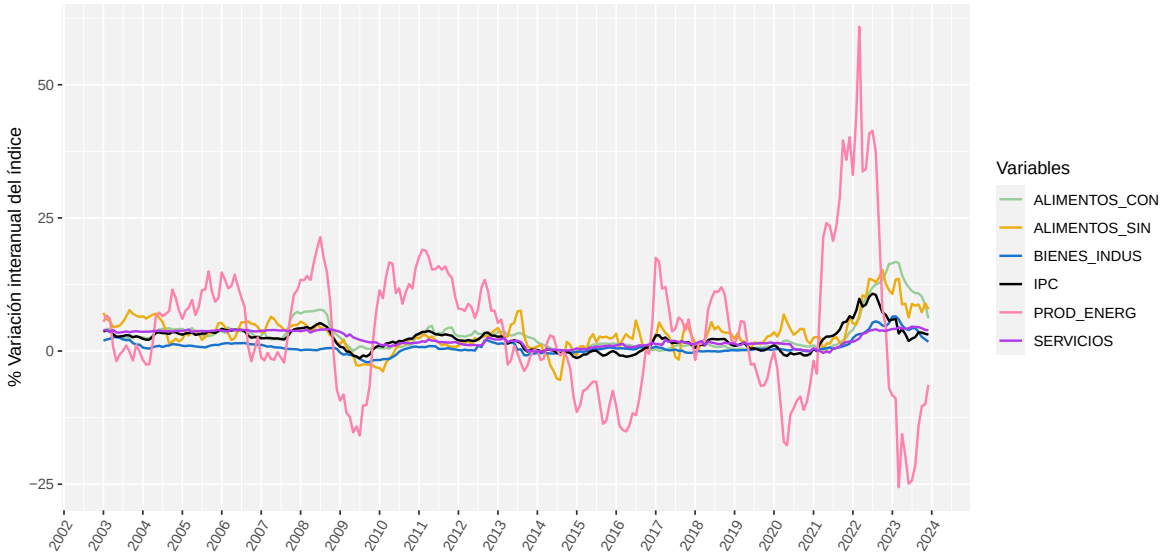

Figura 5.3: Variaciones interanuales de las series del IPC general y sus componentes de grupos especiales.

Atendiendo a ambos gráficos, no parece que estas series sean estacionarias, al apreciarse cierta tendencia y heterocedasticidad en las mismas, lo cual era esperable. Además, se observa estacionalidad en alguna de ellas, especialmente en la serie del IPC de bienes industriales se puede ver un patrón repetitivo marcado según la época del año (que podría deberse por ejemplo a los períodos de rebajas al estar incluidos en esta variable los precios de la ropa, de muebles, electrodomésticos, ...). Así, antes de modelar estas series, cabe profundizar más en estos aspectos. Esto se abordará en las Secciones 5.2.1 y 5.2.2.

Ahora bien, el objetivo es ajustar modelos que incluyan no solo las variables que queremos predecir, sino también variables regresoras, con el fin de aprovechar información que es publicada antes que la inflación. En consecuencia, será necesario estudiar también el comportamiento de estas variables. En la Figura 5.4 se puede ver la representación de las mismas. Nótese que el inicio de las series difiere, por ejemplo los datos del precio del petróleo comienzan en 1988, mientras que solo se dispone de datos del precio de los carburantes desde 2007.

Claramente ninguna de estas series es estacionaria. Más concretamente, el precio del petróleo y el de los carburantes en España presentan una tendencia creciente y una variabilidad cambiante. En cuanto al índice FAO, también se puede apreciar un crecimiento en su media y períodos de mayor variabilidad que otros. Destacar que, como era esperable, el patrón de comportamiento de las series del precio del petróleo y del precio de las gasolinas es muy similar (mismos períodos de subidas y bajadas) y que en las tres series comentadas hay un repunte en 2022, que podría estar motivado por la crisis energética debida a la Guerra de Ucrania.

Por otra parte, las pernoctaciones presentan una tendencia algo menos marcada y una clara estacionalidad con picos pronunciados en las temporadas vacacionales y caídas en épocas de temporada baja. En este caso además se ve un período atípico, provocado por la pandemia del Covid-19. Este impacto produjo una gran caída en las pernoctaciones a partir del mes de marzo de 2020, coincidiendo con el inicio del confinamiento en España. Posteriormente, se puede apreciar una lenta recuperación, que no alcanzó las cifras prepandemia hasta mediados de 2022.

Por último, en cuanto al precio de la electricidad, venía presentando un nivel más o menos constante hasta 2021, con subidas y bajadas quizás provocadas por la presencia de estacionalidad. Sin embargo, se aprecia entorno a 2022 un repunte muy alto en los precios, coincidiendo de nuevo con el estallido de la guerra. En 2023 los precios bajan, no obstante no llegaron todavía a los niveles anteriores a dicha crisis.

Figura 5.4: De izquierda a derecha: serie del precio del petróleo en euros, serie del precio de las gasolinas en España (teniendo en cuenta el descuento del Gobierno), serie del índice FAO, serie del número de pernoctaciones en España y serie de precios de la electricidad en España.

5.2.1. Corrección de estacionalidad

Una fuerte estacionalidad puede dar lugar a ruido a la hora de ajustar un modelo estadístico, es por esto que, antes de nada se comprueba la presencia o no de estacionalidad en las series. Para ello se recurre a la librería `seastests` (Ollech, 2021).

En la Tabla 5.3 se recogen los p-valores asociados a algunos test de estacionalidad. Atendiendo a estos, se concluye que todas las series, salvo la del IPC de alimentos elaborados y la del IPC de productos energéticos, tienen estacionalidad y por tanto, se les corregirá antes de ajustar los correspondientes modelos.

	QS-Test	Friedman-Test	Kruskall-Wallis-Test	Welch-Test	Estacionalidad
ALIMENTOS_SIN	0.00	0.00	0.00	0.00	Sí
ALIMENTOS_CON	1.00	0.00	0.08	0.38	No
PROD_ENERG	1.00	0.59	0.14	0.12	No
SERVICIOS	0.00	0.00	0.00	0.00	Sí
BIENES_INDUS	0.00	0.00	0.00	0.00	Sí

Tabla 5.3: P-valores resultantes de los tests de estacionalidad sobre las series del IPC.

En cuanto a las variables regresoras, se aplican los mismos test, cuyos resultados se recogen en la Tabla 5.4. En este caso, se obtiene que las únicas series que presentan estacionalidad son la de pernoctaciones y la del precio de la electricidad, como ya se preveía observando las gráficas.

	QS-Test	Friedman-Test	Kruskall-Wallis-Test	Welch-Test	Estacionalidad
FAO	1.00	0.63	0.85	0.94	No
PREC_PETRO	1.00	0.03	0.07	0.16	No
PREC_GASOLINA_DESC	1.00	0.29	0.17	0.22	No
PERNOCTACIONES	0.00	0.00	0.00	0.00	Sí
ELECTRICIDAD	1.00	0.00	0.00	0.10	Sí

Tabla 5.4: P-valores resultantes de los tests de estacionalidad sobre las series asociadas a las variables regresoras.

Para la corrección de las series temporales en `R` existen diferentes métodos. En este caso se recurre a la función `decompose`. Dada una serie temporal, Y_t, esta la descompone en tres componentes,

$$Y_t = T_t + S_t + e_t,$$ [2]

que son

- La **tendencia** (T_t), que representa la evolución de la serie a largo plazo y se obtiene como una media móvil.
- La **variación estacional** (S_t), que recoge comportamientos regulares o repetitivos en ciertos períodos de tiempo (producidos por ejemplo por las vacaciones, las rebajas, etc.). Una vez eliminada la componente de la tendencia, se obtiene como una media de los valores resultantes.
- Una **componente aleatoria** (e_t), que alberga efectos resultantes de hechos no previsibles y se calcula eliminando la tendencia y la componente estacional.

Así, de ahora en adelante, las series que así lo precisen se considerarán corregidas de estacionalidad. Esto es, sustrayendo la componente S_t. A modo de ejemplo, se puede ver en la Figura 5.5 la serie del IPC de bienes industriales original, junto con la serie sin componente estacional.

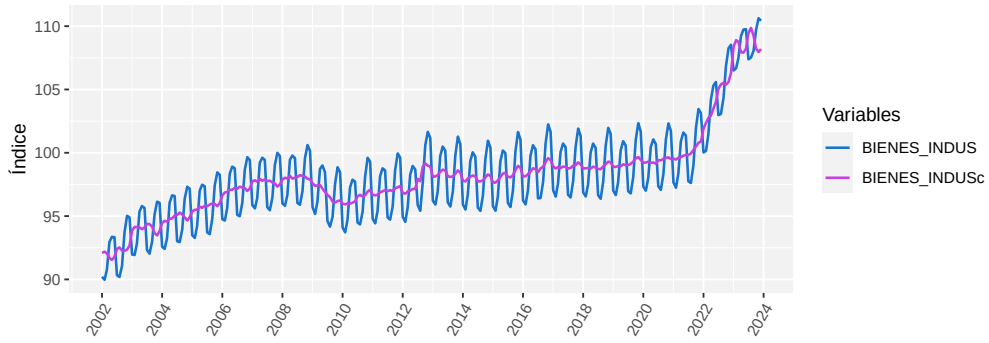

Figura 5.5: Serie del IPC de bienes industriales en España original y serie corregida de estacionalidad.

[2] Si bien en este caso se ha considerado un modelo aditivo, también cabe la posibilidad de considerarlo multiplicativo $Y_t = T_t \cdot S_t \cdot e_t$ según el caso.

5.2.2. Estacionariedad

Para verificar la falta de estacionariedad que se veía en las gráficas, se realizan test de estacionariedad sobre las series ya corregidas de la estacionalidad detectada en el apartado anterior, empleando la función `adf.test` de la librería `tseries` (Trapletti and Hornik, 2023). En la Tabla 5.5 se recogen los resultados del test de Dickey-Fuller para las series del IPC[3], que nos dicen que ninguna de estas es estacionaria, como ya preveíamos. En el caso del IPC de productos energéticos, se rechaza la hipótesis nula de no estacionariedad a un nivel de significación del 10 %, pero no al 5 % ni al 1 %, mientras que en el resto de casos no hay evidencias significativas para rechazar la hipótesis nula de no estacionariedad para ninguno de los niveles habituales (1, 5 y 10 %).

	Dickey-Fuller Test	**Estacionariedad**
ALIMENTOS_SINc	0.99	No
ALIMENTOS_CON	0.58	No
PROD_ENERG	0.08	No
SERVICIOSc	0.55	No
BIENES_INDUSc	0.51	No

Tabla 5.5: P-valores resultantes de los tests de estacionariedad de Dickey-Fuller sobre las series del IPC.

Atendiendo a los resultados, surge la necesidad de transformar las series para conseguir estacionariedad. En primer lugar, veíamos que las series de las componentes del IPC son heterocedásticas, luego como es habitual en series econométricas se aplicarán logaritmos para solucionar ese problema.

A su vez, veíamos en la Figura 5.2 que las series presentan tendencia. Si representamos gráficamente las autocorrelaciones del logaritmo de las series de las componentes, vemos además que estas son altas y tardan en decaer, lo que sugiere aplicar una diferencia regular. Puede verse en la Figura 5.6 el gráfico de autocorrelaciones del logaritmo de la serie de servicios corregida, en la que se observa el comportamiento mencionado.

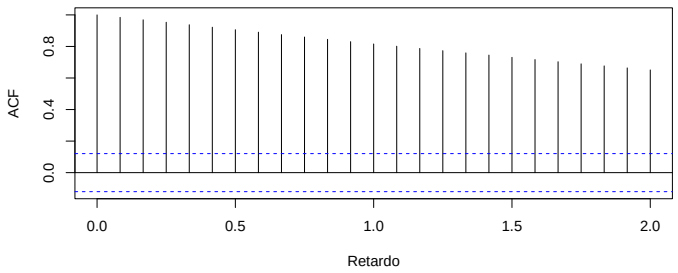

Figura 5.6: Gráfico de autocorrelación del logaritmo de la serie del IPC de Servicios corregida de estacionalidad.

[3]Las series cuyo nombre se acompaña de la letra "c" indican que han sido corregidas de estacionalidad previamente.

Atendiendo a este razonamiento, si estas series se transforman aplicando logaritmos y una diferencia regular y se obtienen de nuevo los p-valores asociados al test de Dickey-Fuller (Tabla 5.6), se concluye que dicha transformación es suficiente para alcanzar la estacionariedad de las series. Nótese que esta transformación se aproxima a las variaciones mensuales de las series.

	Dickey-Fuller Test	**Estacionariedad**
diff(log(ALIMENTOS_SINc))	0.01	Sí
diff(log(ALIMENTOS_CON))	0.04	Sí
diff(log(PROD_ENERG))	0.01	Sí
diff(log(SERVICIOSc))	0.01	Sí
diff(log(BIENES_INDUSc))	0.02	Sí

Tabla 5.6: P-valores resultantes de los tests de estacionariedad de Dickey-Fuller sobre las series del IPC transformadas mediante un logaritmo y una diferencia regular.

En cuanto a las variables regresoras, se puede ver en la Tabla 5.7 que salvo las pernoctaciones y el precio de la electricidad en España, las demás variables no son estacionarias.

	Dickey-Fuller Test	**Estacionariedad**
FAO	0.29	No
PREC_PETRO	0.14	No
PREC_GASOLINA_DESC	0.49	No
PERNOCTACIONESc	0.01	Sí
ELECTRICIDADc	0.04	Sí

Tabla 5.7: P-valores resultantes de los tests de estacionariedad de Dickey-Fuller sobre las series asociadas a las variables regresoras.

No obstante, igual que en el caso de las componentes del IPC, tras aplicar logaritmos y una diferencia regular, se soluciona el problema de no estacionariedad (Tabla 5.8).

	Dickey-Fuller Test	**Estacionariedad**
diff(log(FAO))	0.01	Sí
diff(log(PREC_PETRO))	0.01	Sí
diff(log(PREC_GASOLINA_DESC))	0.01	Sí

Tabla 5.8: P-valores resultantes de los tests de estacionariedad de Dickey-Fuller sobre las series asociadas a las variables regresoras tras aplicar logaritmos y una diferencia regular.

5.2.3. Análisis de causalidad y correlación entre las variables

Por último, una vez transformadas las series, como paso previo al ajuste de los modelos se analiza la correlación entre las diferentes variables. Por un lado, se analiza si las variables relativas a las componentes están correlacionadas entre sí, y después, se estudian las correlaciones con las variables exógenas consideradas. Esto permite saber, no solo si dos variables presentan correlación, si no también si dicha correlación es contemporánea.

En la Figura 5.7 se pueden ver las correlaciones cruzadas contemporáneas entre las series de las componentes del IPC, llegando a que, aparentemente, todas están muy correlacionadas. Ahora bien, como se adelantaba en la Sección 2.3.1, si las series no son de ruido blanco o una es estacionaria y otra de ruido blanco, se pueden producir correlaciones espúreas.

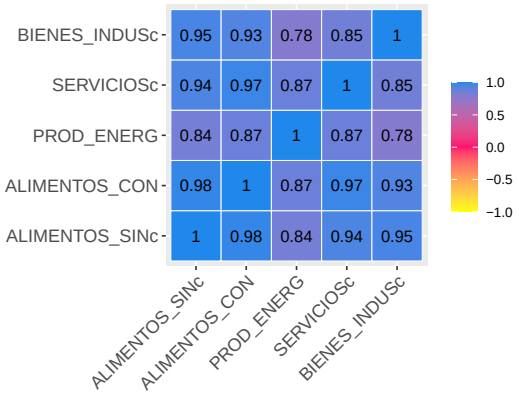

Figura 5.7: Correlaciones contemporáneas entre las series de las componentes del IPC sin componente estacional.

Como hemos visto, las componentes del IPC no son estacionarias y mucho menos, de ruido blanco, pero se pueden transformar mediante un proceso de preblanqueado para analizar su correlación.

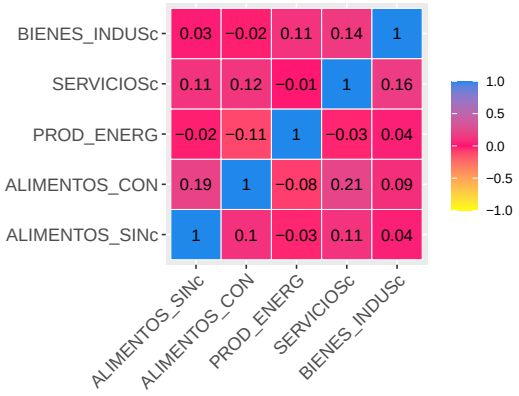

Figura 5.8: Correlaciones contemporáneas entre las series de las componentes del IPC sin componente estacional, estacionarias y preblanqueadas.

En la Figura 5.8 se representa la matriz de correlaciones cruzadas sobre las series correctamente tratadas. Se puede ver que, efectivamente las componentes están correlacionadas, pero dichas correlaciones no son tan elevadas como parecía en el primer análisis.

Las matrices anteriores se refieren a correlaciones contemporáneas entre las componentes del IPC. No obstante, con el fin de encontrar las mejores relaciones para incluir en los modelos, también es interesante analizar si las series están correlacionadas entre sí pero con cierto retardo.

En esta línea, si representamos un gráfico de correlaciones cruzadas sobre las series preblanqueadas para diferentes retardos de las series de tiempo, podemos ver cuál es el retardo en el que se produce una mayor correlación.

Por ejemplo, en la Figura 5.9 se representa la correlación cruzada entre las componentes de alimentos sin elaboración y alimentos con elaboración, bebidas y tabaco, y puede verse que la mayor correlación se da con un decalaje de 2 meses.

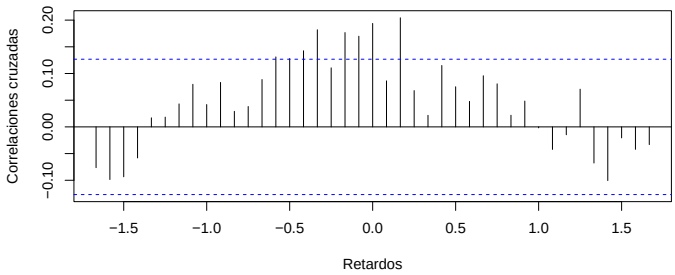

Figura 5.9: Correlaciones cruzadas entre el IPC de alimentos sin elaboración y el IPC de alimentos con elaboración, bebidas y tabaco, tras el proceso de preblanqueado.

Dicho análisis aplica no solo para la correlación entre componentes, sino también entre componentes y otras variables. Así, en la Figura 5.10 se representa el gráfico de correlaciones cruzadas según el retardo entre el precio de los carburantes y el IPC de productos energéticos. En este caso, se ve que existe una correlación contemporánea considerablemente marcada.

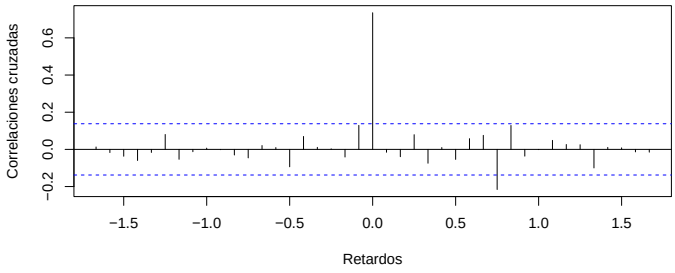

Figura 5.10: Correlaciones cruzadas entre el precio de los carburantes en España y el IPC de productos energéticos, tras el proceso de preblanqueado.

En relación con esta idea, cabe notar que a la hora de relacionar otras variables económicas con cada una de las componentes del IPC, se combina el análisis de correlación entre las mismas junto con la lógica y el sentido económico de los expertos de la entidad. Esto es, así como relacionar el IPC de productos energéticos con el precio de los carburantes es lógico, quizás no tendría mucho sentido relacionar las pernoctaciones en España con el IPC de bienes industriales, aunque su correlación fuera elevada.

5.3. Hipótesis de alto nivel sobre las variables regresoras

Con el fin de obtener predicciones de las componentes del IPC, si vamos a incluir variables exógenas en los modelos será necesario contar con predicciones para dichas variables. En este punto surgen dos posibilidades, o bien generar un modelo para predecir cada una de esas variables y utilizar esas predicciones en los modelos, o bien establecer ciertas hipótesis de alto nivel sobre estas variables siguiendo un criterio experto sobre el entorno macroeconómico.

La primera de las opciones tiene el inconveniente de que al estar "encadenando" predicciones podría estar aumentando el error de predicción. De acuerdo con este razonamiento y también basándose en la necesidad del banco de plantear distintos escenarios macroeconómicos que cuantifiquen el impacto que tienen dichas variables influyentes, se sigue el segundo camino. Así, en esta sección se presentan las hipótesis de alto nivel sobre las variables que se incluirán en algunos de los modelos como explicativas.

En primer lugar, para fijar las hipótesis relacionadas con la evolución de los precios del petróleo y la electricidad, se toman como referencia las hipótesis fijadas por el Banco Central Europeo en su escenario macroeconómico publicado en marzo de 2024 (ECB, 2024). De este modo, se fija que los precios medios a lo largo del 2024 serán de 80 dólares para el petróleo y 68.5 euros para la electricidad.

En cuanto al índice FAO, se proyecta un retorno progresivo hacia los precios medios del año 2021 y por último, para las pernoctaciones se supone que se situarán un 3 % por encima de los años prepandemia.

No obstante, como contraste, también se han ajustado modelos ARIMA a estas variables para comparar las predicciones vía modelos con el criterio de los expertos del área y las proyecciones propuestas por otros organismos.

Capítulo 6

Modelos de regresión dinámica para el IPC

Una vez obtenidos los datos de interés de estudio y analizadas las principales características de los mismos, se aplican las metodologías expuestas en la Parte I. En particular, en el presente capítulo se ajustarán los modelos estudiados en el Capítulo 2 mediante el uso de `R`, recurriendo a la librería `forecast` (Hyndman et al., 2023).

En la Sección 6.1 se presentan los modelos escogidos para el ajuste; y, a continuación, se realizan tareas de validación de los modelos ajustados, que se recogen en la Sección 6.2. Por último, en la Sección 6.3 se muestran los resultados de la estimación de los modelos, tanto por componentes como del IPC general.

6.1. Selección de los modelos

Como se ha dicho, se ajustará un modelo para cada una de las componentes del IPC descritas en el Capítulo 1, separando la categoría de servicios en servicios de restaurantes y hoteles y otro tipo de servicios, debido a su alto peso sobre el total del IPC[1].

En la Figura 6.2 se muestra un esquema a modo de resumen de las relaciones entre las variables implicadas en los modelos, sin especificar los correspondientes retardos considerados.

Nótese que las variables utilizadas para los modelos son corregidas de estacionalidad debidamente, de acuerdo con lo estudiado en el capítulo anterior.

En primer lugar, para ajustar el modelo correspondiente con el **IPC de productos energéticos** se utiliza el precio de las gasolinas (considerando el descuento del Gobierno), el cuál se predice a partir de otro modelo. Para este último, como variables regresoras se toman el precio del petróleo por las variaciones positivas del mismo y el precio del petróleo por las variaciones negativas, las cuales se pueden ver representadas en la Figura 6.1. De este modo, se recoge el impacto asimétrico que tienen los precios del petróleo sobre el precio de las gasolinas, basándonos en las evidencias reflejadas por numerosos estudios, como por ejemplo Atil et al. (2014) o Sun et al. (2019), entre otros.

[1] El IPC de servicios representa en 2024 un 46.86 % del IPC general, del cuál un 13.93 % proviene del IPC de restaurantes y hoteles y el 32.93 % restante del IPC del resto de servicios.

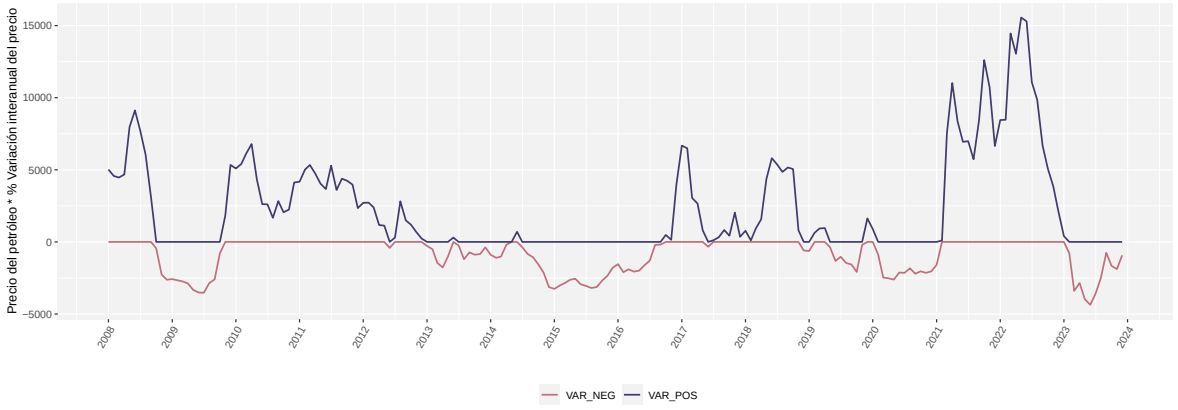

Figura 6.1: Precio del petróleo por las variaciones interanuales positivas y negativas del precio del mismo.

A continuación, se ajusta el modelo para el **IPC de alimentos sin elaborar**, explicando esta variable a partir del IPC de productos energéticos y el índice FAO, ambos retardados 10 meses. A su vez, el IPC de alimentos sin elaborar entra como variable regresora en el modelo del **IPC de alimentos elaborados** con un decalaje de 2 meses (de acuerdo con lo que se veía en la Figura 5.9).

En cuanto al **IPC de Servicios**, por un lado el referido a los restaurantes y hoteles se predice a partir de las pernoctaciones y del IPC de alimentos elaborados. Nótese que, las pernoctaciones no entran tal cual en este modelo. Como se veía antes en los gráficos, presentan una fuerte estacionalidad y además un período muy atípico en la época correspondiente con la pandemia del Covid-19. Lo que se hace es recortar esta serie hasta enero de 2020 y predecir los meses correspondientes con la pandemia mediante un modelo independiente. Con esta serie ya corregida y utilizando los datos reales post-pandemia, se le quita la estacionalidad y esta serie transformada es la que entra como regresora en el modelo. Mientras tanto, por otro lado, el IPC del resto de servicios se predice mediante un modelo ARIMA sin variables regresoras, que explica dicho IPC únicamente a partir de la serie histórica.

Por último, el **IPC de bienes industriales** se predice a partir del precio de la electricidad, con 10 meses de decalaje.

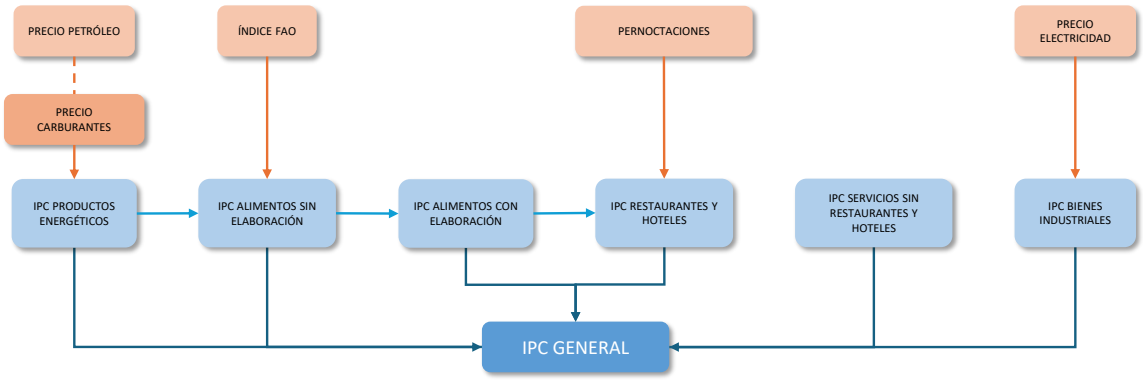

Figura 6.2: Esquema de las variables implicadas en los modelos.

Los detalles de estos modelos se recogen en la Tabla 6.1, en la que se resumen las transformaciones aplicadas a los datos y las variables implicadas que se acaban de mencionar, así como los órdenes de los modelos seleccionados y algunas medidas de bondad de ajuste.

Para seleccionar los órdenes más adecuados de los modelos, se ha creado una función de `R`, que consiste en seleccionar aquellos que presenten un mayor coeficiente de determinación y un menor BIC, teniendo en cuenta a su vez que sean modelos válidos y que los resultados de sus predicciones sean adecuados de acuerdo con la lógica económica.

Modelo	Transformación	Respuesta	Explicativas	Inicio	Fin	p	d	q	P	D	Q	BIC
Precio de las gasolinas	var. i .a.	PREC_GASOLINA	PREC_PETRO * (var. i.a. PREC_PETRO positivas)	2007M1	2023M12	0	1	1	0	0	1	916.24
			PREC_PETRO * (var. i.a. PREC_PETRO negativas)	2007M1	2023M12							
IPC de Productos energéticos	log	PROD_ENERG	log(PREC_GASOLINA_DESC)	2007M1	2023M12	1	0	0	0	0	0	−959.01
IPC de Alimentos sin elaborar	log	ALIMENTOS_SINc	log(lag(FAO, -10))	2002M11	2023M12	1	1	0	1	0	1	−1619.86
			log(lag(PROD_ENERG,-10))	2002M11	2023M12							
IPC de Alimentos con elaboración, bebidas y tabaco	log	ALIMENTOS_CON	log(lag(ALIMENTOS_SINc,-2))	2002M3	2023M12	1	1	0	0	1	1	−2070.17
Pernoctaciones	log	PERNOCTACIONES	-	2002M1	2020M1	1	0	1	2	1	1	−725.64
IPC de Restaurantes y hoteles	log	SERVICIOS_RESTc	log(ALIMENTOS_CON)	2002M1	2023M12	1	1	1	0	1	1	−2302.24
			log(PERNOCTACIONESc)	2002M1	2023M12							
IPC de Servicios sin restaurantes y hoteles	log	SERVICIOS_OTROSc	-	2002M1	2023M12	0	2	1	1	0	1	−2410.78
IPC de Bienes industriales	log	BIENES_INDUSc	log(lag(ELECTRICIDADc, -10))	2002M1	2023M12	2	1	2	0	1	1	−1732.98

Tabla 6.1: Detalles de los modelos ajustados: variables implicadas, transformaciones aplicadas, período de inicio y fin, órdenes del modelo y BIC.

6.2. Validación

Los modelos elegidos pasan la validación. Para comprobarlo, se llevan a cabo pruebas sobre los residuos de los modelos ajustados, cuyos p-valores se recogen en la Tabla 6.2.

Modelo	T test	Dickey Fuller	Shapiro Wilks	Ljung-Box	Significación
Precio gasolinas	0.90	0.01	0.00	0.89	0
IPC Productos Energéticos	0.51	0.01	0.00	0.64	0
IPC Alimentos sin elaboración	0.04	0.01	0.00	0.92	0
IPC Alimentos con elaboración	0.66	0.01	0.00	0.39	0
Pernoctaciones	0.11	0.01	0.01	0.34	0
IPC Restaurantes y hoteles	0.76	0.01	0.00	0.88	0
IPC Servicios sin restaurantes y hoteles	0.96	0.01	0.00	0.17	0
IPC Bienes Industriales	0.65	0.01	0.00	0.84	0

Tabla 6.2: P-valores resultantes de la fase de validación de los residuos de los modelos e indicador de significación de los coeficientes de los mismos.

Más concretamente, se utilizan las funciones `t.test` para contrastar que los residuos tengan media cero, `adf.test` para aplicar el test de Dickey-Fuller sobre los residuos, `shapiro.test` para contrastar la normalidad de los mismos y `Box.test` para llevar a cabo los test de Ljung-Box.

Observando los resultados, se puede decir que no hay evidencias estadísticamente significativas para rechazar las hipótesis sobre los residuos de los modelos para los niveles de significación habituales. Tan solo mencionar que en el caso del modelo de los alimentos no elaborados, no se rechaza que la media de los residuos sea nula a un nivel del 1 % de significación, pero sí a un nivel del 5 %, resultado mejorable pero que se considera válido atendiendo al resto de estadísticos.

Se incluye además un indicador de significación de los coeficientes, que se define como 0, si todos los coeficientes son significativos y 1, si alguno no lo es. Pero en este caso, los coeficientes de todos los modelos ajustados son significativos a los niveles de significación usuales.

Además, se representan las funciones de autocorrelación simple y parcial de los residuos de los modelos en las Figuras 6.3 y 6.4, respectivamente. En general, las autocorrelaciones se mantienen dentro de las bandas, salvo casos puntuales que podrían considerarse debidos a un error.

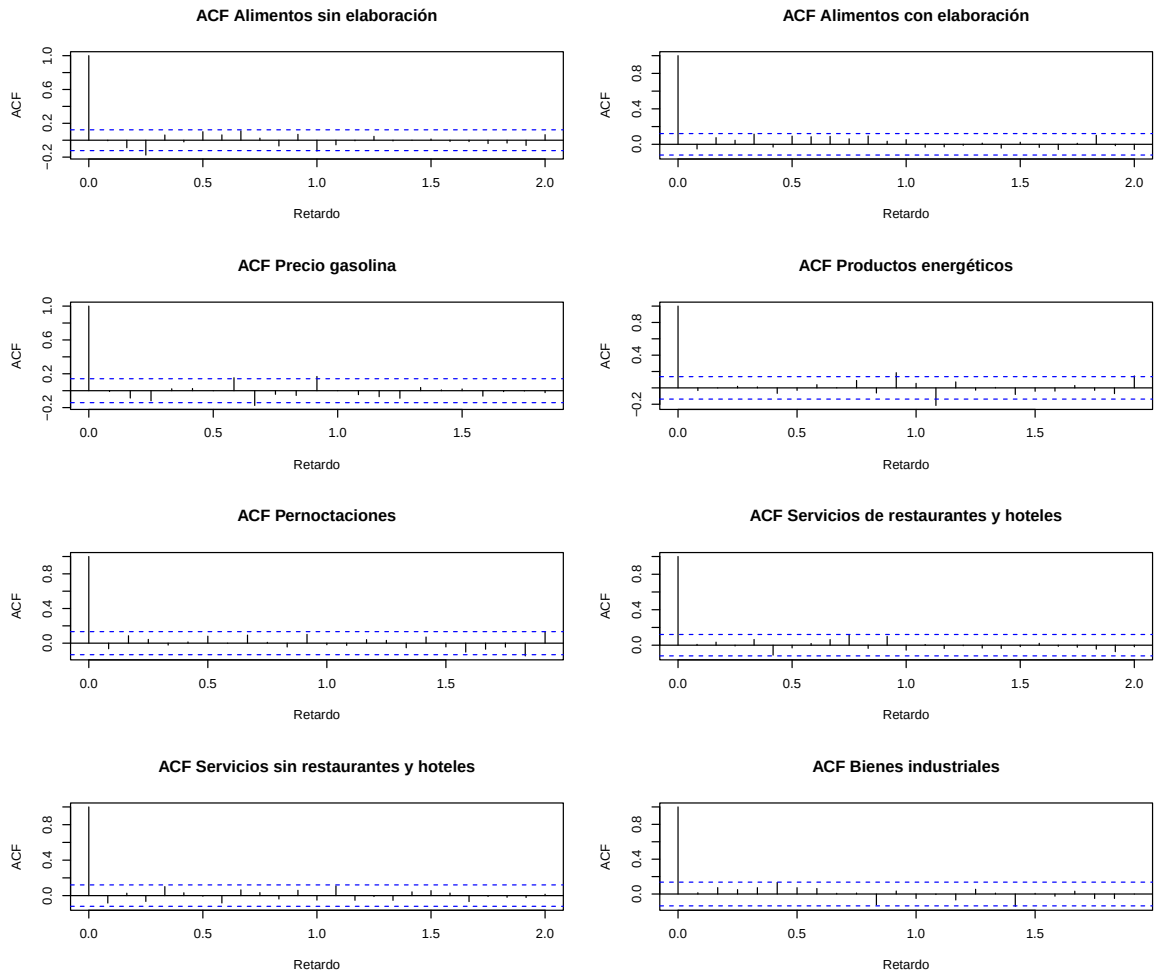

Figura 6.3: Funciones de autocorrelación simple de los residuos de los modelos.

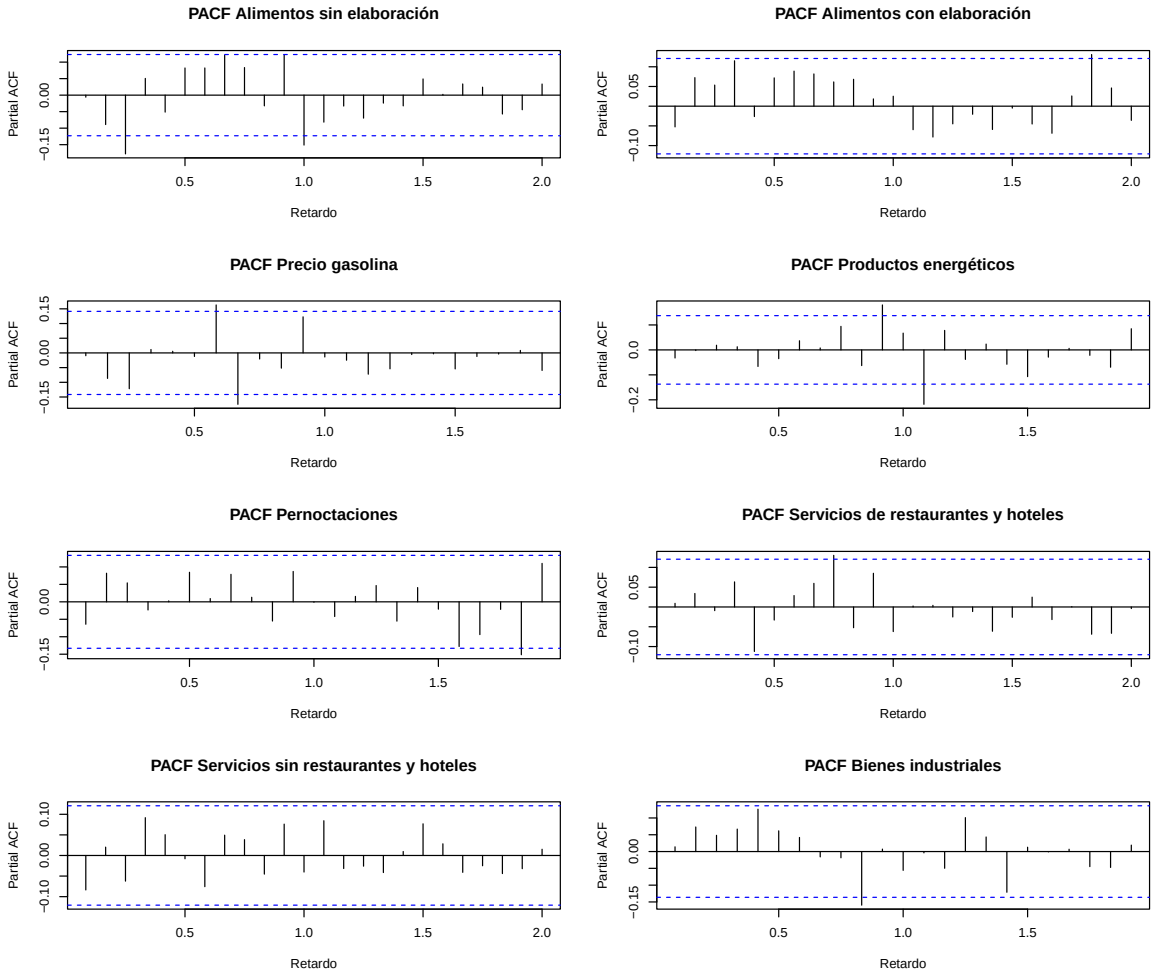

Figura 6.4: Funciones de autocorrelación parcial de los residuos de los modelos.

6.3. Estimación y ajuste

Una vez seleccionados los modelos y pasada la validación de los mismos se estiman sus coeficientes. En la Tabla 6.3 se recogen los coeficientes estimados de todos los modelos ajustados, junto con el error estándar asociado a la estimación y el p-valor asociado al contraste de significación de los mismos.

Nótese que el coeficiente asociado a las variables regresoras, en aquellos casos en los que se aplica el logaritmo y una diferencia regular ($d = 1$), puede interpretarse como el efecto que tendría sobre el IPC de la correspondiente componente, un aumento de 1 punto porcentual en la variación mensual de dicha variable. Así, por ejemplo, un incremento de un 1 % en el IPC de alimentos elaborados supondría que la tasa de variación del IPC de restaurantes y hoteles aumentase 0.12 puntos porcentuales, suponiendo que las pernoctaciones se mantienen constantes. Por otro lado, si en el modelo no se aplican diferencias regulares, dicho coeficiente nos daría la relación directa entre las variables. Por ejemplo, en el caso del modelo del IPC de productos energéticos, la interpretación sería que si se incrementa en 1 punto el logaritmo del precio de la gasolina, entonces el IPC de productos energéticos se vería multiplicado por $e^{0.98} = 2.66$.

Modelo	Coeficientes	Estimación	Error estándar	valor z	p-valor
Precio gasolinas	ma1	0.22	0.09	2.59	0.01
	sma1	-0.65	0.08	-8.59	0.00
	PREC_PETRO*VAR_POS	0.00	0.00	10.57	0.00
	PREC_PETRO*VAR_NEG	0.00	0.00	8.43	0.00
IPC Productos energéticos	ar1	0.90	0.03	27.90	0.00
	log(PREC_GASOLINA_DESC)	0.98	0.00	272.03	0.00
IPC Alimentos sin elaborar	ar1	0.12	0.06	1.88	0.06
	sar1	0.86	0.05	15.89	0.00
	sma1	-0.62	0.07	-8.59	0.00
	log(lag(FAO,-10))	0.05	0.02	2.07	0.04
	log(lag(PROD_ENERG,-10))	-0.06	0.02	-2.74	0.01
IPC Alimentos con elaboración	ar1	0.56	0.05	10.51	0.00
	sma1	-0.99	0.15	-6.47	0.00
	log(lag(ALIMENTOS_SINc,-2))	0.03	0.02	1.79	0.07
Pernoctaciones	ar1	0.97	0.02	60.34	0.00
	ma1	-0.57	0.06	-10.21	0.00
	sar1	-1.01	0.11	-9.19	0.00
	sar2	-0.61	0.07	-9.14	0.00
	sma1	0.39	0.14	2.81	0.00
IPC Servicios restaurantes y hoteles	ar1	0.94	0.04	25.99	0.00
	ma1	-0.83	0.06	-13.67	0.00
	sma1	-0.58	0.05	-10.82	0.00
	log(ALIMENTOS_CON)	0.12	0.04	3.07	0.00
	log(PERNOCTACIONESc)	0.01	0.00	3.38	0.00
IPC Servicios otros	ma1	-0.96	0.02	-51.20	0.00
	sar1	0.80	0.08	10.05	0.00
	sma1	-0.59	0.10	-6.17	0.00
IPC Bienes industriales	ar1	0.80	0.14	5.56	0.00
	ar2	-0.82	0.09	-9.22	0.00
	ma1	-0.49	0.14	-3.50	0.00
	ma2	0.79	0.08	10.50	0.00
	sma1	-0.39	0.09	-4.24	0.00
	log(lag(ELECTRICIDADc,-9))	0.00	0.00	1.68	0.09

Tabla 6.3: Ajuste de los modelos de regresión dinámica: estimación y significación de sus coeficientes.

En la Figura 6.5 se representan las series originales (o corregidas de estacionalidad, si aplica), junto con los valores ajustados de cada uno de los modelos.

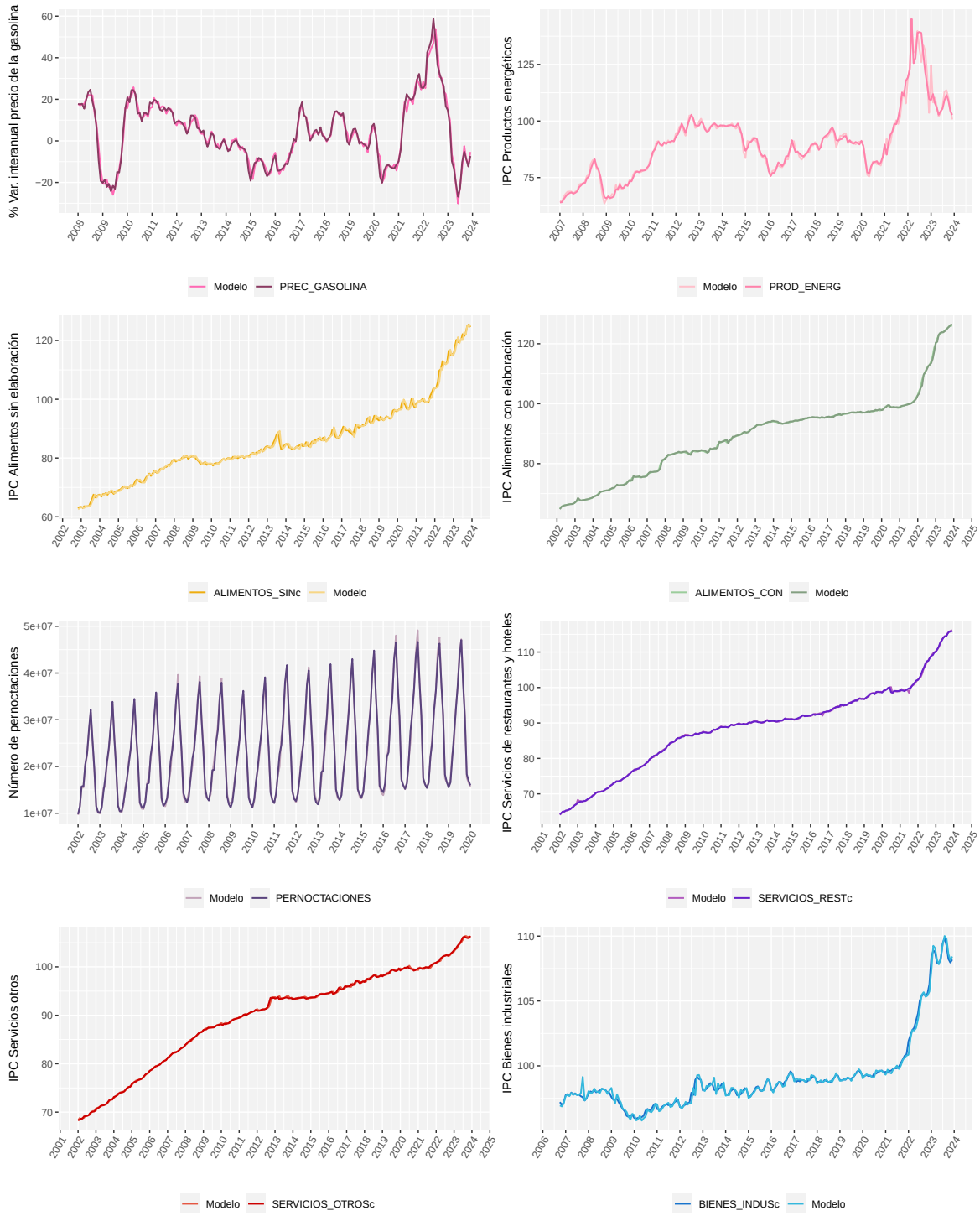

Figura 6.5: Series reales y valores ajustados de los modelos de regresión.

Ahora bien, vistos los modelos por componentes, nuestro principal objetivo de interés es obtener el ajuste del IPC general.

Nótese que no todas las series que se han modelizado son las series originales, si no que a algunas se les ha aplicado algún tipo de transformación. Así, es esperable que si se obtiene el ajuste del IPC a partir de los valores ajustados de los modelos, este no se aproxime a la serie original del IPC.

En su defecto, atendiendo a las anotaciones realizadas en el Capítulo 1, se calculan las correspondientes repercusiones mensuales de cada una de las componentes, una vez consideradas las correcciones de estacionalidad realizadas. Dichas repercusiones, en suma, nos devuelven las variaciones mensuales del índice general, con las que se puede construir el índice. De este modo reconstruimos el IPC general, obteniendo el índice que se compone por las componentes corregidas.

Hacemos lo mismo con los valores ajustados de las series resultantes de los modelos, obteniendo el ajuste del IPC general, que será comparable con el índice anteriormente reconstruido.

En la Figura 6.6 se recoge la representación de dicho IPC general, junto con su ajuste, tanto en niveles como en variaciones interanuales. A simple vista se puede ver que, en general, el ajuste es bueno y el período en el que comete un mayor error de estimación coincide con los últimos años, lo cual puede deberse al carácter atípico de estos.

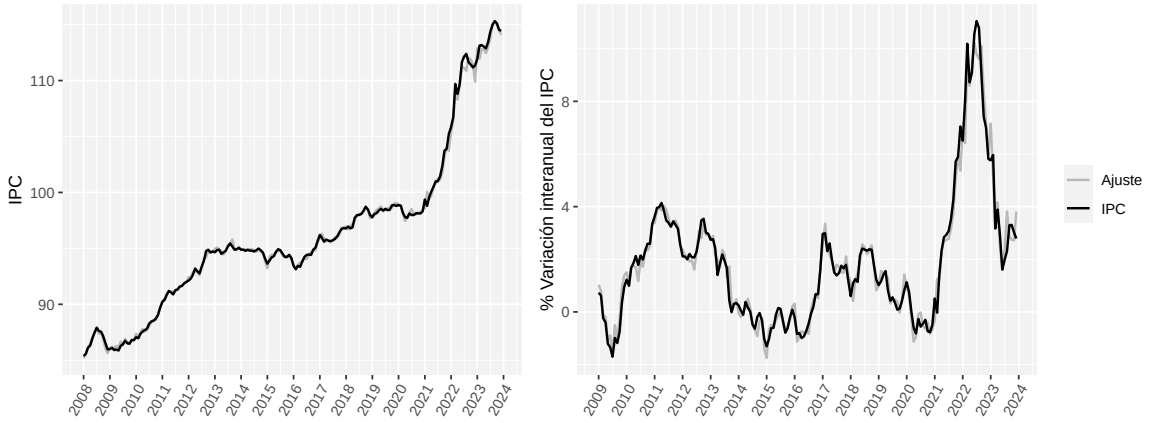

Figura 6.6: Series del IPC reconstruido y de sus variaciones interanuales junto con el ajuste a partir de los modelos de regresión dinámica seleccionados.

Capítulo 7

Modelos VAR y VECM para el IPC

Tras ajustar modelos univariantes a las series de las componentes del IPC y sabiendo que están correlacionadas entre sí, surge la idea de recurrir a modelos multivariantes, que recojan las posibles interacciones entre las diferentes componentes, además de incluir otras variables exógenas.

Así, en el presente capítulo se trabajará en el marco de los modelos VAR y VECM, presentados en el Capítulo 3. En particular, en la Sección 7.1 se comenzará seleccionando un modelo VAR cuyas variables endógenas sean todas las componentes, incluyendo a su vez variables exógenas en el mismo. Una vez seleccionado el modelo, se llevarán a cabo tareas de validación de los residuos del modelo, se estimarán sus coeficientes y se mostrará su ajuste. Paralelamente, en la Sección 7.2, se estudiará si las componentes del IPC están cointegradas con el fin de ajustar un modelo VECM, que también será validado y estimado. Además, en ambos casos se hará la descomposición de la varianza del error de predicción y se representarán algunas funciones de tipo impulso-respuesta.

7.1. Modelo VAR

En este caso el objetivo será ajustar un modelo VAR a una serie de tiempo multivariante que incluya las cinco componentes del IPC mencionadas, introduciendo a su vez las variables regresoras consideradas en el capítulo anterior como exógenas en el modelo. Para ello, se utiliza la librería `vars` (Pfaff, 2008b).

En línea con los modelos de regresión dinámica, se utilizan las series corregidas de estacionalidad y transformadas tomando logaritmos. Además, se consideran las variables exógenas retardadas de igual manera que en el capítulo anterior.

Nótese también, que al estar modelando todas las variables de forma simultánea, la longitud de todas las series debe ser la misma. Así, se consideran todas las variables desde enero de 2007 (debido a que el precio de las gasolinas solo está disponible desde esta fecha) hasta diciembre de 2023.

7.1.1. Selección

Para seleccionar el orden del modelo VAR más adecuado a nuestros datos, en primer lugar, se estiman por OLS modelos $VAR(m)$ con $m = 1, \ldots, 10$ y se obtienen los criterios de Akaike (AIC), de Hamilton-Quinn (HQ) y de Schwarz (SC) para cada uno de los valores del orden del modelo.

Los valores de dichos criterios para cada orden pueden verse reflejados en la Figura 7.1, en la que podemos ver que los valores $m = 4$, $m = 4$ y $m = 1$ minimizan el AIC, el HQ y el SC, respectivamente. Por consiguiente, se toma el orden del modelo como $p = 4$, es decir, se ajustará un $VAR(4)$.

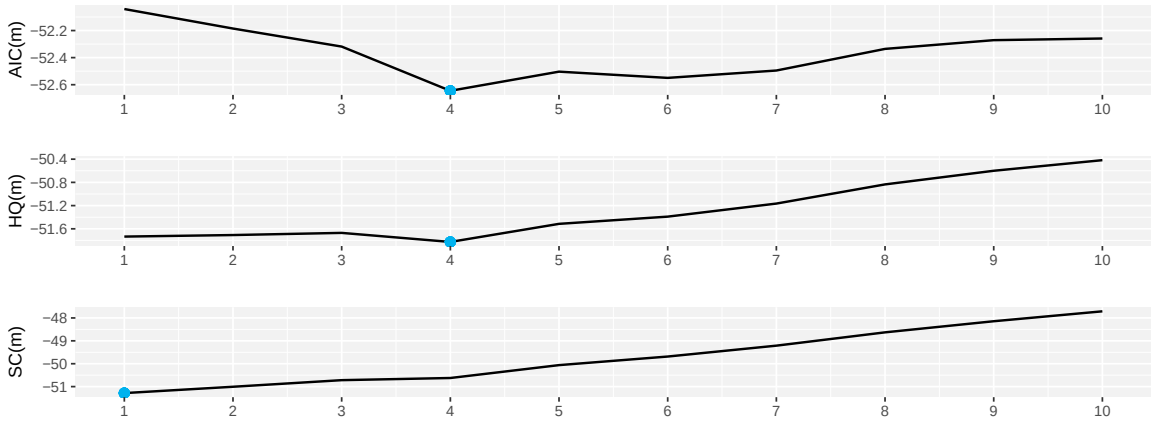

Figura 7.1: Variación de los criterios de información de Akaike (AIC), de Hamilton-Quinn (HQ) y de Schwarz (SC) según el orden del modelo VAR.

En este punto, se observa que todas las variables exógenas incluidas en el modelo salvo el precio de las gasolinas resultan no significativas. En consecuencia, se eliminan las variables relativas al índice FAO, las pernoctaciones y el precio de la electricidad, manteniéndose el precio de los carburantes con descuento como única variable exógena en el modelo.

Si se repite el proceso de selección del orden tras descartar estas variables, se llega a los mismos resultados, tal y como se puede ver en la Tabla 7.1.

	1	**2**	**3**	**4**
$\boldsymbol{AIC(m)}$	-52.08	-52.26	-52.40	-52.72
$\boldsymbol{HQ(m)}$	-51.88	-51.88	-51.86	-52.00
$\boldsymbol{SC(m)}$	-51.58	-51.33	-51.05	-50.95

Tabla 7.1: Valores de los criterios de información de Akaike (AIC), de Hamilton-Quinn (HQ) y de Schwarz (SC) según el orden m del modelo VAR, tras descartar algunas variables exógenas.

Se ajustará entonces un modelo $VAR(4)$ con todas las componentes del IPC como variables endógenas y el precio de las gasolinas con descuento como exógena.

7.1.2. Validación

Seleccionado el modelo a ajustar, se comprueba que se trata de un modelo válido. Por ende, se analizan los residuos del mismo, atendiendo a las pruebas de diagnosis descritas en la Sección 3.1.3.

En la Tabla 7.2 se recogen los resultados del test de incorrelación de *Breusch-Godfrey*, el test *LM-ARCH* multivariante para la heterocedasticidad condicional y el test de normalidad multivariante de *Jarque-Bera*.

	Estadístico	**p-valor**
Breusch-Godfrey	1000.00	0.13
LM-ARCH	2529.77	0.22
Jarque-Bera	913.42	0.00

Tabla 7.2: Estadísticos y p-valores resultantes de la fase de validación de los residuos del modelo VAR.

Atendiendo a los resultados de los test se puede ver que no hay evidencias estadísticamente significativas para rechazar la incorrelación y la homocedasticidad de los residuos. No obstante, sí las hay para rechazar la normalidad de los mismos. Esto no es un problema, al no ser esta hipótesis indispensable para verificar la validez del modelo.

Adicionalmente, se puede comprobar la estabilidad estructural del proceso, representando gráficamente el proceso de fluctuación empírico, que permite detectar cambios estructurales en los coeficientes estimados del modelo. En la Figura 7.2 se pueden ver dichos ajustes para cada una de las variables endógenas del modelo ajustado, observándose que en ninguno de los casos se superan los límites determinados por las bandas de confianza (las líneas rojas), lo que confirma que el modelo es estructuralmente estable.

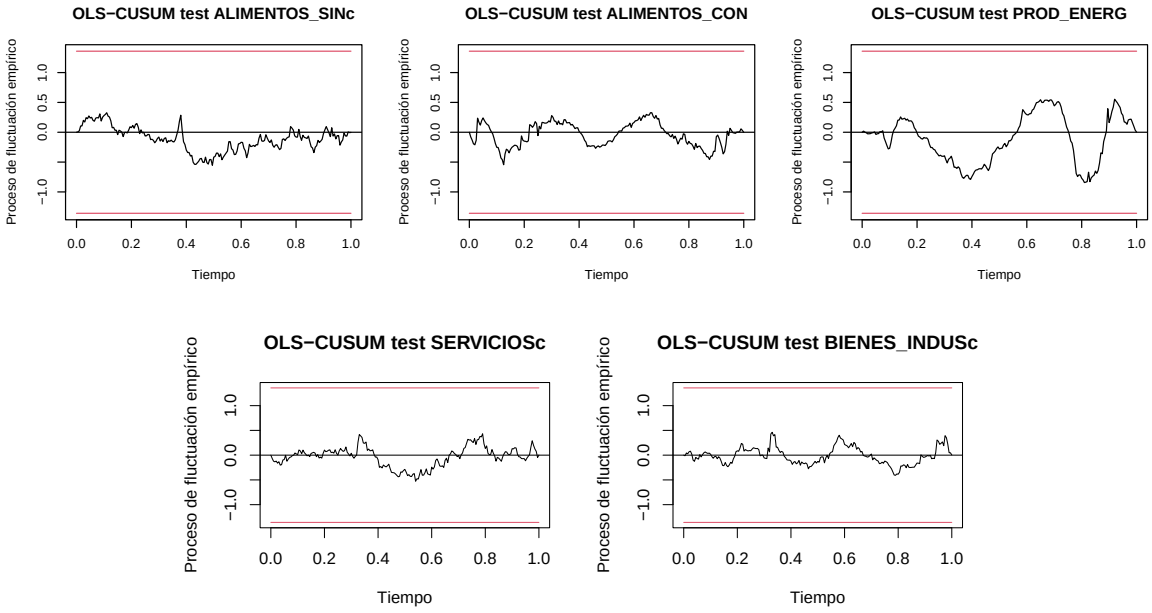

Figura 7.2: Ajustes de *OLS-CUSUM* para comprobar la estabilidad estructural de cada una de las variables endógenas en el modelo $VAR(4)$ ajustado.

Asimismo, también puede ser interesante contrastar las causalidades entre las componentes del IPC en el modelo, siguiendo la definición de *Granger*. En este contexto se realizan los contrastes recogidos en la siguiente tabla.

Contraste	**Estadístico**	**p-valor**
H_0 : ALIMENTOS_SINc no causa ALIMENTOS_CON, PROD_ENERG, SERVICIOSc, BIENES_INDUSc	4.30	0.00
H_0 : ALIMENTOS_CON no causa ALIMENTOS_SINc, PROD_ENERG, SERVICIOSc, BIENES_INDUSc	2.26	0.00
H_0 : PROD_ENERG no causa ALIMENTOS_SINc, ALIMENTOS_CON, SERVICIOSc, BIENES_INDUSc	1.68	0.05
H_0 : SERVICIOSc no causa ALIMENTOS_SINc, ALIMENTOS_CON, PROD_ENERG, BIENES_INDUSc	2.86	0.00
H_0 : BIENES_INDUSc no causa ALIMENTOS_SINc, ALIMENTOS_CON, PROD_ENERG, SERVICIOSc	5.56	0.00

Tabla 7.3: Estadísticos y p-valores resultantes de los contrastes de causalidad en el sentido de Granger (1969) en el modelo $VAR(4)$.

De acuerdo con los resultados de la Tabla 7.3, en cualquiera de los casos hay evidencias significativas para rechazar la hipótesis nula de no causalidad. En consecuencia, se puede decir que todas las variables causan al resto.

7.1.3. Estimación y ajuste

Así como en el capítulo anterior sobre los modelos de regresión se mostraron las estimaciones de todos los coeficientes de los modelos, junto con el ajuste gráfico de cada uno de ellos, en este caso, debido al elevado número de coeficientes implicados, no se incluyen todos. Al ajustar un modelo $VAR(4)$ con cinco variables endógenas y una exógena, el número de coeficientes del modelo asciende hasta $5 \cdot (4 \cdot 5 + 1) = 105$.

No obstante, se muestra una de las ecuaciones estimadas del modelo a modo ilustrativo y se recogen en la Figura 7.3 los ajustes para cada una de las componentes.

Si denotamos por

$$\boldsymbol{y}_t = \begin{pmatrix} y_{1,t} \\ y_{2,t} \\ y_{3,t} \\ y_{4,t} \\ y_{5,t} \end{pmatrix} = \begin{pmatrix} \log(\texttt{ALIMENTOS\_SINc}) \\ \log(\texttt{ALIMENTOS\_CON}) \\ \log(\texttt{PROD\_ENERG}) \\ \log(\texttt{SERVICIOSc}) \\ \log(\texttt{BIENES\_INDUSc}) \end{pmatrix}, \tag{7.1}$$

y $x_t = \log(\texttt{PREC\_GASOLINA\_DESC})$, entonces, la ecuación estimada para el IPC de alimentos sin elaborar viene dada por

$$
\begin{aligned}
y_{1,t} = & 0.958y_{1,t-1} - 0.002y_{2,t-1} + 0.047y_{3,t-1} - 0.037y_{4,t-1} - 0.239y_{5,t-1} \\
& -0.215y_{1,t-2} + 0.710y_{2,t-2} - 0.066y_{3,t-2} + 0.008y_{4,t-2} + 0.838y_{5,t-2} \\
& +0.011y_{1,t-3} - 0.888y_{2,t-3} + 0.056y_{3,t-3} - 0.164y_{4,t-3} - 0.311y_{5,t-3} \\
& +0.261y_{1,t-4} + 0.174y_{2,t-4} - 0.029y_{3,t-4} + 0.188y_{4,t-4} - 0.297y_{5,t-4} \\
& -0.001x_t.
\end{aligned}
$$

Las ecuaciones del modelo sirven también para analizar las relaciones entre las variables. Así, podemos ver por ejemplo que el valor del IPC de alimentos sin elaborar depende en gran medida del mismo el mes anterior o que está poco influenciado por el valor del precio de la gasolina de forma directa.

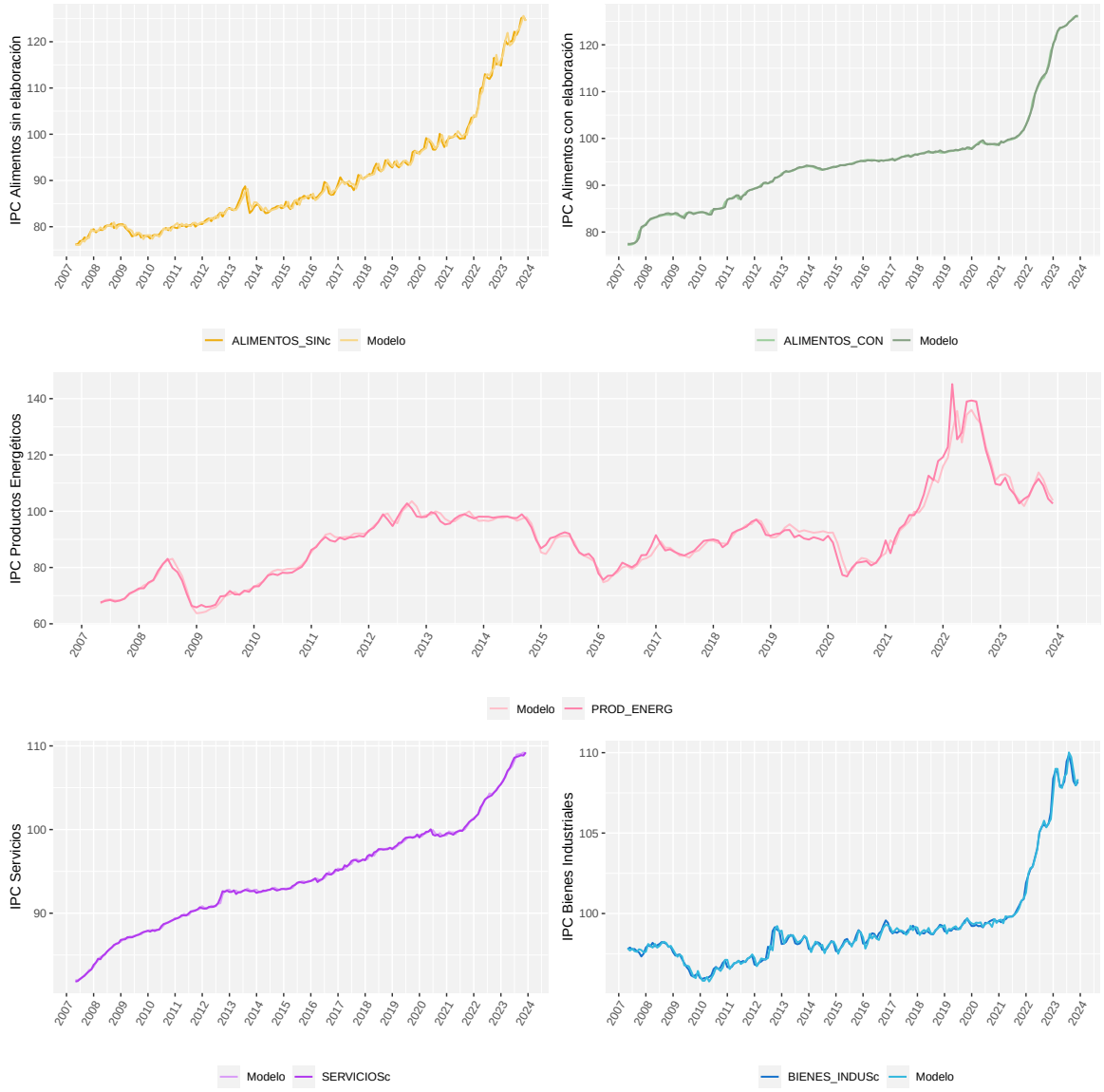

Figura 7.3: Series reales y ajustes del modelo $VAR(4)$.

De forma paralela al capítulo anterior, tras obtener el ajuste del modelo $VAR(4)$ para cada una de las componentes del IPC debidamente corregidas, se reconstruye el ajuste del IPC general. Como ya se mencionaba anteriormente, este ajuste se corresponderá con la serie del IPC resultante de componer las series corregidas.

En la Figura 7.4 se recoge la representación del IPC general reconstruido a partir de las componentes corregidas, junto con el ajuste del modelo VAR, en niveles en el primer caso y en variaciones interanuales en el segundo.

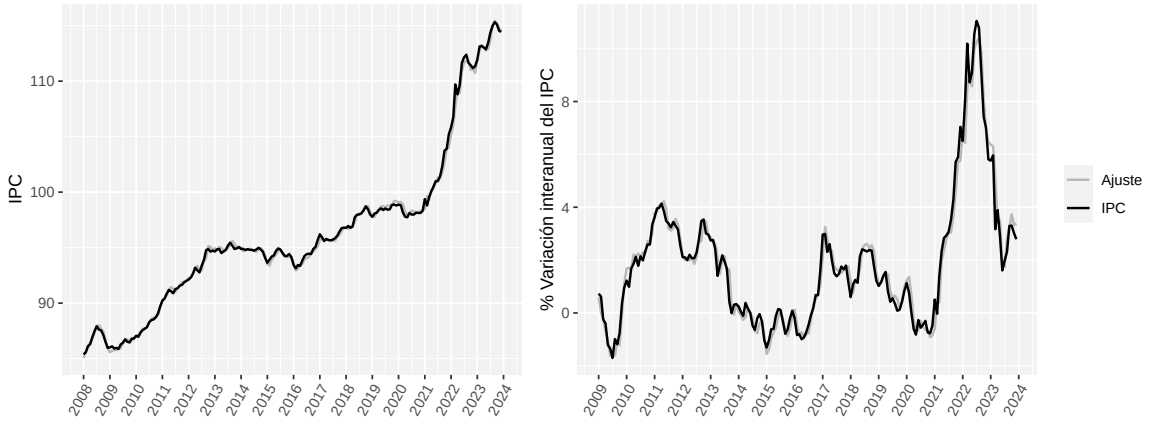

Figura 7.4: Series del IPC reconstruido y de sus variaciones interanuales, junto con el ajuste resultante del modelo $VAR(4)$ por componentes.

7.1.4. IRF y FEVD

Otro tema interesante a abordar cuando se trata con modelos VAR son las funciones impulso-respuesta (IRF) y la descomposición de la varianza del error de predicción (FEVD), herramientas que resultarán de utilidad a la hora de analizar impactos.

Como ya se adelantaba en el Capítulo 3, las funciones de impulso-respuesta sirven para cuantificar el impacto de una variable endógena sobre otra a cierto horizonte h. Así, a modo de ejemplo, se incluyen en la Figura 7.5 las funciones impulso-respuesta con variable de impulso el IPC de productos energéticos a horizonte $h = 12$ (un año).

Se puede observar que un *shock* en el IPC de productos energéticos produce un impacto permanente de crecimiento en el resto de componentes en los próximos doce meses, mientras que, en el caso de la propia componente de productos energéticos tiene un efecto transitorio que tiende a cero a partir de horizontes superiores a 7 meses.

Por ejemplo, en el caso del IPC de alimentos sin elaboración, se podría interpretar como que un aumento de un punto en el logaritmo del IPC de productos energéticos supondría, tras un mes, un aumento de 0.05 en el logaritmo del IPC de alimentos sin elaboración. Esto se traduce en que un incremento del 172 % en el IPC de productos energéticos,

$$e^{log(\texttt{PROD\_ENERG})+1} = 2.72 \cdot \texttt{PROD\_ENERG},$$

implica una subida de un 5 % en el IPC de alimentos sin elaboración,

$$e^{log(\texttt{ALIMENTOS\_SINc})+0.05} = 1.05 \cdot \texttt{ALIMENTOS\_SINc}.$$

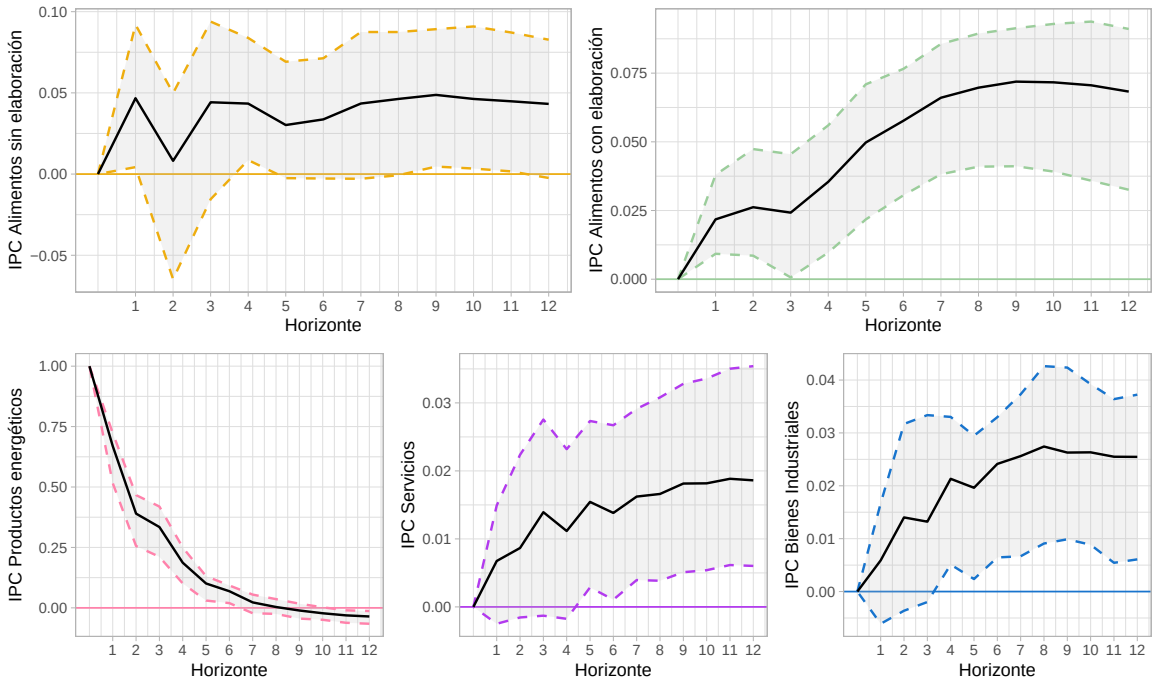

Figura 7.5: Funciones de impulso-respuesta con impulso el IPC de productos energéticos con el modelo $VAR(4)$ ajustado.

En la misma línea, la descomposición de la varianza del error permite estudiar cuál es la proporción de la varianza del error al predecir una variable endógena del modelo a horizonte h que se debe a otra variable endógena.

En la Figura 7.6 se recoge dicha descomposición, en la que se puede ver que la mayoría de dicha varianza se debe a la propia variable en todos los casos. Aún así, se observa que a medida que aumenta el horizonte de predicción, la proporción debida al resto de las variables va aumentando.

Además, por ejemplo, en el caso del IPC de alimentos elaborados, se puede apreciar que, además del propio IPC de alimentos elaborados, también depende en gran medida de los alimentos sin elaboración y de los productos energéticos, lo cual es una relación lógica en el sentido económico.

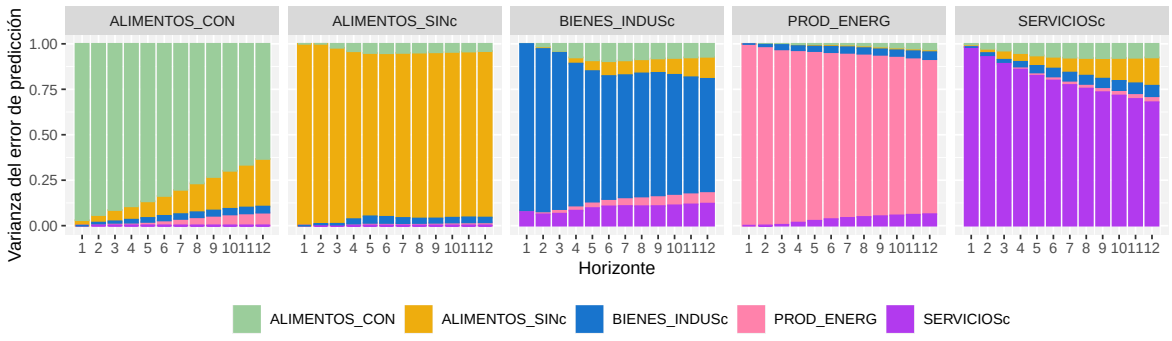

Figura 7.6: Descomposición de la varianza del error del modelo $VAR(4)$.

7.2. Modelo VECM

Tal y como se vio en el capítulo referente al análisis exploratorio de los datos, las series de tiempo de las componentes del IPC no son estacionarias, si no que son integradas de orden 1, $I(1)$. Este hecho hace que el modelo $VAR(4)$ ajustado en la sección anterior no sea estable. En efecto, si analizamos los autovalores de la matriz $\hat{\boldsymbol{A}}$ de la expresión del modelo VAR dada por (3.3), se obtiene que

$$\begin{aligned}|\boldsymbol{\lambda}| = (&1.01, 1.00, 0.96, 0.92, 0.89, 0.89, 0.81, 0.75, 0.63, 0.63,\\ &0.60, 0.60, 0.53, 0.53, 0.44, 0.34, 0.34, 0.33, 0.33, 0.17),\end{aligned}$$

es decir, existen autovalores cuyo valor absoluto es mayor que uno, lo cual es coherente con que las series sean $I(1)$.

A pesar de que las series no son estacionarias, se ajustó un modelo VAR sobre las series transformadas únicamente por logaritmos. Podría parecer razonable haber ajustado el modelo sobre las series diferenciadas mediante una diferencia regular, para que así fuesen estacionarias. No obstante, de ese modo se perdía mucha información sobre las relaciones entre las variables, e igualmente, se encontró un modelo VAR válido para los logaritmos de las componentes.

Ahora bien, dadas las características de los datos, podría resultar más adecuado ajustar un modelo VECM. Para confirmar esta idea, antes de nada debemos comprobar si las series del IPC, que ya sabemos que son $I(1)$, son también cointegradas.

Para comprobarlo, se recurre al test de *Phillips-Ouliaris* (Phillips and Ouliaris, 1990), implementado a partir de la función `po.test` de la librería `tseries` (Trapletti and Hornik, 2023), con hipótesis nula que la variable $\boldsymbol{y}_t$ no esté cointegrada.

Aplicando el test sobre la serie multivariante de los logaritmos de las componentes del IPC se obtiene que el estadístico de contraste del test es de -68.15, con un p-valor asociado de 0.01. En consecuencia, hay evidencias estadísticamente significativas para rechazar la hipótesis nula del test, concluyendo que efectivamente la serie de logaritmos de las componentes está cointegrada. Entonces, se está en las condiciones idóneas para ajustar un modelo VECM.

7.2.1. Selección

Partiendo de lo antes estudiado, como el modelo más adecuado era un $VAR(4)$ y el modelo VECM se define para un orden $p-1$, se ajustará un $VECM(3)$.

Por otra parte, como ya se ha comentado, los modelos VECM cuentan a mayores con relaciones de cointegración, luego será necesario seleccionar el rango de cointegración más adecuado. Para ello, se utiliza el test de Johansen (1995), que se lleva a cabo con la función `ca.jo` de la librería `urca` (Pfaff, 2022).

Mencionar que, de forma paralela al estudio anterior, se comienza incluyendo en el modelo las componentes del IPC como variables endógenas y todas las variables exógenas consideradas en los modelos de regresión del Capítulo 6. Con dichas variables, se selecciona el rango de cointegración y se ajusta el modelo, observándose que el índice FAO y las pernoctaciones en España no son significativas. Así, se decide ajustar el modelo con variables exógenas el precio de las gasolinas con descuento y el precio de la electricidad, descartando las otras dos variables.

En la Tabla 7.4, se recogen los resultados del test del rango de cointegración, considerando este último conjunto de variables. A la vista de los resultados, se rechaza la hipótesis de nula de no cointegración para $r = 0$, $r \leq 1$ y $r \leq 2$, pero no se rechaza para a lo sumo 3 ecuaciones de cointegración. Por

tanto se toma $r = 3$, es decir, se considera un modelo $VECM(3)$ con $r = 3$ ecuaciones de cointegración.

	Estadístico	**Cuantiles**		
		0.1	0.05	0.01
$\boldsymbol{r \leq 4}$	0.87	6.50	8.18	11.65
$\boldsymbol{r \leq 3}$	15.45	12.91	14.90	19.19
$\boldsymbol{r \leq 2}$	25.86	18.90	21.07	25.75
$\boldsymbol{r \leq 1}$	51.75	24.78	27.14	32.14
$\boldsymbol{r = 0}$	102.27	30.84	33.32	38.78

Tabla 7.4: Estadísticos y cuantiles teóricos asociados al test del rango de cointegración de Johansen (1995) para el modelo $VECM(3)$.

7.2.2. Validación

La fase de validación de un modelo VECM es análoga a la de los modelos VAR. Así, se llevan a cabo los mismos contrastes sobre los residuos del modelo, cuyos resultados se recogen en la Tabla 7.5.

	Estadístico	**p-valor**
Breusch-Godfrey	1000.00	0.13
LM-ARCH	2549.24	0.15
Jarque-Bera	915.68	0.00

Tabla 7.5: Estadísticos y p-valores resultantes de la fase de validación de los residuos del modelo $VECM(3)$.

Observando los p-valores de los test, se puede ver que no hay evidencias en contra de la incorrelación y homocedasticidad de los residuos, pero sí en contra de la normalidad de los mismos. De nuevo, esto no es problemático, puesto que la gaussianidad de los residuos es una propiedad deseable pero no es estrictamente necesaria para la validez estadística del modelo.

7.2.3. Estimación y ajuste

En cuanto a la estimación de los coeficientes del modelo VECM, igual que en el caso del VAR, asciende a un número muy elevado. Así, a modo ilustrativo, se incluye la estimación de las ecuaciones de corrección de error o ecuaciones de cointegración y la estimación relativa al IPC de alimentos sin elaborar. Se utiliza la notación dada por (7.1) y a mayores, se considera

$$\boldsymbol{x}_t = \begin{pmatrix} x_{1,t} \\ x_{2,t} \end{pmatrix} = \begin{pmatrix} \log(\texttt{PREC\_GASOLINA\_DESC}) \\ \log(\texttt{ELECTRICIDADc}) \end{pmatrix}.$$

Así, la ecuación estimada del modelo para el IPC de alimentos sin elaboración es

$$\begin{aligned} \Delta y_{1,t} = &- 2.718 - 0.148ec_{1,t} - 0.151ec_{2,t} + 0.015ec_{3,t} - 0.009x_{1,t} - 0.002x_{2,t} \\ &- 0.099\Delta y_{1,t-1} - 0.031\Delta y_{2,t-1} + 0.056\Delta y_{3,t-1} - 0.127\Delta y_{4,t-1} + 0.198\Delta y_{5,t-1} \\ &- 0.349\Delta y_{1,t-2} + 0.658\Delta y_{2,t-2} - 0.009\Delta y_{3,t-2} - 0.087\Delta y_{4,t-2} + 0.871\Delta y_{5,t-2} \\ &- 0.320\Delta y_{1,t-3} - 0.201\Delta y_{2,t-3} + 0.046\Delta y_{3,t-3} - 0.225\Delta y_{4,t-3} + 0.632\Delta y_{5,t-3}, \end{aligned} \tag{7.2}$$

donde las ecuaciones de corrección de error o de cointegración vienen dadas por

$$\boldsymbol{ec}_t = \begin{pmatrix} ec_{1,t} \\ ec_{2,t} \\ ec_{3,t} \end{pmatrix} = \boldsymbol{\beta}\boldsymbol{y}_{t-1} = \begin{pmatrix} 1.00 & 0.00 & 0.00 & -1.41 & -1.67 \\ 0.00 & 1.00 & 0.00 & -0.62 & -2.16 \\ 0.00 & 0.00 & 1.00 & -0.87 & 1.47 \end{pmatrix} \cdot \begin{pmatrix} y_{1,t-1} \\ y_{2,t-1} \\ y_{3,t-1} \\ y_{4,t-1} \\ y_{5,t-1} \end{pmatrix}.$$

Nótese que la estimación de los coeficientes que acompañan a los términos de corrección de error coincide con los correspondientes elementos de la matriz $\boldsymbol{\alpha}$ definida en la teoría. Estos miden cómo cambia la variable endógena correspondiente cuando hay una desviación en el período inmediatamente anterior.

En particular, en el caso de los alimentos sin elaborar, vemos que el signo de los coeficientes asociados a dichos términos es negativo en los dos primeros casos, lo que se traduce en que cuando $ec_{1,t} > 0$ y $ec_{2,t} > 0$ es previsible que el cambio en el IPC de alimentos sin elaboración sea negativo durante el período siguiente.

A partir de la ecuación estimada (7.2), también se pueden observar las relaciones entre el cambio del IPC de alimentos sin elaboración con el cambio de las diferentes componentes en los $p-1 = 3$ períodos anteriores, dadas por los coeficientes estimados. Por ejemplo, atendiendo a la ecuación del modelo para los alimentos sin elaboración, se podría decir que si el IPC de productos energéticos aumentó en el período anterior, el IPC de alimentos sin elaboración en este período aumentaría, puesto que el coeficiente que relaciona ambas tasas de cambio es positivo.

Por otra parte, en la Figura 7.7 se representan los valores ajustados del modelo para cada una de las componentes, junto con las series reales de índices corregidas de estacionalidad.

Asimismo, en la Figura 7.8, se representa el IPC reconstruido a partir de las series corregidas, tanto en niveles como en variaciones interanuales, junto con el ajuste resultante de considerar la contribución de cada una de las componentes del IPC ajustadas con el modelo $VECM(3)$.

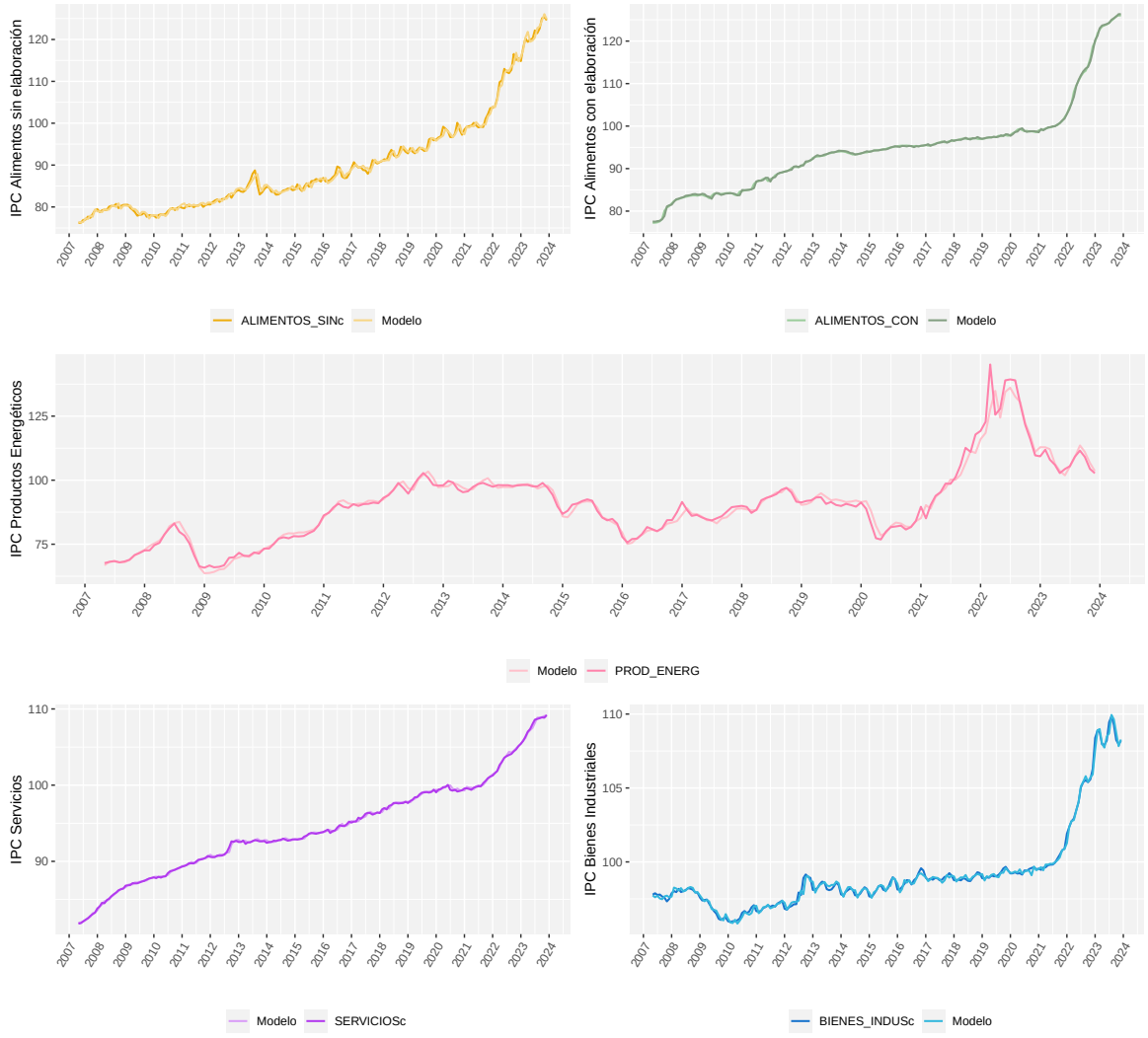

Figura 7.7: Series reales y ajustes del modelo VECM.

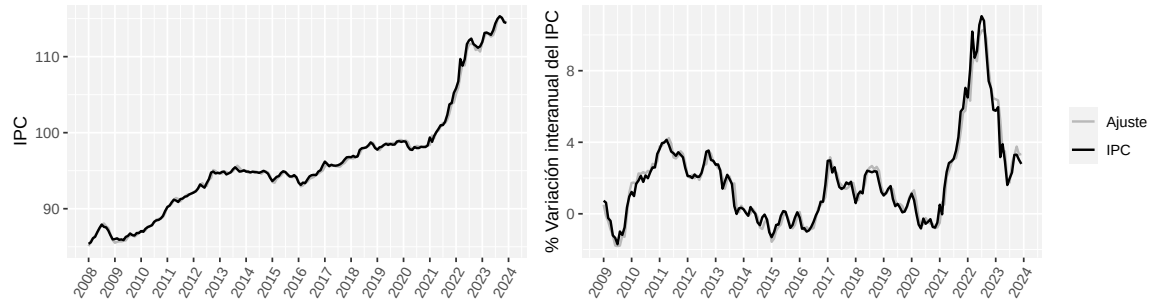

Figura 7.8: Series del IPC reconstruido y de sus variaciones interanuales, junto con el ajuste resultante del modelo $VECM(3)$ con 3 ecuaciones de cointegración por componentes.

7.2.4. IRF y FEVD

Por último, igual que con el modelo VAR, puede ser interesante disponer de las funciones impulso-respuesta y de la descomposición de la varianza del error de predicción para un horizonte dado con el modelo VECM, con el fin de llevar a cabo análisis de impactos.

En la Figura 7.9 se incluye la representación de las funciones impulso-respuesta con variable de impulso el IPC de productos energéticos a horizonte un año, en similitud con el caso de los modelos VAR.

El efecto que se observa ante un *shock* en los productos energéticos (un aumento de una unidad en el logaritmo del IPC de productos energéticos) es muy similar al que se veía con el modelo anterior. Por un lado, produce un aumento en el IPC del resto de componentes que se mantiene con el tiempo, un comportamiento esperable desde el sentido económico de los precios, al estar los productos energéticos relacionados con el resto de precios indirectamente. Mientras que, en el caso del propio IPC de productos energéticos, el efecto es transitorio, yéndose a cero a partir de horizontes superiores a 9 meses.

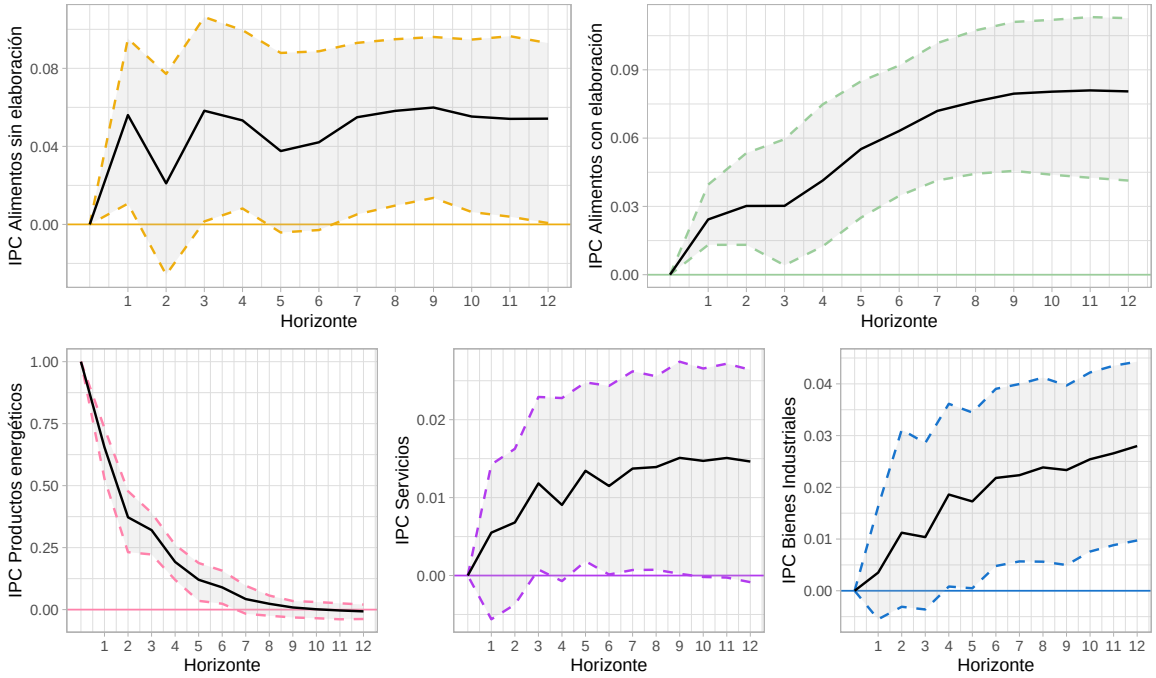

Figura 7.9: Funciones de impulso-respuesta con impulso el IPC de productos energéticos con el modelo $VECM(3)$ con 3 ecuaciones de cointegración ajustado.

En cuanto a la descomposición de la varianza del error, recogida en la Figura 7.10, si bien se observa un comportamiento similar al que se veía con el modelo VAR ajustado, parece que la contribución a la varianza del error de predicción de las demás variables al predecir otra variable endógena es superior en este caso.

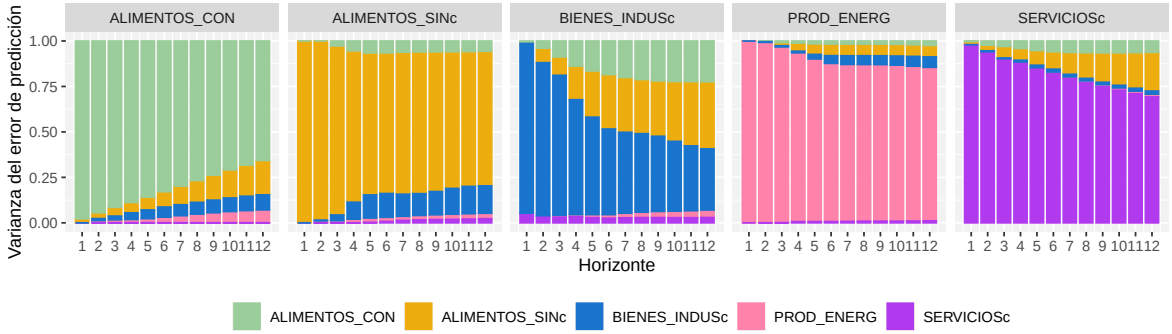

Figura 7.10: Descomposición de la varianza del error del modelo VECM.

Capítulo 8

Redes neuronales LSTM para el IPC

Como última alternativa para abordar el problema de ajustar y predecir el IPC y sus componentes, se recurre a un modelo multivariante de redes neuronales LSTM.

Se inicia este capítulo introduciendo en la Sección 8.1 cómo implementar este tipo de redes en `R`, al no ser modelos directamente disponibles en las librerías habituales. Hay que tener en cuenta que, a diferencia de los modelos anteriores, en el contexto de las redes neuronales es necesario un preprocesamiento de los datos implicados, lo cual se trata en la Sección 8.2.

A continuación, en la Sección 8.3 se selecciona un modelo de redes neuronales LSTM con variables respuesta todas las componentes del IPC e incluyendo a su vez variables explicativas de entrada. Además, se explica el proceso seguido para la selección del modelo atendiendo a los hiperparámetros implicados.

Tras esto, como es habitual en los modelos de aprendizaje estadístico, en la Sección 8.4 se evalúa el modelo con una muestra de entrenamiento y una de test y, por último, se muestra el ajuste del modelo escogido en la Sección 8.5.

8.1. Implementación de las redes LSTM

En el caso de las redes neuronales LSTM, es necesario recurrir a dos herramientas destacadas en el ámbito del *deep learning*: TensorFlow y Keras. Por un lado, **TensorFlow** es una biblioteca de código abierto que facilita el acceso a diferentes algoritmos y modelos para abordar problemas de aprendizaje automático, mientras que, **Keras** es una API (Interfaz de Programación de Aplicaciones) de redes neuronales en lenguaje *Python* (Van Rossum and Drake Jr, 1995), que permite acceder a las herramientas de TensorFlow.

En este contexto, para poder acceder a este tipo de herramientas en `R`, se necesitan instalar varios paquetes. Por un lado, la integración entre `R` y *Python* se consigue con el paquete `reticulate` (Ushey et al., 2024). Por otra parte, el paquete `keras` (Allaire and Chollet, 2023) proporciona una interfaz en la que nos aprovecharemos de los beneficios de `R`, pero con la capacidad de *Python* y el paquete `tensorflow` (Allaire and Tang, 2024) contiene los algoritmos necesarios para la implementación de los modelos. El procedimiento necesario para la instalación en `R` es el que sigue.

```
install.packages("reticulate")
library(reticulate)

install.packages("tensorflow")
tensorflow::install_tensorflow()

install.packages("keras")

library(tensorflow)
library(keras)
```

Nótese que para poder instalar TensorFlow con la función `install_tensorflow` es necesario tener instalada alguna versión de Anaconda (Anaconda Software Distribution, 2020).

8.2. Preprocesamiento de los datos

En primer lugar, cabe hablar sobre las variables consideradas. Igual que en los casos anteriores, se toman como variables respuesta las cinco componentes del IPC y, como variables explicativas todas las variables exógenas consideradas en los capítulos anteriores, junto con las componentes del IPC retardadas 12 meses. En este caso, por analogía con los casos anteriores, se consideran las series transformadas por logaritmos y adecuadamente corregidas de estacionalidad.

Ahora bien, para ajustar una red neuronal correctamente es necesario que todos los datos de entrada se encuentren en la misma escala, así pues, antes de nada, será necesario aplicar alguna transformación en los datos. En este punto surgen dos opciones:

- **Normalizar las series**. Consiste en reescalar las series de forma que todas las observaciones estén entre 0 y 1. Para ello, se aplica la transformación

$$X_t' = \frac{X_t - \text{mín}\,(X_t)}{\text{máx}\,(X_t) - \text{mín}\,(X_t)} \in [0,1]\,.$$

- **Estandarizar las series**. Se reescalan las series de forma que la media de los valores observados sea 0 y la desviación típica 1, para ello basta centrar los datos restándoles su media y dividirlos por su desviación típica,

$$X_t' = \frac{X_t - \mu_{X_t}}{\sigma_{X_t}}.$$

Una vez que los datos tienen la escala adecuada, como es habitual en los problemas de aprendizaje estadístico, se consideran dos muestras, una de entrenamiento, que servirá para entrenar los posibles modelos y otra de test, para validar si los modelos entrenados tienen un buen ajuste. En particular, se reservan 12 meses, correspondientes con el año 2023, como muestra de test, constituyendo el resto de observaciones la muestra de entrenamiento.

Para elegir qué transformación de las mencionadas es más adecuada, se prueban ambas y se analiza con cuál se obtienen mejores resultados. En este caso, se realiza el proceso de selección del modelo aplicando ambas transformaciones y se observa que el error cuadrático medio tanto en la muestra de entrenamiento como en la de test es menor si se utiliza la estandarización.

Para terminar con el preprocesamiento de los datos, es importante tener en cuenta su formato. Las series que se toman como datos de entrada en el modelo deberán ser *arrays* con tres dimensiones (al estar los modelos implementados en lenguaje *Python*), recogiendo la longitud de las series de tiempo, el número de muestras y el número de variables.

8.3. Selección del modelo

Para seleccionar el modelo más adecuado se ha seguido un proceso en varias etapas:

1. Una **etapa de entrenamiento**, en la que se entrena la red neuronal para las posibles combinaciones de los hiperparámetros implicados en la misma.
2. Una **etapa de evaluación**, en la que se obtienen predicciones de los modelos entrenados con la muestra de test y así, se evalúa qué combinación de hiperparámetros proporciona un mejor ajuste.

En la primera de las etapas, el primer paso es fijar los posibles valores para cada uno de los hiperparámetros, teniendo en cuenta sus implicaciones. Por un lado, hay que determinar el número de neuronas o nodos de la capa oculta de la red. Cuantas más neuronas, la red será más poderosa, sin embargo, al aumentar el número de neuronas aumenta el número de parámetros a estimar y en consecuencia, el tiempo de computación, luego es importante llegar a un equilibrio.

Además, es necesario seleccionar el tamaño de los *batch* (*batch size*) y el número de *epochs*. El *batch size* hace referencia al número de observaciones que se incluyen en la red neuronal para que se entrene en cada iteración. Entonces, que dicho tamaño sea pequeño se traduce en que la red tenga en memoria pocos datos y el entrenamiento sea rápido, mientras que, en el caso contrario se permite a la red tener más datos en memoria, llevando eso a un proceso de entrenamiento más costoso. En cuanto al número de *epochs*, se refiere al número de veces que pasa cada *bacth* por la red neuronal. De nuevo, si se toma un número muy elevado, puede llevar a un sobreajuste (*overfitting*) y si es demasiado pequeño, puede haber infraajuste (*underfitting*) [1].

Adicionalmente, debe escogerse un valor para la tasa de aprendizaje (*learning rate*), que indica con qué velocidad se actualizan los pesos del modelo en cada *batch*; y, opcionalmente, para la tasa de abandono (*dropout rate*). En los modelos de redes neuronales LSTM cabe la posibilidad de añadir una capa de *dropout*. Esta sirve para evitar el sobreajuste, al ir desconectando neuronas con cada entrenamiento. Lo que indica el valor del *dropout rate* es el porcentaje de neuronas que se deshabilitan en cada entrenamiento de la red neuronal.

Ahora, teniendo en cuenta estas consideraciones, se recogen en la Tabla 8.1 los posibles valores para los hiperparámetros implicados que se han tomado para el ajuste de la red.

Hiperparámetro	**Posibles valores**
Neuronas	{32, 64, 96, 128, . . . , 512}
Batch size	{8, 16, 32, 64, 128}
Epochs	{10, 20, 40, 80, 100}
Learning rate	{0.01, 0.001, 0.0001}
Dropout rate	{0.1, 0.2, 0.3}

Tabla 8.1: Posibles valores considerados para los hiperparámetros implicados en la red neuronal LSTM.

[1] El *overfitting* y *underfitting* son dos problemas muy comunes en el entrenamiento de redes neuronales. El primero ocurre cuando el algoritmo de aprendizaje se sobreentrena con ciertos datos, provocando dificultades en el ajuste cuando se le introducen nuevos datos al modelo. A su vez, el *underfitting* se da en la situación opuesta, cuando el modelo cuenta con un entrenamiento muy pobre, dificultando la obtención de resultados correctos.

Como ya se explicaba en el Capítulo 4, además de los hiperparámetros en las redes neuronales intervienen una función de activación y una función de pérdida. Mencionar que en este caso se han considerado la función tangente hiperbólica ($tanh$) y el error cuadrático medio ($RMSE$), respectivamente.

A su vez, para compilar el modelo es necesario especificar el algoritmo de optimización. En este caso se recurre al algoritmo de optimización de Adam, que es una extensión del método de descenso de gradiente propuesto por Kingma and Ba (2014).

Ahora bien, para la selección de los hiperparámetros se siguen dos vías. En primer lugar, el número de neuronas y las tasas de aprendizaje y abandono se seleccionan recurriendo al paquete `kerastuneR` (Abdullayev, 2024). Esta librería simplifica el complejo proceso de selección de hiperparámetros óptimos para redes basadas en TensorFlow y Keras, al tener implementados los algoritmos de búsqueda aleatoria, optimización bayesiana y búsqueda mediante hiperbanda.

Por otro lado, en cuanto al *batch size* y el número de *epochs*, se decide recurrir a una búsqueda en cuadrícula (*grid search*), de forma que se crea una malla de todas las posibles combinaciones de estos hiperparámetros y se entrena el modelo para cada caso.

Una vez entrenados los modelos, contaríamos con un modelo con un número óptimo de neuronas, de *learning rate* y de *dropout rate* para cada configuración de *bacth size* y *epochs*. Para determinar cuál es el mejor, se tienen en cuenta dos criterios. Por un lado, se evalúan los modelos ajustados en la muestra de test, con el fin de minimizar el error cuadrático medio en dicha muestra. Y, como criterio de selección adicional se tiene en cuenta la lógica económica de los resultados, en el sentido de que, por ejemplo, un modelo que prediga que el IPC (en términos de variaciones interanuales) en 2024 es negativo no encaja con la situación actual.

Con todo esto, el modelo seleccionado se corresponde con una red neuronal con 192 neuronas, una tasa de aprendizaje de 0.01 y de abandono de 0.3, empleando 100 *epochs* y un tamaño de *batch* de 16 observaciones.

8.4. Evaluación en la muestra de entrenamiento y de test

Como se decía, para seleccionar los hiperparámetros del modelo se siguió una combinación de dos criterios, minimizar el error cuadrático medio en la muestra de test y la lógica económica, resultando el modelo ya especificado.

En este contexto, en la Figura 8.1 se puede ver como va variando el error cuadrático medio en cada paso del entrenamiento del modelo, tanto en la muestra de entrenamiento como en la de test. Por un lado, como es de esperar, el error en la muestra de entrenamiento es próximo a 0 a lo largo de todo el proceso. Y, por otro, el error en la muestra de test, si bien supera al error de entrenamiento en todos los casos, parece que se mantiene en niveles bastante bajos. Finalmente, resulta que el modelo escogido presenta un error cuadrático medio de 0.01 en la muestra de entrenamiento, que se eleva hasta 0.05 cuando se evalúa sobre la muestra de test.

Ahora bien, para interpretar la magnitud de estos errores hay que tener en cuenta la escala que se está manejando. Antes de ajustar el modelo, se realizó un preprocesamiento de los datos, estandarizando los mismos al restarles su media y dividirlos por su desviación típica. Así, para tener una idea real del error que comete el modelo en términos del IPC, cabe transformar las predicciones del modelo en cada una de las muestras a la escala original del IPC y compararlas con los datos reales, calculando el error cuadrático medio en esta escala.

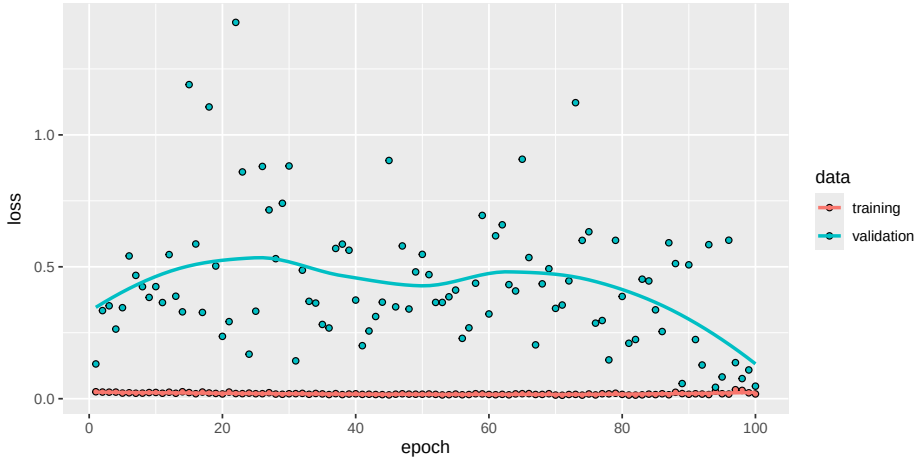

Figura 8.1: Error cuadrático medio en la muestra de entrenamiento y de test en cada paso (*epoch*) del entrenamiento del modelo.

Aplicando dicha transformación, se obtienen los errores recogidos en la Tabla 8.2. Si bien es cierto que son superiores a los obtenidos previamente, son bastante razonables dada la escala en la que estamos trabajando.

Puede verse que en la muestra de test seleccionada, el modelo comete un mayor error a la hora de predecir la componente de alimentos sin elaborar y se aproxima mucho más a la realidad en el caso de los servicios o los bienes industriales.

	Muestra de entrenamiento	**Muestra de test**
ALIMENTOS_SINc	2.03	24.88
ALIMENTOS_CON	1.49	5.16
PROD_ENERG	5.90	7.44
SERVICIOSc	0.39	0.60
BIENES_INDUSc	0.18	0.81
Media	2.00	7.78

Tabla 8.2: Error cuadrático medio en la muestra de entrenamiento y de test con la red neuronal LSTM escogida.

8.5. Ajuste del modelo

Una vez seleccionado el modelo utilizando las muestras de entrenamiento y de test, se ajusta el modelo a la serie completa con el fin de después predecir.

En paralelo con los capítulos anteriores, representamos en la Figura 8.2 las correspondientes series de las componentes del IPC corregidas, junto con los valores ajustados de la red neuronal.

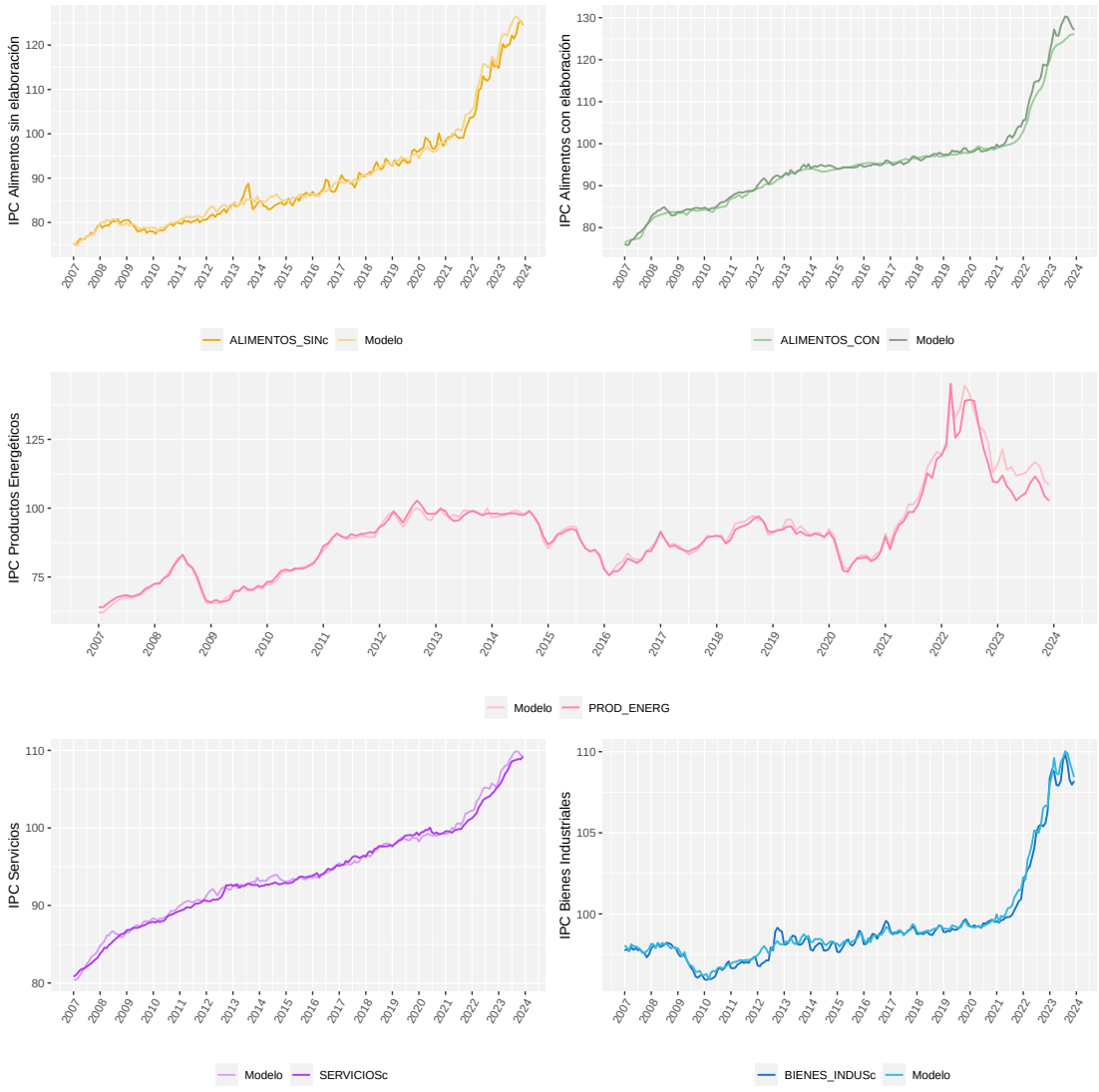

Figura 8.2: Series reales y ajustes del modelo de redes neuronales $LSTM$ seleccionado.

Además, igual que antes, en la Figura 8.3 se representa el IPC general reconstruido a partir de las componentes corregidas, en niveles y en variaciones interanuales, incluyendo el ajuste del mismo, resultante del cálculo de las repercusiones de cada una de las componentes ajustadas con la red neuronal.

Nótese que a diferencia de los modelos propuestos en los capítulos anteriores, en los que se mostraba algún ejemplo de ecuación estimada del modelo y su interpretación a partir de los coeficientes estimados, en este caso no es posible. Solo disponemos de la estimación de los pesos, pero estos carecen de interpretación.

En general, si bien las redes neuronales suelen ser buenas a la hora de predecir, no cuentan con una interpretación de las relaciones subyacentes en el modelo. Esto se debe a su complejidad y el gran número de conexiones involucradas en las mismas.

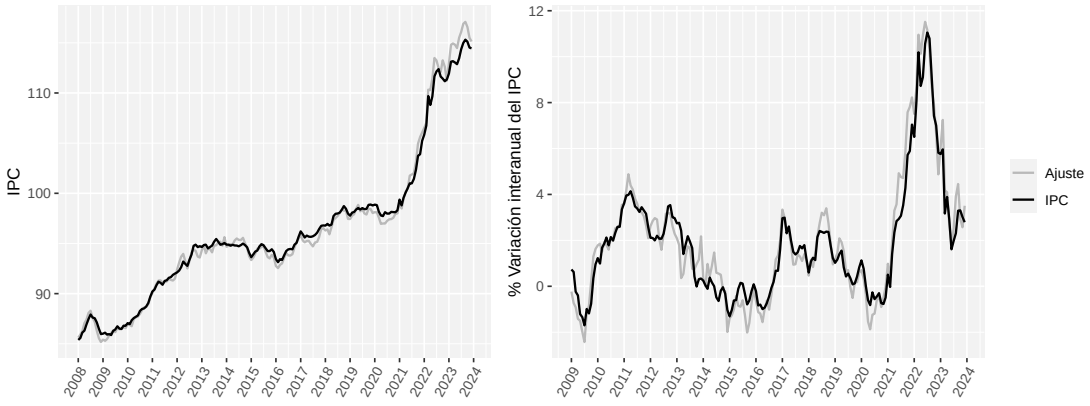

Figura 8.3: Series del IPC reconstruido y de sus variaciones interanuales junto con el ajuste resultante del modelo $LSTM$ por componentes.

Capítulo 9

Comparativa y rendimiento de los modelos ajustados

Tras ajustar los distintos modelos propuestos, además de predecir el IPC y sus componentes durante el año 2024, resulta de interés comparar dichas predicciones con otras fuentes que realizan proyecciones macroeconómicas, así como con los datos reales disponibles. Así, en la Sección 9.1 se recogen las predicciones de los modelos ajustados, junto con dichas comparativas. Además, se realiza un ejercicio de *backtesting*, para validar la calidad y precisión de los modelos.

Por otro lado, como ya se adelantaba, con el fin de analizar el rendimiento de los modelos, es importante estudiar su sensibilidad ante impactos. En la Sección 9.2, se observa cómo afecta un *shock* en los precios del petróleo sobre el IPC y sus distintas componentes, lo que permite conocer cómo de sensibles son los modelos ante cambios en las variables regresoras y como de útil es el precio del petróleo en la simulación de impactos.

9.1. Rendimiento de los modelos ajustados

En los capítulos anteriores se presentaron los modelos seleccionados en cada uno de los casos, mostrando los resultados de la validación y ajuste de cada uno de ellos, pero sin obtener predicciones, objetivo primordial de este estudio. Esto se hizo con el fin de mostrar las predicciones de todos los modelos de manera simultánea y así poder llevar a cabo comparaciones entre las mismas.

Así, a continuación se incluyen las predicciones del IPC general y a su vez, como apoyo para evaluar la calidad predictiva de los modelos, se incluyen los resultados de un ejercicio de *backtesting*.

9.1.1. Predicciones

A partir de todos los modelos ajustados y fijando el horizonte de predicción a $h = 12$ meses, se obtienen las previsiones para el año 2024 de cada una de las componentes del IPC, adecuadamente transformadas. Con estas predicciones, se construyen las predicciones del IPC general, que se pueden ver representadas en las Figuras 9.1 y 9.2, tanto en niveles como en variaciones interanuales, y distinguiendo cada uno de los modelos empleados. Los resultados concretos de las predicciones, resumidos de forma anual, pueden consultarse en la Tabla 9.1, que comentaremos más tarde.

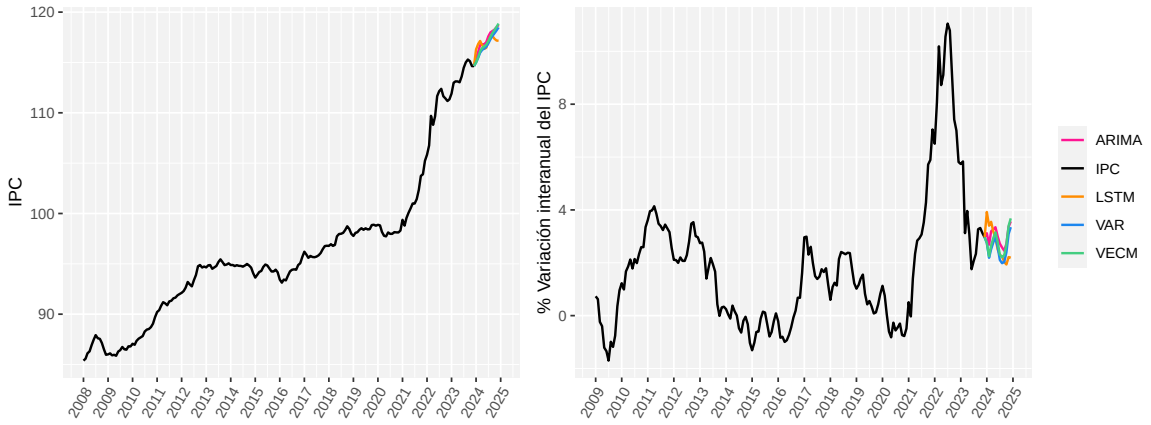

Figura 9.1: Serie del IPC general junto con las predicciones obtenidas con los distintos modelos ajustados en niveles (a la izquierda) y en porcentaje de variación interanual (a la derecha).

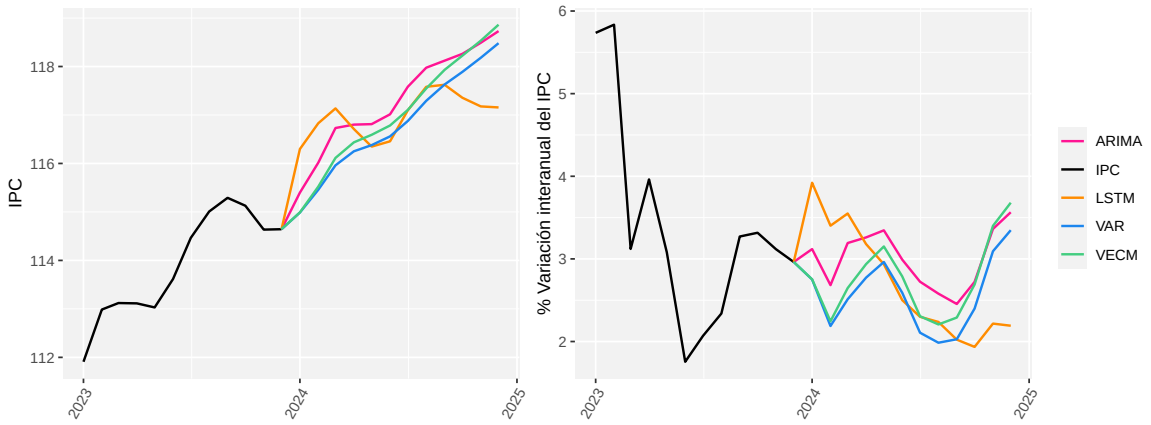

Figura 9.2: Serie del IPC general desde 2023 junto con las predicciones obtenidas con los distintos modelos ajustados en niveles (a la izquierda) y en porcentaje de variación interanual (a la derecha).

En este punto, además de las predicciones en si mismas, interesa saber si las previsiones propuestas están en línea con las previsiones que dan otros organismos o bien, con los datos reales ya disponibles.

Comparativa con otras fuentes

Existen diversas fuentes que elaboran proyecciones macroeconómicas de la economía española, y en particular, del IPC. Así, con el fin de evaluar el rendimiento de los modelos ajustados en términos predictivos, puede ser interesante comparar las predicciones obtenidas con los modelos descritos a lo largo del trabajo con las predicciones publicadas por algunas de dichas fuentes.

En la Tabla 9.1 se incluyen las predicciones del IPC por componentes y del IPC general en términos de variaciones interanuales para el 2024, por un lado obtenidas mediante los modelos estadísticos

propuestos y, por otro, publicadas por el Banco de España (Banco de España, 2024), Bankinter (Departamento de análisis Bankinter, 2024), la Fundación de las Cajas de Ahorros (Funcas, 2024) y Caixa Bank Research (Martín Vilató, 2024); y se puede ver que, efectivamente, las predicciones de los modelos se encuentran en línea con las predicciones de las otras cuatro fuentes.

	Predicciones 2024							
	ARIMA	**VAR**	**VECM**	**LSTM**	**FUNCAS**	**BdE**	**Caixa Bank**	**Bankinter**
IPC Alimentos sin elaboración	5.6	5.7	6.4	5.9	7.1	-	-	-
IPC Alimentos con elaboración	3.0	3.5	4.7	5.0	4.7	-	-	-
IPC Productos energéticos	1.7	0.8	2.3	1.6	1.8	-	-	-
IPC Servicios	3.3	3.0	2.4	2.6	3.4	-	-	-
IPC Bienes industriales	2.0	0.6	0.9	0.3	0.4	-	-	-
IPC general	3.0	2.6	2.8	2.7	3.2	2.7	3.0	3.3

Tabla 9.1: Predicciones del IPC y de sus componentes de grupos especiales para 2024 en variaciones interanuales obtenidas. por un lado. con los modelos ajustados y por otro, a partir de otras fuentes.

Observación 9.1. *De las cuatro fuentes incluídas, tan solo FUNCAS publica predicciones desglosadas por componentes.*

Comparativa con los datos reales disponibles

De forma adicional, como los modelos se ajustan sobre series temporales de frecuencia mensual y se han considerado datos hasta diciembre de 2023, se pueden comparar las predicciones de los meses del primer trimestre de 2024 con los datos reales del IPC, al ya estar estos publicados.

En la Tabla 9.2 se recogen las predicciones del IPC general en términos de variaciones interanuales de los modelos para los meses de enero a abril, junto con los datos reales publicados por el INE. Se observa que las predicciones de los modelos univariantes por componentes son las que más se aproximan a los datos reales hasta el momento.

	ARIMA	**VAR**	**VECM**	**LSTM**	**Reales**
Enero 2024	3.1	2.8	2.8	3.9	3.4
Febrero 2024	2.7	2.2	2.2	3.4	2.8
Marzo 2024	3.2	2.5	2.6	3.5	3.2
Abril 2024	3.3	2.8	2.9	3.2	3.3

Tabla 9.2: Predicciones del IPC para el primer trimestre del 2024 en variaciones interanuales obtenidas con los modelos ajustados y los correspondientes datos reales.

9.1.2. *Backtesting*

El *backtesting* es un procedimiento estadístico que sirve para validar la calidad o la precisión de un modelo. Consiste en considerar datos históricos y recortarlos hasta cierto período, con el fin de ajustar los correspondientes modelos seleccionados con esos datos y comparar la predicción de los modelos en el período recortado frente a los datos reales en esos instantes de tiempo.

En este caso, se tomaron las series históricas recortadas hasta septiembre de 2023 y se usaron los meses de octubre, noviembre y diciembre para el *backtesting*. En el caso del modelo de pernoctaciones, los meses que se usaron para calcular esta medida fueron octubre, noviembre y diciembre de 2019.

En la Tabla 9.3 se recogen los porcentajes de variación de la media de las predicciones de cada uno de los modelos en esos tres meses frente a la media de los datos reales en estos mismos meses. Puede verse que en general los modelos cometen errores pequeños, siendo el modelo de redes neuronales LSTM en el caso del IPC de productos energéticos el que presenta una mayor desviación en los meses seleccionados y el modelo VAR en el IPC de alimentos con elaboración el que más se aproxima a la realidad.

	ARIMA	**VAR**	**VECM**	**LSTM**
IPC Alimentos sin elaborar	−0.18	−1.08	0.26	−2.73
IPC Alimentos con elaboración	0.14	0.04	0.67	−0.82
IPC Productos energéticos	1.05	6.27	4.97	−6.97
IPC Servicios	0.25 [1]	0.49	0.23	1.27
IPC Bienes industriales	0.83	1.04	0.65	−0.13

Tabla 9.3: Porcentajes de variación de la media de las predicciones en los meses de octubre, noviembre y diciembre de los modelos recortados hasta septiembre de 2023 frente a la media de los datos reales en esos tres meses.

9.2. Análisis de sensibilidad a impactos

Un ejercicio importante relacionado con los modelos estadísticos, y en especial en el ámbito macroeconómico, es medir la sensibilidad de los mismos ante cambios en las variables explicativas.

En particular, interesa que los modelos sean sensibles a *shocks* en los precios energéticos, ya que estos están estrechamente vinculados a las tensiones geopolíticas. El precio del petróleo ha venido presentando una volatilidad muy elevada en los últimos años, lo que afecta directa o indirectamente al resto de precios del mercado.

Desde la entidad es de gran importancia saber a qué niveles de inflación nos iríamos con un determinado precio del petróleo porque inflaciones muy altas reducen tanto el nivel de consumo, como el de ahorro. Luego resulta interesante analizar si los modelos seleccionados son sensibles ante cambios en el precio del petróleo.

[1] En el caso de los modelos de regresión dinámica (ARIMA), se consideró un modelo para el IPC de restaurantes y hoteles y otro para el resto de servicios. El 0.25 % representa la media de esta medida de *backtesting* entre ambos modelos, siendo un 0.11 % en el caso del IPC de restaturantes y hoteles y un 0.39 % en el IPC del resto de servicios.

Para ello, en primer lugar, se realiza un ejercicio de estrés de los precios, incrementando las hipótesis de alto nivel realizadas para el precio del petróleo en distinta medida, y se analiza cómo afecta a los precios de los carburantes. En la Tabla 9.4 se recogen algunos resultados relativos a este estudio.

Incremento en el precio del petróleo	**Incremento en el precio de los carburantes**
+10 %	+3.2 %
+50 %	+16.2 %
+100 %	+41.2 %

Tabla 9.4: Aumento en las predicciones del precio de los carburantes en 2024 suponiendo que el precio del petróleo es un 10 %, 50 % y 100 % superior a las hipótesis de alto nivel fijadas para dicho precio.

Más concretamente, se consideraron precios del petróleo un 10 %, un 50 % y un 100 % superiores a las hipótesis fijadas para los meses de 2024. Estos incrementos se tradujeron en un incremento del 3.2 %, 16.2 % y 41.2 %, respectivamente, en el precio de los carburantes, en comparación con los precios resultantes del escenario de partida.

A continuación, con los precios de los carburantes estresados, se estudió el impacto sobre las diferentes componentes del IPC y sobre el IPC general con cada uno de los modelos. En la Tabla 9.5 se recogen los resultados correspondientes con los tres posibles escenarios planteados.

Antes de analizar los resultados es importante notar que algunas variables explicativas entran en los modelos con cierto decalaje, como veíamos en los capítulos anteriores. Así, por ejemplo en el caso del modelo de regresión dinámica para el IPC de alimentos sin elaboración, en el que entraban los productos energéticos con un retraso de 10 meses, será difícil ver resultados de este impacto en términos anuales, al estar afectando el *shock* únicamente a los meses de noviembre y diciembre de 2024. Al entrar los productos energéticos con 10 meses de decalaje, los datos correspondientes con los meses de enero a octubre de 2024 serían datos reales (corresponderían con los meses de marzo a diciembre de 2023).

	+10 % en el precio del petróleo				**+50 % en el precio del petróleo**				**+100 % en el precio del petróleo**			
	ARIMA	**VAR**	**VECM**	**LSTM**	**ARIMA**	**VAR**	**VECM**	**LSTM**	**ARIMA**	**VAR**	**VECM**	**LSTM**
Alimentos sin elaboración	5.6	5.9	6.6	6.3	5.5	6.6	7.4	7.6	5.3	7.8	8.7	9.3
Alimentos con elaboración	3.0	3.7	4.8	5.4	3.0	4.2	5.4	6.6	3.0	5.2	6.4	8.1
Productos energéticos	4.9	3.5	5.0	2.6	17.9	14.5	16.1	6.1	42.8	35.1	37.0	12.2
Servicios	3.3	3.0	2.5	2.8	3.3	3.1	2.6	3.3	3.3	3.3	2.9	3.9
Bienes industriales	2.0	0.7	1.0	0.5	2.0	1.0	1.3	1.1	2.0	1.5	1.8	1.9
IPC general	3.3	2.9	3.1	3.0	4.6	4.2	4.5	4.0	7.0	6.7	7.0	5.4
Diferencia frente a base	0.3	0.3	0.3	0.3	1.6	1.6	1.7	1.3	4.0	4.1	4.2	2.7

Tabla 9.5: Predicciones del IPC y de sus componentes de grupos especiales para 2024 en variaciones interanuales suponiendo que el precio del petróleo es un 10 %, 50 % y 100 % superior a las hipótesis de alto nivel fijadas por los expertos del área para dicho precio en 2024.

En términos generales, se podría decir que todos los modelos ajustados son sensibles a impactos en el precio del petróleo, ya que el IPC aumenta en todos los casos de manera significativa.

En efecto, si se duplicara el precio del petróleo previsto para 2024 (+100 % en el precio del petróleo),

conllevaría una subida de 4 puntos porcentuales en la inflación según los modelos de regresión dinámica; un incremento de 4.1 puntos porcentuales con el modelo VAR; una subida de 4.2 puntos porcentuales de acuerdo con el modelo VECM; y, un aumento de 2.7 puntos según la red neuronal LSTM. De ese modo, se estaría alcanzando una inflación media en 2024 de entre un 5.4 y un 7 %, frente al 2.6-3 % previsto, según el modelo considerado.

Cabe notar que los impactos sobre el IPC general son similares en todos los casos, siendo el modelo de redes neuronales LSTM algo menos sensible que el resto.

Sin embargo, componente a componente, se puede ver que mientras que los modelos de regresión dinámica concentran la mayoría del impacto sobre el IPC de productos energéticos, en los modelos VAR, VECM y LSTM el impacto se distribuye más entre las diferentes componentes. Posiblemente, esto se debe por un lado a lo que comentábamos del retraso en las variables explicativas que entran en los modelos y, por otro, a que los modelos multivariantes recogen mejor las interacciones entre las distintas componentes.

Capítulo 10

Aplicación para el análisis de la inflación

Como se ha visto, a lo largo del trabajo se han desarrollado distintos modelos para predecir la inflación en España y analizar impactos, haciendo uso de `R`. Ahora bien, con el fin de llevar a cabo análisis de estos resultados por parte de los especialistas del área económica del banco, resulta de interés crear algún tipo de herramienta accesible para cualquiera, independientemente de sus conocimientos de programación en `R`.

Con este objetivo, se ha desarrollado una aplicación web interactiva en la que se incluyen todos los resultados del trabajo mediante el paquete `Shiny` (Chang et al., 2023). La estructura básica de este tipo de aplicaciones se introduce en la Sección 10.1 y, posteriormente en la Sección 10.2 se describen algunos detalles de la aplicación desarrollada. Finalmente, en la Sección 10.3 se explica como compartir esta aplicación para poder ejecutarla sin necesidad de tener `R` instalado en el equipo.

10.1. Estructura de construcción básica de una aplicación en Shiny

Las aplicaciones `Shiny` permiten crear una interfaz gráfica interactiva en la que se muestran los resultados de un código de `R`. Para ello, emplean un sistema de programación reactivo que permite que el usuario escoja ciertos valores de entrada, denominados *inputs* (que pueden ir cambiando), y se generen ciertos valores de salida, denominados *outputs* (que se actualizan de manera inmediata).

Todas las aplicaciones `Shiny` tienen la misma estructura, que viene dada por tres componentes fundamentales que interactúan entre ellas:

- Un **objeto de interfaz de usuario** (`ui`), que controla el diseño y la apariencia de la aplicación. Contiene las instrucciones de lo que se debe mostrar en la misma y recibe los *inputs* que introduce el usuario.
- Una **función de servidor** (`server`), que contiene el código `R` que hace que la aplicación funcione. Recibe los valores de entrada y los transforma mediante código en *outputs*, que envía a la interfaz del usuario. Como los valores de entrada pueden cambiar, al poder ser modificados por el usuario, se dice que los elementos del servidor son reactivos.

- Una **llamada a la función** `shinyApp`, para crear la aplicación.

El código de la aplicación se desarrolla en un *script* de `R` dentro de algún directorio, en el que también se incluyen otros archivos adicionales que puedan resultar necesarios para la ejecución de la misma. Dicho código, como mínimo debería incluir lo siguiente:

```
# Cargamos la librería
library(shiny)

# Creamos la interfaz gráfica
ui <- ui = fluidPage(...)

# Creamos el servidor
server <- function(input, output){...}

# Creamos la aplicación
shinyApp(server=server, ui=ui)
```

Conocido el esquema básico de funcionamiento de estas aplicaciones, solo falta indagar en todas las posibilidades y estructuras que se pueden construir. Por un lado, los *inputs* de la aplicación se crean con funciones `*Input()` en la parte de la interfaz del usuario (por ejemplo, `sliderInput()`, `selectInput()`, ...), cuya sintaxis viene dada por:

```
*Input(inputId=" ", label=" ", ...)
```

El argumento `inputId` fija el nombre con el que se va a llamar a dicho *input* para las distintas operaciones en el servidor, mientras que `label` hace referencia a la etiqueta que se le muestra al usuario. Además de estos dos argumentos, se incluyen distintas opciones específicas de cada *input*.

Por otra parte, para mostrar un *output* se utiliza una función `*Output()` (como `plotOutput()`, `DataTableOutput()`, ...). Estos *outputs* deben definirse en la parte del servidor:

```
server <- function(input, output){
  output$nombre<- #code
}
```

Pueden crearse mediante una función de tipo `*render()`, asociada al tipo de *output* deseado (como `renderPlot()`, `renderDataTable()`, ...), que sirve para construir un *output* reactivo que se muestra en la interfaz. Asimismo, para acceder a los *inputs* desde el servidor, se utiliza la notación `input$`.

Para más detalles acerca de las posibles funciones tanto de *input* como de *output* y un desarrollo más extenso acerca de la construcción de aplicaciones `Shiny`, puede consultarse Wickham (2021).

10.2. Aplicación desarrollada para la inflación

Como ya se adelantaba, el objetivo de la aplicación es brindar la posibilidad de que cualquiera pueda acceder a los resultados del presente trabajo e incluso interactuar, sin necesidad de utilizar código `R`.

La aplicación desarrollada permite al usuario fijar las hipótesis de las variables de entrada de los diferentes modelos y ver cómo estas impactan sobre la predicción de la inflación y de cada una de sus

componentes. Asimismo, incluye todos los detalles de cada uno de los modelos ajustados a lo largo del trabajo, así como una pestaña con la comparativa de todos los modelos y sus previsiones. En la Figura 10.1 puede verse una captura de pantalla de la pestaña de inicio de la aplicación, que cuenta con un menú de navegación que permite acceder a los diferentes resultados.

Figura 10.1: Pestaña de inicio de la aplicación.

En la primera de las pestañas, mostrada en la Figura 10.2, se encuentran los detalles de las variables consideradas, tanto de las propias componentes de la inflación como de las variables regresoras utilizadas en los modelos. Se muestran los datos de dichas variables recogidos en tablas, así como en gráficos interactivos que permiten ver la evolución de las mismas. Además, se cuenta con una pestaña en la que se pueden descargar los datos utilizados en una hoja de cálculo de Excel.

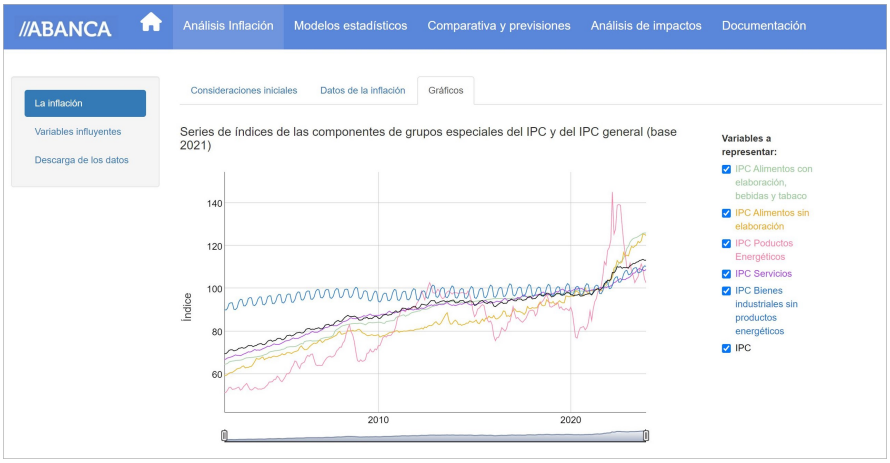

Figura 10.2: Pestaña de análisis de la inflación de la aplicación.

Al contar con datos actualizados, esto permite al analista estudiar de forma muy visual las información más reciente.

En la Figura 10.3 se muestra una captura de la siguiente pestaña. En ella se recogen los detalles de todos los modelos ajustados, acompañados por los resultados de las correspondientes validaciones y los gráficos de ajuste de los modelos.

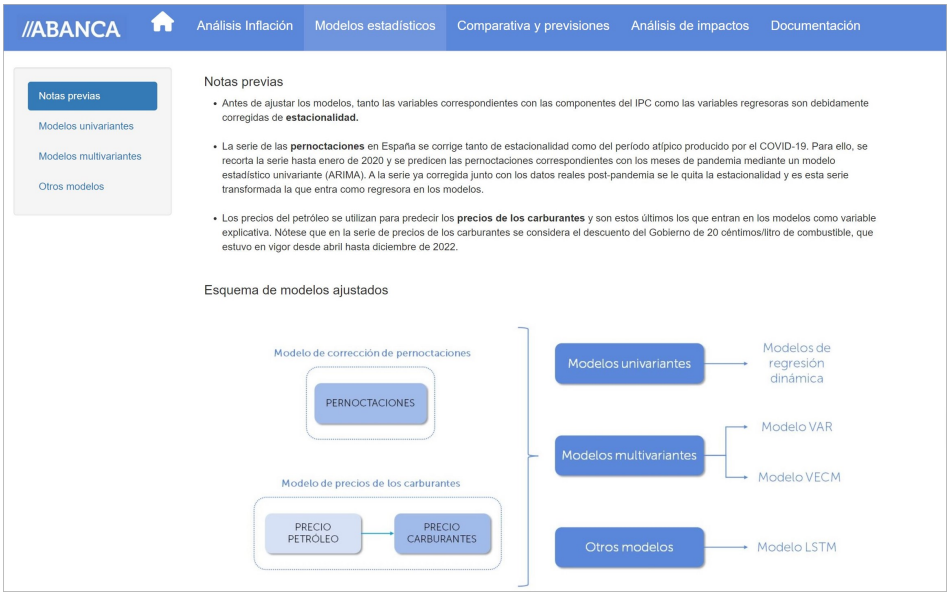

Figura 10.3: Pestaña de los modelos de la aplicación.

Para continuar, en la siguiente pestaña (Figura 10.4) se engloban los resultados de previsiones y comparativa de las mismas. Se indican las hipótesis de alto nivel fijadas sobre las variables regresoras, con la posibilidad de descargar un archivo Excel con las mismas; se muestran las predicciones de cada uno de los modelos en el año 2024, tanto númerica como gráficamente; se comparan con las previsiones de otros organismos, así como con los datos reales disponibles hasta el momento; y se muestran los resultados del ejercicio de *backtesting*.

Figura 10.4: Pestaña de comparativa y previsiones de la aplicación.

La pestaña relativa al análisis de impactos es quizás la más interactiva para el usuario. En esta se permite que el usuario decida cuánto varía el precio del petróleo frente al escenario base fijado y vea automáticamente cómo afecta al precio de los carburantes y a las componentes del IPC, así como al IPC general. Asimismo incluye la posibilidad de subir un archivo de hipótesis, similar al incluído en la parte de hipótesis de alto nivel, con las hipótesis que quiera sobre las variables regresoras y vea qué impacto produce sobre la inflación. Puede verse una captura de la misma en la Figura 10.5.

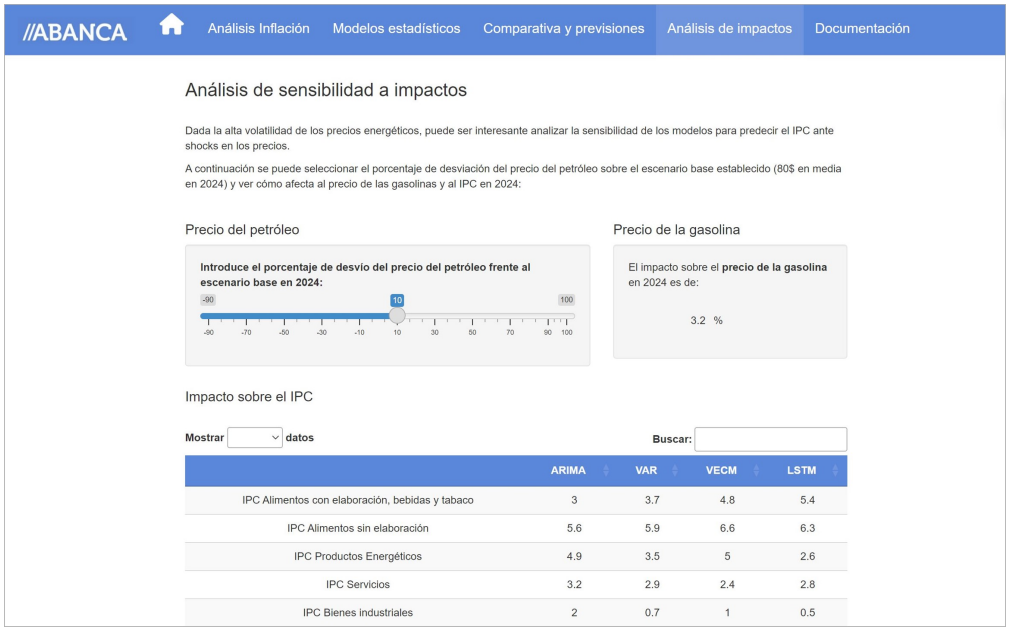

Figura 10.5: Pestaña de análisis de impactos de la aplicación.

Por último, en la pestaña de documentación, se recoge la presente memoria, como apoyo teórico a los modelos presentados.

Esta aplicación interactiva constituye una herramienta muy útil para la entidad. No solo permite a los analistas económicos del banco realizar ejercicios de análisis, sino que también les otorga mucha autonomía al llevar a cabo simulaciones y analizar las predicciones de los modelos.

10.3. Ejecución de la app en cualquier ordenador

A priori, al haber creado la aplicación con el paquete `shiny`, es necesario tener `R` instalado en el ordenador para poder visualizar y utilizar la herramienta.

En este caso, para ABANCA suponía una desventaja que solo pudiera ser ejecutada en ordenadores con `R` instalado. El objetivo primordial de la aplicación es que los analistas económicos del banco que no tienen conocimientos de programación puedan acceder a los resultado del proyecto y analizarlos con autonomía, sin embargo, no disponen de este *software* en su equipo.

Así, para resolver este problema, surgió la idea de desarrollar una solución que permita distribuir la aplicación vía `zip`, siendo ejecutable en cualquier ordenador, sin importar lo que tenga instalado. Para ello, se recurrió al proyecto `DektopDeployR`, desarrollado por Pang, Lee (2020).

En concreto, los pasos seguidos para crear la aplicación de escritorio fueron:

1. Descargar el repositorio de `DektopDeployR` en local en una carpeta con el nombre de la aplicación. En este caso será "`AppIPC`".
2. Instalar R-Portable en la carpeta "`/AppIPC/dist/`". R-Portable es un proyecto de código abierto que sirve para utilizar `R` sin instalarlo en el ordenador.
3. Guardar en la carpeta "`/AppIPC/app/shiny`" todos los scripts y archivos necesarios para su funcionamiento. Como mínimo será necesario incluir los *scripts* `server.R` y `ui.R`, con las función de servidor y la interfaz gráfica definidas para configurar la aplicación, y un *script* `global.R` con el código necesario para los modelos y datos implementados en la misma.
4. Editar el archivo "`/AppIPC/app/app.R`", para llamar correctamente a nuestra aplicación:

```
shiny::runApp("./app/shiny", launch.browser=TRUE)
```

5. Especificar las dependencias de paquetes, editando el archivo "`/AppIPC/app/packages.txt`". Añadimos en cada línea el nombre de cada paquete necesario para ejecutar la aplicación.

Con esto, ya tenemos en el escritorio nuestra aplicación, que podemos compartir mediante un archivo `zip`. Para ponerla en marcha, basta ejecutar el archivo "`/AppIPC/appname.exe`" y se abrirá en el navegador del ordenador. Se puede ver una captura de los elementos de la carpeta en la que se recoge la aplicación en la Figura 10.6.

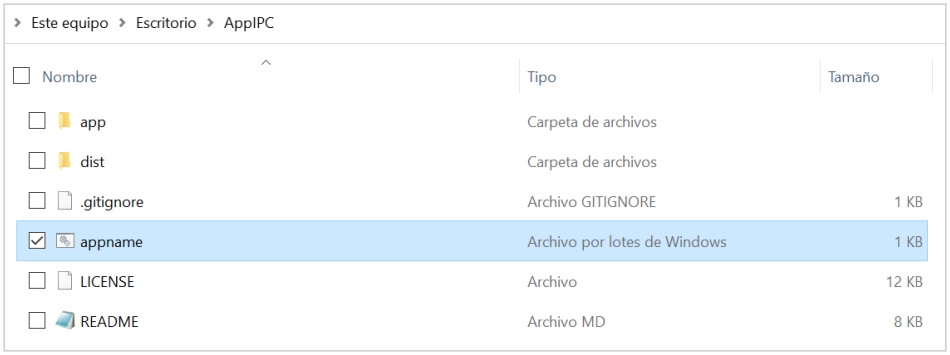

Figura 10.6: Captura de pantalla de la carpeta "`AppIPC`" con la aplicación para el análisis de la inflación lista para ser ejecutada o compartida a cualquier usuario.

Gracias a esto, la aplicación diseñada puede ser compartida con cualquier analista económico de ABANCA, permitiéndole analizar los datos actualizados de la inflación y demás variables implicadas en los modelos, así como los diferentes resultados desarrollados en el proyecto (detalles de los modelos, predicciones de los mismos, análisis de sensibilidad interactivo, ...).

Capítulo 11

Conclusiones y líneas futuras

Para finalizar la memoria de este proyecto, en este capítulo se presentan las conclusiones del estudio realizado, derivadas de los resultados incluidos en el Capítulo 9.

En particular, en la Sección 11.1 se revisan los resultados obtenidos a lo largo del proyecto, comparando las metodologías utilizadas para abordar el problema propuesto. Y, por último, en la Sección 11.2 se recogen diferentes líneas de trabajo futuras que surgen con el desarrollo del mismo.

11.1. Conclusiones

El objetivo de este trabajo es proponer metodologías que permitan a ABANCA predecir la inflación en España y cuantificar impactos. Para ello, se proponen cuatro metodologías diferentes en cuanto a la explicabilidad de sus parámetros y predictibilidad de las mismas, algunas de las cuales están estrechamente relacionadas.

En primer lugar, se han considerado modelos univariantes de regresión dinámica para cada una de las cinco componentes de grupos especiales del IPC, logrando resultados satisfactorios. Por un lado, sus predicciones encajan con las previsiones de otros organismos, así como con la realidad en lo que llevamos de año y con las opiniones de los expertos en el área de Planificación Estratégica y PMO de ABANCA. En media, con estos modelos se obtiene que el IPC general para 2024 será de un 3 %.

Además, los modelos son estadísticamente válidos, pues sus residuos tienen media nula, son estacionarios y están incorrelados. Presentan buenos resultados de *backtesting*, al no superar sus predicciones variaciones del 1.5 % con respecto a los datos reales, en términos absolutos, en ninguna de las componentes. Y son suficientemente sensibles a impactos, llegando a producir un incremento de 4 puntos porcentuales en el IPC general, ante un escenario en el que se duplique el precio del petróleo.

A continuación, motivado por la correlación entre las componentes del IPC ajustadas, se utilizaron modelos multivariantes, que permitieron la interacción de todas las variables entre sí. En este punto surgieron dos enfoques complementarios. Por un lado, los modelos VAR y VECM, que son la extensión multivariante de los modelos univariantes desarrollados, y por otro, los modelos de redes neuronales LSTM, que forman parte de las técnicas de aprendizaje automático.

De nuevo, en los tres casos se obtuvieron resultados muy buenos, con modelos estadísticamente correctos, que proporcionan predicciones que están en línea con las de otras entidades y con la propia realidad. En el caso de los modelos VAR y VECM, sus residuos superan los contrastes de incorrelación

y homocedasticidad y, en lo que se refiere a la red LSTM, el error cuadrático medio del modelo al considerar una muestra de entrenamiento y otra de test es razonable. Relativo a las predicciones, en términos anuales, se obtuvo una variación interanual del 2.6 % con el modelo VAR, 2.8 % con el modelo VECM y 2.7 % con las redes LSTM, para el IPC general en el año 2024.

En cuanto a los resultados de *backtesting* y sensibilidad ante impactos, los tres responden de forma adecuada, aunque la red LSTM ajustada resulta levemente peor en estos aspectos. En términos absolutos, los modelos VAR y VECM presentan un *backtesting* que no supera el 6.3 % para ninguna de las componentes y elevan el IPC general hasta 4.1 y 4.2 puntos porcentuales, respectivamente, si se duplica el precio del petróleo previsto para 2024. Asimismo, la red LSTM propuesta presenta un *backtesting* inferior al 7 % en todas las componentes, y en el caso de que el petróleo duplique su precio en 2024 (frente al escenario de precio del petróleo previsto), produce un incremento de 2.7 puntos porcentuales sobre el IPC general del año.

Atendiendo a estos resultados, podría decirse que cualquiera de las metodologías es adecuada para abordar el problema propuesto, al ser las cuatro estadísticamente correctas y producir resultados que concuerdan con la lógica económica. No obstante, dada la importancia de la interpretabilidad de los resultados para la entidad y la sensibilidad a impactos, los modelos clásicos (modelos de regresión dinámica, VAR y VECM) resultan mejores para esta necesidad.

Así como los modelos univariantes y los modelos VAR y VECM nos permiten conocer las correspondencias entre las variables implicadas en los mismos, permitiéndonos analizar si las relaciones subyacentes entre ellas tienen sentido desde un punto de vista económico, en el caso de las redes LSTM no ocurre lo mismo. Con las redes no somos capaces de estudiar su funcionamiento en cuanto a las interacciones entre variables. Además, vimos que la red escogida no es tan sensible como el resto de modelos, habiendo una diferencia de hasta 2.2 puntos en el impacto que se produce sobre el IPC general al duplicar el precio del petróleo.

Por último, señalar que además de las metodologías desarrolladas, se ha creado una herramienta web basada en *Shiny* que va a poder ser utilizada por cualquier analista del banco. Se trata de una aplicación interactiva, que permite estudiar de manera muy visual, tanto la información actualizada de las variables implicadas en este proyecto, como los resultados de los modelos ajustados. Por estar el área de Planificación Estratégica de ABANCA constituida por un equipo multidisciplinar, esta herramienta aporta un valor adicional significativo al proyecto al permitir el análisis de los resultados sin conocimientos específicos de programación.

11.2. Líneas futuras

A partir de la realización de este Trabajo de Fin de Máster han surgido ciertas líneas futuras a tener en cuenta.

En relación con el objetivo del mismo, predecir la inflación en España y analizar impactos, surge desarrollar estas mismas ideas para el caso concreto de Galicia, al ser uno de los focos principales de negocio de la entidad.

En lo referido a los datos empleados para los modelos, por un lado, además de las variables económicas seleccionadas como influyentes, podría explorarse la inclusión de otras variables relevantes. Por ejemplo, el precio de otras materias primas o algún indicador de tensiones geopolíticas.

Por otro lado, en lo que se refiere a la corrección de las series previa a la modelización, cabe estudiar otras vías. Se propone utilizar el *software* `JDemetra+` a partir de su interfaz disponible para `R` (Sax and Eddelbuettel, 2018), que permite corregir las series con las herramientas de TRAMO-SEATS y

X12-ARIMA/X13-ARIMA-SEATS, ampliamente utilizadas por organismos oficiales, como Eurostat (2015), IGE (2022) o INE (2019).

En cuanto a las metodologías escogidas, se considera desarrollar el mismo estudio empleando otro tipo de modelos que hayan sido satisfactorios, de acuerdo con el estado del arte de modelos para predecir la inflación. Por ejemplo, podría ser interesante comparar el rendimiento de los modelos propuestos en este proyecto con un modelo BVAR o con un *random forest*.

Bibliografía

Abdullayev, T. (2024). *kerastuneR: Interface to 'Keras Tuner'*. R package version 0.1.0.7. https://CRAN.R-project.org/package=kerastuneR.

Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Akaike, H. (1998). Information Theory and an Extension of the Maximum Likelihood Principle. In *Selected Papers of Hirotugu Akaike*, pages 199–213. Springer.

Allaire, J. and Chollet, F. (2023). *keras: R Interface to 'Keras'*. R package version 2.13.0.

Allaire, J. and Tang, Y. (2024). *tensorflow: R Interface to 'TensorFlow'*. R package version 2.16.0. https://CRAN.R-project.org/package=tensorflow.

Allen, R. G. D., Fürst, G. M. W., Roy, R., Loftus, P. J., Castellano, V., Barberi, B., and Khamis, S. H. (1963). Price Index Numbers. *Revue de l'Institut International de Statistique / Review of the International Statistical Institute*, 31(3):281–306.

Anaconda Software Distribution (2020). *Anaconda Documentation*. Anaconda Inc. https://docs.anaconda.com/.

Aneiros, G. (2022). *Series de tiempo*. Apuntes de la asignatura, Universidade da Coruña.

Atil, A., Lahiani, A., and Nguyen, D. K. (2014). Asymmetric and nonlinear pass-through of crude oil prices to gasoline and natural gas prices. *Energy Policy*, 65:567–573.

Banco de España (2024). Proyecciones macroeconómicas de la economía española (2024-2026). https://www.bde.es/f/webbe/SES/Secciones/Publicaciones/InformesBoletinesRevistas/BoletinEconomico/24/T1/Fich/be2401-it-Proy.pdf. Accedido 26 de mayo de 2024.

Barkan, O., Benchimol, J., Caspi, I., Cohen, E., Hammer, A., and Koenigstein, N. (2023). Forecasting CPI inflation components with hierarchical recurrent neural networks. *International Journal of Forecasting*, 39(3):1145–1162.

Benalal, N., Diaz del Hoyo, J. L., Landau, B., Roma, M., and Skudelny, F. (2004). To aggregate or not to aggregate? Euro area inflation forecasting. *Euro Area Inflation Forecasting (July 2004)*.

Bergstra, J. and Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*, 13(2).

Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. (2015). *Time Series Analysis: Forecasting and Control*. John Wiley & Sons.

Brown, R. L., Durbin, J., and Evans, J. M. (1975). Techniques for Testing the Constancy of Regression Relationships Over Time. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 37(2):149–163.

Ceasar, L. (2006). Forecasting swiss inflation using VAR models. *Swiss National Bank Economic Studies*, 2.

Chan, K.-S. and Cryer, J. D. (2008). *Time Series Analysis With Applications in R*. Springer.

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., and Borges, B. (2023). *shiny: Web Application Framework for R*. R package version 1.7.5. https://CRAN.R-project.org/package=shiny.

Chow, G. C. (1960). Tests of Equality Between Sets of Coefficients in Two Linear Regressions. *Econometrica: Journal of the Econometric Society*, pages 591–605.

Cowpertwait, P. S. and Metcalfe, A. V. (2009). *Introductory Time Series with R*. Springer Science & Business Media.

Dagum, E. B. and Bianconcini, S. (2016). *Seasonal Adjustment Methods and Real Time Trend-Cycle Estimation*. Springer.

Departamento de análisis Bankinter (2024). Previsión IPC España para 2024 y 2025. https://www.bankinter.com/blog/economia/previsiones-ipc-espana. Accedido 26 de mayo de 2024.

Dong, J. (2020). Forecasting modeling for China's inflation.

Doornik, J. A. and Hendry, D. F. (1997). Modelling Dynamic Systems Using PcFiml 9.0 for Windows. *International Thomson Business Press*.

Durán, R., Garrido, E., Godoy, C., and de Dios Tena, J. (2012). Predicción de la inflación en México con modelos desagregados por componente. *Estudios Económicos*, pages 133–167.

ECB (2024). ECB Staff Macroeconomic Projections for the Euro Area. https://www.ecb.europa.eu/pub/projections/html/ecb.projections202403_ecbstaff~f2f2d34d5a.es.html. Accedido 26 de mayo de 2024.

Engle, R. F. and Granger, C. W. (1987). Co-integration and Error Correction: Representation, Estimation, and Testing. *Econometrica: journal of the Econometric Society*, pages 251–276.

Eurostat (2015). *ESS guidelines on seasonal adjustment*. https://ec.europa.eu/eurostat/documents/3859598/6830795/KS-GQ-15-001-EN-N.pdf.

Fernández-Casal, R., Julián, C.-B., and Oviedo, M. (2021). Aprendizaje Estadístico. https://rubenfcasal.github.io/aprendizaje_estadistico/.

Friedman, M. (1937). The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *Journal of the American Statistical Association*, 32(200):675–701.

Funcas (2024). IPC diciembre 2023. Previsiones hasta 2024. https://www.funcas.es/textointegro/ipc-diciembre-2023-previsiones-hasta-diciembre-2024/. Accedido 26 de mayo de 2024.

Godfrey, L. G. (1988). *Misspecification Tests in Econometrics: the Lagrange Multiplier Principle and Other Approaches*. Number 16. Cambridge University Press.

González Mínguez, J., Hurtado, S., Leiva-León, D., and Urtasun, A. (2022). The spread of inflation from energy to other components. *Economic Bulletin/Banco de España, 2023/Q1, 02*.

Granger, C. W. (1969). Investigating Causal Relations by Econometric Models and Cross-Spectral Methods. *Econometrica: journal of the Econometric Society*, pages 424–438.

Granger, C. W. (1981). Some Properties of Time Series Data and Their Use in Econometric Model Specification. *Journal of Econometrics*, 16(1):121–130.

Graves, A. (2013). Generating Sequences with Recurrent Neural Networks. *arXiv preprint arXiv:1308.0850*.

Hamilton, J. D. (2020). *Time Series Analysis*. Princeton University Press.

Hannan, E. J. and Quinn, B. G. (1979). The Determination of the Order of an Autoregression. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(2):190–195.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The Elements of Statistical Learning: Data mining, Inference, and Prediction*, volume 2. Springer.

Hochreiter, S. and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Hosking, J. R. (1980). The Multivariate Portmanteau Statistic. *Journal of the American Statistical Association*, 75(371):602–608.

Hurvich, C. M. and Tsai, C.-L. (1989). Regression and Time Series Model Selection in Small Samples. *Biometrika*, 76(2):297–307.

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E., and Yasmeen, F. (2023). *forecast: Forecasting functions for time series and linear models*. R package version 8.21.1. https://pkg.robjhyndman.com/forecast/.

IGE (2022). *Banco de series de conxuntura. Ciclotendencia e series corrixidas de estacionalidade e calendario*. Instituto Galego de Estatística.

INE (2016). *Índice de Precios de Consumo. Base 2016. Metodología*. Instituto Nacional de Estadística.

INE (2019). *Estándar del INE para la corrección de efectos estacionales y efectos de calendario en las series coyunturales*. Instituto Nacional de Estadística.

INE (2020). *Metodología de cálculo del scanner data en el IPC e IPCA*. Instituto Nacional de Estadística.

INE (2022). *Principales novedades metodológicas del Índice de Precios de Consumo Base 2021*. Instituto Nacional de Estadística.

Jarque, C. M. and Bera, A. K. (1987). A Test for Normality of Observations and Regression Residuals. *International Statistical Review / Revue Internationale de Statistique*, 55(2):163–172.

Johansen, S. (1995). Likelihood-based Inference in Cointegrated Vector Autoregressive Models. *Oxford University Press*.

Kingma, D. P. and Ba, J. (2014). Adam: A method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*.

Kruskal, W. H. and Wallis, W. A. (1952). Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association*, 47(260):583–621.

Lenza, M., Moutachaker, I., and Paredes, J. (2023). Density forecasts of inflation: a quantile regression forest approach.

Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., and Talwalkar, A. (2018). Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization. *Journal of Machine Learning Research*, 18(185):1–52.

Ljung, G. M. and Box, G. E. (1978). On a Measure of Lack of Fit in Time Series Models. *Biometrika*, 65(2):297–303.

López, L., Párraga Rodríguez, S., and Santabárbara, D. (2022). Box 4. The pass-through of higher natural gas prices to inflation in the euro area and in Spain. *Economic Bulletin/Banco de España, 3/2022, p. 49-52*.

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer Science & Business Media.

Lütkepohl, H. and Krätzig, M. (2004). *Applied Time Series Econometrics*. Cambridge University Press.

Maravall, A. (2011). Seasonality Tests and Automatic Model Identification in TRAMO-SEATS. *Bank of Spain: Madrid, Spain*.

Martín Vilató, Z. (2024). InflaciON, inflaciOFF: perspectivas para 2024. https://www.caixabankresearch.com/es/economia-y-mercados/inflacion/inflacion-inflacioff-perspectivas-2024. Accedido 26 de mayo de 2024.

McCulloch, W. S. and Pitts, W. (1990). A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bulletin of Mathematical Biology*, 52:99–115.

Ollech, D. (2021). *seastests: Seasonality Tests*. R package version 0.15.4. https://CRAN.R-project.org/package=seastests.

Ooms, J. (2014). The jsonlite Package: A Practical and Consistent Mapping Between JSON Data and R Objects. *arXiv:1403.2805 [stat.CO]*.

Pang, Lee (2020). DesktopDeployR. https://github.com/wleepang/DesktopDeployR. Accedido 26 de mayo de 2024.

Paranhos, L. (2021). Predicting Inflation with Recurrent Neural Networks. *arXiv preprint arXiv:2104.03757*.

Pascanu, R., Mikolov, T., and Bengio, Y. (2013). *On the Difficulty of Training Recurrent Neural Networks*.

Pfaff, B. (2008a). *Analysis of Integrated and Cointegrated Time Series with R*. Springer Science & Business Media.

Pfaff, B. (2008b). VAR, SVAR and SVEC Models: Implementation Within R Package vars. *Journal of Statistical Software*, 27(4).

Pfaff, B. (2022). *urca: Unit Root and Cointegration Tests for Time Series Data*. R package version 1.3-3. https://CRAN.R-project.org/package=urca.

Phillips, P. C. and Ouliaris, S. (1990). Asymptotic Properties of Residual Based Tests for Cointegration. *Econometrica: journal of the Econometric Society*, pages 165–193.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Said, S. E. and Dickey, D. A. (1984). Testing for Unit Roots in Autoregressive-Moving Average Models of Unknown Order. *Biometrika*, 71(3):599–607.

Sax, C. and Eddelbuettel, D. (2018). Seasonal Adjustment by X-13ARIMA-SEATS in R. *Journal of Statistical Software*, 87(11):1–17.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, pages 461–464.

Shapiro, S. S. and Wilk, M. B. (1965). An Analysis of Variance Test for Normality (complete samples). *Biometrika*, 52(3-4):591–611.

Sharma, S., Sharma, S., and Athaiya, A. (2017). Activation Functions in Neural Networks. *Towards Data Sci*, 6(12):310–316.

Shumway, R. H. and Stoffer, D. S. (2000). *Time Series Analysis and Its Applications*, volume 3. Springer.

Snoek, J., Larochelle, H., and Adams, R. P. (2012). Practical Bayesian Optimization of Machine Learning Algorithms. *Advances in Neural Information Processing Systems*, 25.

Sun, Y., Zhang, X., Hong, Y., and Wang, S. (2019). Asymmetric pass-through of oil prices to gasoline prices with interval time series modelling. *Energy Economics*, 78:165–173.

Torres, J. (2020). *Python Deep Learning: Introducción práctica con Keras y TensorFlow 2*. Alpha Editorial.

Trapletti, A. and Hornik, K. (2023). *tseries: Time Series Analysis and Computational Finance*. R package version 0.10-54. https://CRAN.R-project.org/package=tseries.

Tsay, R. S. (2005). *Analysis of Financial Time Series*. John Wiley & Sons.

Ushey, K., Allaire, J., and Tang, Y. (2024). *reticulate: Interface to 'Python'*. https://rstudio.github.io/reticulate/, https://github.com/rstudio/reticulate.

Van Rossum, G. and Drake Jr, F. L. (1995). *Python tutorial*. Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands.

Welch, B. L. (1951). On the Comparison of Several Mean Values: An Alternative Approach. *Biometrika*, 38(3/4):330–336.

Werbos, P. (1974). New Tools for Prediction and Analysis in the Behavioral Science. *Ph. D. dissertation, Harvard University*.

Wickham, H. (2021). *Mastering shiny*. "O' Reilly Media, Inc.".

Wijffels, J. and Belmans, O. (2023). *taskscheduleR: Schedule R Scripts and Processes with the Windows Task Scheduler*. R package version 1.8. https://CRAN.R-project.org/package=taskscheduleR.

Wold, H. (1938). *A Study in the Analysis of Stationary Time Series*. PhD thesis, Almqvist & Wiksell.

Zellner, A. (1962). An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias. *Journal of the American Statistical Association*, 57(298):348–368.

Zubieta Huaygua, G. (2016). Análisis de los efectos de la inflación en el crecimiento económico: evidencia para la economía boliviana. *Revista de Análisis del Banco Central de Bolivia*, 24:9.