Trabajo Fin de Máster

Predicción en Tiempo Real del PIB de España

Fernando Rebolo García

Máster en Técnicas Estadísticas

Curso 2022-2023

Propuesta de Trabajo Fin de Máster

Título en galego: Predición en Tempo Teal do PIB de España
Título en español: Predicción en Tiempo Real del PIB de España
English title: Real Time Forecasting of Spain's GDP
Modalidad: Modalidad B
Autor/a: Fernando Rebolo García, Universidade da Coruña
Director/a: Rubén Fernández Casal, Universidade da Coruña; Guillermo López Taboada, Universidade da Coruña
Tutor/a: Teresa Veiga Rodríguez, ABANCA; Sergio Díaz Canosa, ABANCA
Breve resumen del trabajo: Predicción del PIB utilizando variables medidas a una frecuencia más alta, para ello las series de tiempo utilizadas se transforman mediante las metodologías de TRAMO-SEATS y Log-diferenciación. Con las series resultantes se construyen y comparan 4 modelos de predicción Bosques aleatorios, Redes neuronales *feedforward* de una sola capa oculta, modelos de regresión *Sparse Group* MIDAS y modelos de regresión MIDAS y restringidos.
Recomendaciones:
Otras observaciones:

Don Rubén Fernández Casal, Profesor contratado Doctor de la Universidade da Coruña, don Guillermo López Taboada, Catedrático de universidad de la Universidade da 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

Predicción en Tiempo Real del PIB de España

fue realizado bajo su dirección por don Fernando Rebolo García 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 14 de [Julio] de 2023.

Firmado digitalmente por FERNANDEZ CASAL RUBEN - 33317724Q
Fecha: 2023.07.14 17:40:20 +02'00'

El director:
Don Rubén Fernández Casal

LOPEZ TABOADA GUILLERMO - 44088774M

Firmado digitalmente por LOPEZ TABOADA GUILLERMO - 44088774M
Nombre de reconocimiento (DN): c=ES, serialNumber=IDCES-44088774M, givenName=GUILLERMO, sn=LOPEZ TABOADA, cn=LOPEZ TABOADA GUILLERMO - 44088774M
Fecha: 2023.07.14 12:12:07 +02'00'

El director:
Don Guillermo López Taboada

VEIGA RODRIGUEZ, TERESA (FIRMA)

Firmado digitalmente por VEIGA RODRIGUEZ, TERESA (FIRMA)
Fecha: 2023.07.14 09:45:16 +02'00'

La tutora:
Doña Teresa Veiga Rodríguez

DIAZ CANOSA SERGIO - 47356028V

Firmado digitalmente por DIAZ CANOSA SERGIO - 47356028V
Fecha: 2023.07.14 09:28:30 +02'00'

El tutor:
Don Sergio Díaz Canosa

El autor:
Don Fernando Rebolo García

Declaración responsable. Para dar cumplimiento a la Ley 3/2022, de 24 de febrero, de convivencia universitaria, referente al plagio en el Trabajo Fin de Máster (Artículo 11, Disposición 2978 del BOE núm. 48 de 2022), **el/la autor/a 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

Agradecer, en primer lugar, a la entidad ABANCA por darme la oportunidad de desarrollar el Trabajo de Fin de Máster en la empresa y aplicar los conocimientos adquiridos lo largo del máster al mundo laboral. En particular, agradecer a Teresa Veiga Rodríguez y Sergio Díaz Canosa el apoyo brindado durante este tiempo. Por otro lado, agradecer al profesorado del Máster en Técnicas Estadísticas por los conocimientos aportados y, en especial, a mis tutores académicos Rubén Fernández Casal y Guillermo López Taboada.

Índice general

Resumen

Resumen en español

El objetivo de este Trabajo de Fin de Máster es obtener una predicción a corto plazo del Producto Interior Bruto (de ahora en adelante PIB) de España, mediante el uso de variables medidas a una frecuencia más alta. Para lograrlo, se emplean las metodologías de TRAMO-SEATS y Log-diferenciación para transformar las series y tratar de eliminar cualquier elemento que distorsione la información que se pueda extraer de ellas como puede ser la tendencia, estacionalidad... Se proponen cuatro modelos de predicción: Bosques Aleatorios, Redes neuronales *feedforward* de una sola capa oculta, Modelos de Regresión *Sparse Group* MIDAS y Modelos de Regresión MIDAS restringidos.

El primer modelo, Bosques Aleatorios, utiliza técnicas de aprendizaje automático basadas en árboles de decisión para realizar la predicción. Este enfoque permite capturar relaciones no lineales y considerar la importancia de cada variable en la predicción del PIB.

El segundo modelo, Redes neuronales *feedforward* de una sola capa oculta, utiliza una arquitectura de red neuronal simple para realizar la predicción. Estas redes son capaces de aprender patrones complejos en los datos y ajustar los pesos de las conexiones entre las neuronas para obtener una mejor predicción.

El tercer modelo, Modelos de Regresión *Sparse Group* MIDAS, utiliza una combinación de regresión y técnicas de mínimos cuadrados penalizados para realizar la predicción. Este enfoque tiene en cuenta la relación entre las variables medidas a diferentes frecuencias y permite seleccionar de manera automática las variables más relevantes para la predicción.

El último modelo propuesto, Modelos de Regresión MIDAS restringidos, utiliza una versión modificada de los modelos MIDAS (*Mixed Data Sampling*) para realizar la predicción del PIB. Estos modelos consideran la información de las variables medidas a diferentes frecuencias, pero con restricciones adicionales que mejoran la precisión de la predicción.

English abstract

The objective of this Master's Thesis is to improve the accuracy in Gross Domestic Product (GDP) prediction by utilizing variables measured at a higher frequency. To achieve this, the time series data is transformed using TRAMO-SEATS and Log-differencing methodologies. Four prediction models are employed: Random Forests, Redes neuronales feedforward de una sola capa oculta, Sparse Group MIDAS Regression Models, and Restricted MIDAS Regression Models.

The first model, Random Forests, utilizes machine learning techniques based on decision trees to make predictions. This approach allows capturing non-linear relationships and considers the importance of each variable in GDP prediction.

The second model, Redes neuronales feedforward de una sola capa oculta, employs a simple neural network architecture for prediction. These networks can learn complex patterns in the data and adjust the weights of connections between neurons to achieve better prediction accuracy.

The third model, *Sparse Group* MIDAS Regression Models, combines regression and penalized least squares techniques for prediction. This approach takes into account the relationship between variables

measured at different frequencies and automatically selects the most relevant variables for prediction.

The final proposed model, Restricted MIDAS Regression Models, employs a modified version of MIDAS (Mixed Data Sampling) models for GDP prediction. These models consider information from variables measured at different frequencies but incorporate additional restrictions to enhance prediction accuracy.

Capítulo 1

Introducción

1.1. Motivación

En el área de Planificación Estratégica y PMO *Project Management Office* de ABANCA Corporación Bancaria S.A. (en adelante, ABANCA), se llevan a cabo diversas tareas, entre las cuales se encuentra el seguimiento de la evolución macroeconómica, el desarrollo de los Planes Estratégicos, la presupuestación anual y los procedimientos necesarios para cumplir con los requisitos del supervisor (*Stress Test*, ICAAP, entre otros). Específicamente, el seguimiento de la evolución macroeconómica se realiza para distintas economías, siendo de especial interés la economía española, ya que es en esta región donde reside la mayor parte del negocio de la entidad.

Para llevar a cabo esta labor, se realizan seguimientos de las principales magnitudes macroeconómicas publicadas por diferentes organismos oficiales e institutos de estadística detalladas en el Cuadro 1.1, así como indicadores coyunturales que ofrecen una imagen precisa del comportamiento de la actividad económica.

Fuente	Abreviatura
Instituto Nacional de Estadística	INE
Eurostat	Eurostat
Red Eléctrica Española	REE
Ministerio de Transportes Movilidad y Agenda Urbana	MITMA
Agencia Estatal de Administración Tributaria	AEAT
Ministerio de Inclusión, Seguridad Social y Migraciones	MISSM
Ministerio de Industria, Comercio y Turismo	MICT
Corporación de Reservas Estrategicas de Productos Petrolíferos	CORES
Ministerio de Asuntos Económicos y Transformación Digital	MINECO
Banco de España	BDE

Cuadro 1.1: Fuentes y forma en la que se denominarán en este trabajo.

En el Cuadro 1.1 puede verse el desglose de las fuentes de los datos y la abreviatura mediante la que cual serán nombradas en este trabajo.

Sin embargo, muchas de las variables macroeconómicas e indicadores coyunturales que se analizan son publicados con cierto retraso, lo que provoca demoras en los análisis realizados con base en ellos.

Además, algunos de estos indicadores se publican trimestralmente, lo cual impide incluir su análisis en el seguimiento mensual que se realiza en la entidad. El PIB y sus principales componentes son uno de estos indicadores, publicados trimestralmente con un desfase de entre 3 y 4 semanas desde el final del trimestre.

Además de por esto, se tiene un especial interés en la modelización del PIB porque es uno de los indicadores más importantes para entender la economía de un país Callen (2008). A nivel nacional, los gobiernos, los inversores, las empresas y los agentes económicos en general, utilizan el PIB para evaluar el rendimiento económico, identificar tendencias y tomar decisiones importantes. A nivel internacional, el PIB se utiliza para comparar el rendimiento económico de diferentes países y determinar su posición en la economía global. Teniendo este interés presente, este trabajo se centrará en el *nowcasting* del PIB español que se define como la predicción en el presente, en el futuro cercano y en el pasado más reciente. El principio básico del *nowcasting* es el empleo de la información que se publica con una mayor frecuencia que la variable objetivo para obtener predicciones más tempranas véase Bańbura et al. (2013).

El objetivo de este trabajo de fin de máster es tratar de proporcionar una predicción del PIB mediante el uso de variables medidas a una frecuencia más alta. Para ello se emplean metodologías como TRAMO-SEATS o Log-diferenciación para transformar los datos y mejorar las propiedades de los 4 modelos de predicción considerados: Bosques Aleatorios, Redes neuronales *feedforward* de una sola capa oculta, Modelos de Regresión *Sparse Group* MIDAS (sg-MIDAS) y Modelos de Regresión MIDAS restringidos.

La memoria se estructura en 5 capítulos que abarcan diferentes aspectos. En el Capítulo 1 se establecen el contexto en el que surge este trabajo y los datos utilizados proporcionando un panorama general de la problemática, los objetivos y los medios disponibles. El Capítulo 2 se centra en la transformación de las variables. Se describen las metodologías utilizadas, como TRAMO-SEATS y Log-diferenciación. El Capítulo 3 expone el marco teórico que sustenta la construcción de los modelos de predicción. Se revisan las teorías y enfoques existentes para comprender la relación entre las variables de alta frecuencia y el PIB. En el Capítulo 4, se lleva a cabo la implementación práctica de los modelos propuestos. Se aplican los métodos de Bosques Aleatorios, Redes neuronales *feedforward* de una sola capa oculta, Modelos de Regresión *Sparse Group* MIDAS y Modelos de Regresión MIDAS restringidos, y se presentan los resultados. Finalmente, en el Capítulo 5 se comentan las conclusiones alcanzadas a partir del trabajo realizado y las posibles líneas de trabajo futuras.

1.2. Datos

En esta sección se dan a conocer los datos empleados en este trabajo con un especial énfasis en la serie de tiempo del PIB.

1.2.1. Variables

Para realizar el *nowcasting* se utilizará el PIB, observado a frecuencia trimestral como variable dependiente y una batería de variables de carácter económico observadas a frecuencia mensual como variables independientes, las series utilizadas tendrán datos desde el 2011 hasta el último dato publicado al que se tenga acceso y su desglose puede verse en los Cuadros 1.2 y 1.3.

Producto Interior Bruto (PIB) es una medida del valor total de los bienes y servicios producidos dentro de un área geográfica. Índice Base 100. Fuente INE.

Pernoctaciones número de pernoctaciones en establecimientos turísticos. El total incluye tanto a los turistas nacionales como a los turistas extranjeros. Unidades. Fuente INE.

Índice de Comercio al por Menor (ICM) indicador económico que mide la variación en el valor total de las ventas minoristas en un país. El índice de comercio al por menor incluye diferentes

categorías de productos, como alimentos y bebidas, ropa y accesorios, artículos para el hogar, electrónica y equipo de oficina, entre otros. Índice Base 100. Fuente INE.

Índice de Cifra de Negocios Sector Servicios indicador que mide la evolución a corto plazo de la cifra de negocios, los ingresos de las empresas del sector servicios y de los sectores que lo componen. Índice Base 100. Fuente INE.

Índice de Producción Industrial (IPI) mide la evolución mensual de la actividad productiva de las ramas industriales, es decir, de las industrias extractivas, manufactureras y de producción y distribución de energía eléctrica, agua y gas. Índice Base 100. Fuente INE.

Índice de Producción de la Construcción mide la evolución mensual de la actividad productiva en el sector de la construcción. Índice Base 100. Fuente Eurostat.

Consumos de electricidad miden el consumo de electricidad en España por los sectores de actividad de forma total y de forma individual. Índice Base 100. Fuente REE.

Edificios Visados todos los usos número de edificios con permiso para su construcción por las autoridades competentes. Unidades. Fuente MITMA.

Viviendas Visadas número de viviendas con permiso para su construcción por las autoridades competentes. Unidades. Fuente MITMA.

Grandes Empresas. Ventas Interiores ventas realizadas por las grandes empresas en el mercado interior (todos los sectores de producción). Índice Base 100. Fuente AEAT.

Trabajadores en Alta Laboral Afiliados a la Seguridad Social número total de trabajadores dados de alta en la Seguridad Social por sectores productivos y el total. Miles. Fuente MISSM.

Consumo aparente de Cemento cantidad de cemento consumido. Miles de toneladas. Fuente MICT.

Consumo de Gasolinas y Gasóleos Auto consumo de combustibles. Miles de toneladas. Fuente CORES.

Consumo de Gas-Oil Agricultura y Pesca cantidad de combustible diésel utilizado por el sector de la agricultura y la pesca. Miles de toneladas. Fuente CORES.

Índice PMI Industria Manufacturera y Servicios el Índice de Gestores de Compras, es un indicador macroeconómico que pretende reflejar la situación económica de un país basándose en los datos recabados por una encuesta mensual de sus empresas más representativas que realizan los gestores de compras. Índice Base 50. Fuente MINECO.

ED. Créditos créditos concedidos a empresas residentes en España. Miles de euros. Fuente BDE.

En los Cuadros 1.2 y 1.3, se indica el nombre original de las variables, el nuevo nombre con el que aparecerán en algunas Figuras y Cuadros para facilitar su visualización, su fuente de origen, la unidad en la que están medidas y si han sido corregidas por la fuente de origen.

Estas variables han sido elegidas en base a la teoría económica, el juicio de economistas y su uso recurrente en artículos motivados en la predicción del PIB.

En el Cuadro 1.4, pueden verse algunos de los artículos donde se predice el PIB, así como, un resumen de las variables utilizadas por cada uno de ellos y el país objetivo del estudio. Las variables más utilizadas por estos autores son las referentes a la producción mientras que las menos utilizadas son las referentes al turismo como se ve al contar los sies y noes del Cuadro 1.4.

Nombre original	Nombre nuevo	Fuente	Medición	Corrección
PRODUCTO INTERIOR BRUTO PM VOLUMEN REVISIÓN ESTADÍSTICA 2019 CVEC	PIB	INE	Base 100	CVEC
TRABAJADORES EN ALTA LABORAL AFILIADOS A LA SEGURIDAD SOCIAL TOTAL	TRAB_TOTAL	MISSM	Miles	Ninguna
INDICE PMI INDUSTRIA MANUFACTURERA CVE ESPAÑA	PMI_INDUS	MINECO	Base 50	CVEC
INDICE PMI ACTIVIDAD SERVICIOS CVE ESPAÑA	PMI_SERV	MINECO	Base 50	CVEC
ÍNDICE DE COMERCIO AL POR MENOR CON ESTACIONES DE SERVICIO	ICM	INE	Base 100	CVEC
ÍNDICE DE COMERCIO AL POR MENOR SIN ESTACIONES DE SERVICIO	ICM_SIN_ESTACIONES	INE	Base 100	CVEC
TRABAJADORES EN ALTA LABORAL TOTAL SISTEMA AGRICULTURA Y PESCA CNAE 09	TRAB_AGRI	MISSM	Miles	Ninguna
TRABAJADORES EN ALTA LABORAL TOTAL SISTEMA INDUSTRIA CNAE 09	TRAB_INDU	MISSM	Miles	Ninguna
TRABAJADORES EN ALTA LABORAL TOTAL SISTEMA CONSTRUCCIÓN CNAE 09	TRAB_CONST	MISSM	Miles	Ninguna
TRABAJADORES EN ALTA LABORAL TOTAL SISTEMA SERVICIOS CNAE 09	TRAB_SERV	MISSM	Miles	Ninguna
CONSUMO APARENTE DE CEMENTO	C_CEMENTO	MICT	Miles de toneladas	Ninguna
EOH NUMERO DE PERNOCTACIONES TOTAL	PERNOC_TOTAL	INE	Unidades	Ninguna
CONSUMO DE GASOLINAS AUTO	C_GASOLINAS	CORES	Miles de toneladas	Ninguna
CONSUMO ELECTRICO TOTAL	C_ELECT_TOTAL	REE	Base 100	CVEC
CONSUMO ELECTRICIDAD INDUSTRIA	C_ELECT_INDUS	REE	Base 100	CVEC
CONSUMO ELECTRICIDAD SERVICIOS	C_ELECT_SERV	REE	Base 100	CVEC
CONSUMO ELECTRICIDAD CONSTRUCCIÓN	C_ELECT_CONST	REE	Base 100	CVEC
CONSUMO ELECTRICIDAD AGRICULTURA	C_ELECT_AGRI	REE	Base 100	CVEC

Cuadro 1.2: Especificaciones de las variables empleadas.

1.2.2. Evolución del PIB y otras variables relevantes

PIB

La serie de tiempo del PIB de España que se puede ver en la Figura 1.1 donde se muestra la serie en niveles, en variaciones trimestrales y en variaciones anuales, ha experimentado una notable evolución desde 1995 hasta la actualidad. Durante este período, se observa un patrón general de crecimiento económico, con fluctuaciones y eventos significativos que han marcado su trayectoria. En la década de

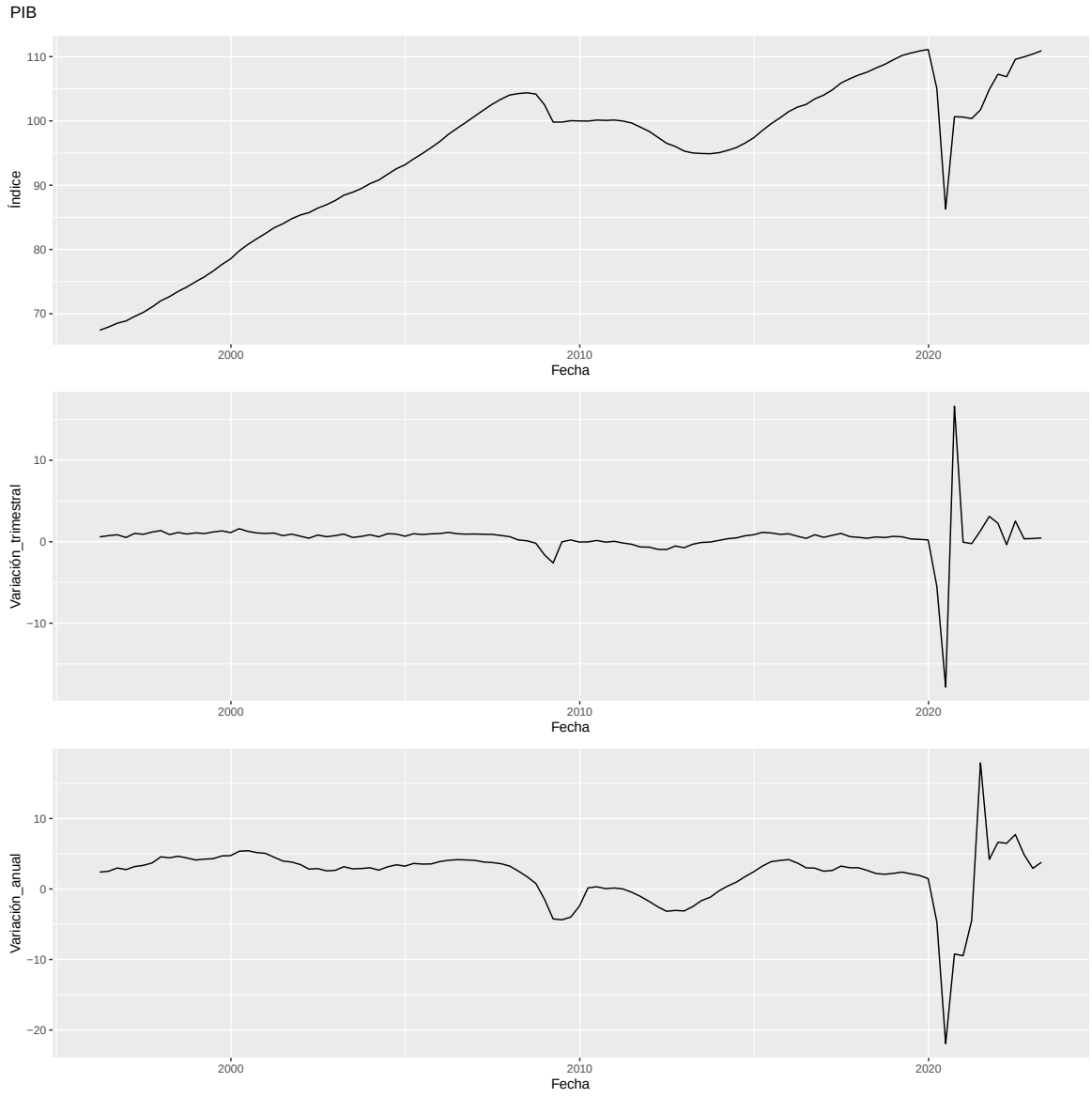

Figura 1.1: Evolución del PIB español de 1995 hasta la actualidad, serie corregida de efectos de estacionalidad y calendario. Niveles, Variaciones trimestrales y Variaciones anuales.

1990, España experimentó un aumento constante en su PIB, respaldado por reformas estructurales, la integración en la Unión Europea y un período de estabilidad económica. Este crecimiento continuó durante los primeros años del nuevo milenio, impulsado por sectores clave como la construcción, el

turismo y los servicios.

Sin embargo, la crisis financiera global de 2008 tuvo un impacto significativo en la economía española. España se vio especialmente afectada debido a la burbuja inmobiliaria y una alta dependencia del sector de la construcción. Como resultado, el PIB experimentó una contracción notable y el país entró en una recesión que duró varios años.

A partir de 2014, España comenzó a mostrar signos de recuperación económica. Las reformas implementadas, las políticas de estímulo y un aumento en las exportaciones contribuyeron a una mejora gradual del PIB. El crecimiento se mantuvo constante en los años siguientes, con un enfoque renovado en sectores como el turismo, la tecnología y las energías renovables.

Es importante destacar que la pandemia de COVID-19, tuvo un impacto sin precedentes en la economía mundial, incluida España. Las restricciones y el confinamiento afectaron negativamente a sectores como el turismo, la hostelería y el comercio minorista, lo que llevó a una contracción significativa del PIB, dicha contracción comienza a ser revertida en el periodo actual.

Trabajadores de alta afiliados a la SS. Sector de la construcción

En la Figura 1.2 puede verse la evolución de esta variable desde 1995 hasta 2023.

Durante el auge de la construcción en la década de 1990 y principios de la década de 2000, el sector experimentó un crecimiento exponencial, lo que se tradujo en un aumento en el número de trabajadores afiliados a la Seguridad Social. Esto a su vez contribuyó al crecimiento del PIB español, ya que el sector de la construcción aportaba una parte considerable de la actividad económica y generaba empleo e inversión.

Sin embargo, la crisis financiera y económica de 2008 tuvo un impacto significativo tanto en el sector de la construcción como en el PIB. La caída en la demanda de viviendas y la contracción en la actividad de construcción se reflejaron en una disminución en la serie de tiempo de los trabajadores afiliados, y el PIB español experimentó una recesión profunda.

A medida que la economía comenzó a recuperarse y el sector de la construcción mostró signos de estabilización, tanto la serie de tiempo de los trabajadores afiliados como el PIB empezaron a mostrar mejoras. Aunque la recuperación puede haber sido más lenta en el sector de la construcción en comparación con otros sectores, su contribución al crecimiento del PIB se hizo evidente a medida que se generaba más actividad y empleo en la industria.

Consumo eléctrico. Total

En la Figura 1.3 puede verse la evolución de esta variable desde 2010 hasta 2023.

El consumo eléctrico total de la economía es un indicador importante que refleja la demanda de energía en todos los sectores económicos, incluyendo la industria, el comercio, los servicios y los hogares. El consumo de electricidad suele estar relacionado con la actividad económica y la producción de bienes y servicios.

En períodos de crecimiento económico, generalmente se observa un aumento en el consumo eléctrico total, ya que las empresas aumentan su producción y los hogares incrementan su consumo de energía. Esto está impulsado por una mayor actividad económica y una mayor demanda de bienes y servicios. Durante la década de 2010, España se enfrentó a una recesión económica debido a la crisis financiera mundial de 2008. Esto tuvo un impacto en el consumo eléctrico, ya que la actividad económica se contrajo y la demanda de electricidad disminuyó. Sin embargo, a medida que la economía comenzó a recuperarse, el consumo eléctrico también mostró signos de crecimiento gradual. Como se comentaba el consumo eléctrico está muy ligado con la economía de un país, por lo tanto, con la contracción de la economía con el COVID también se redujo el consumo electrico.

Índice de Cifra de Negocios. Transporte y Almacenamiento

En la Figura 1.4 puede verse la evolución de esta variable desde la década de los 2000 hasta 2023. Durante la década de 1990, España experimentó un crecimiento económico sólido y un aumento en el

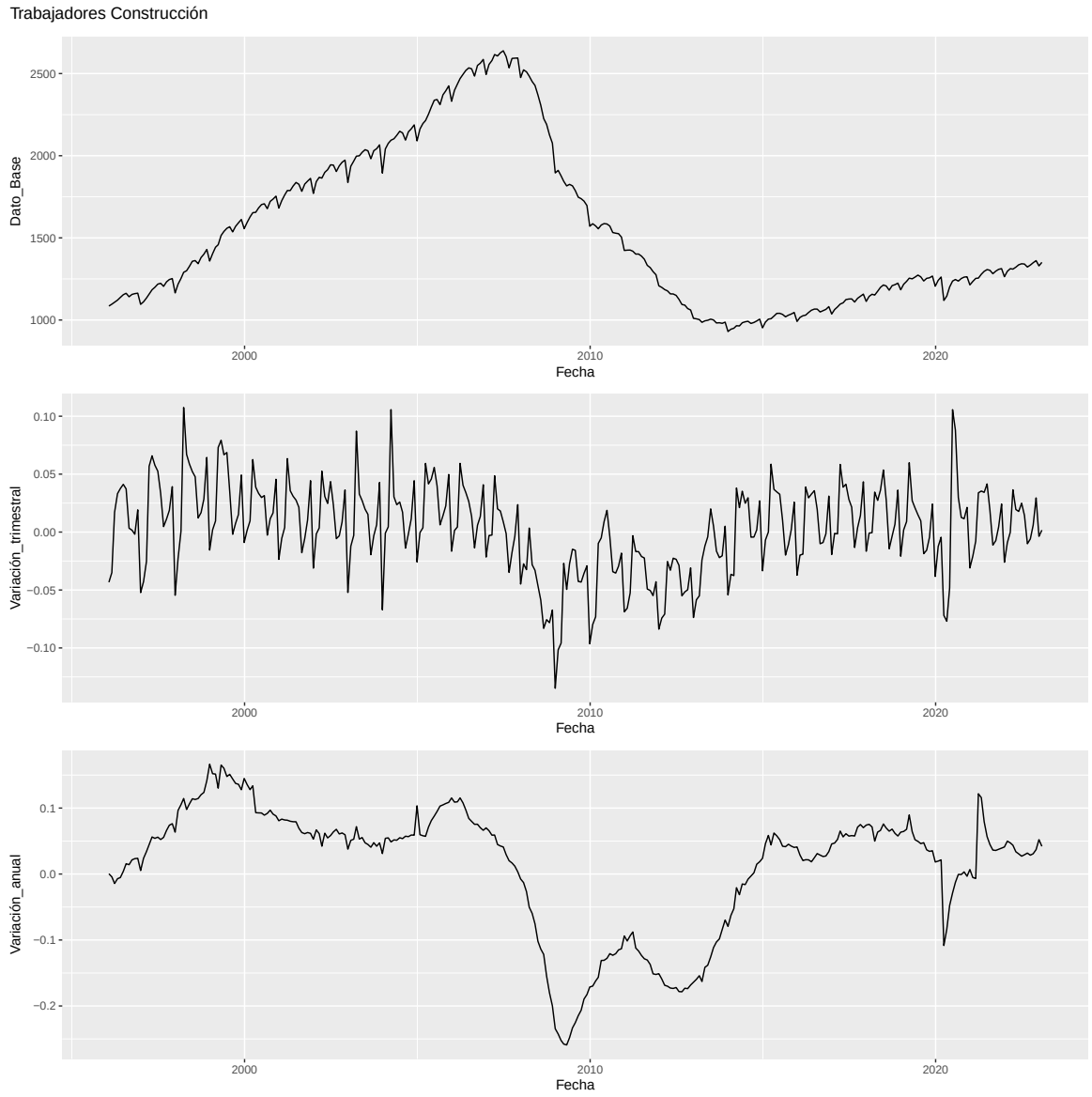

Figura 1.2: Evolución de los Trabajadores afiliados a la SS. del sector de la construcción español de 1995 hasta la actualidad, serie corregida de efectos de estacionalidad y calendario. Niveles, Variaciones trimestrales y Variaciones anuales.

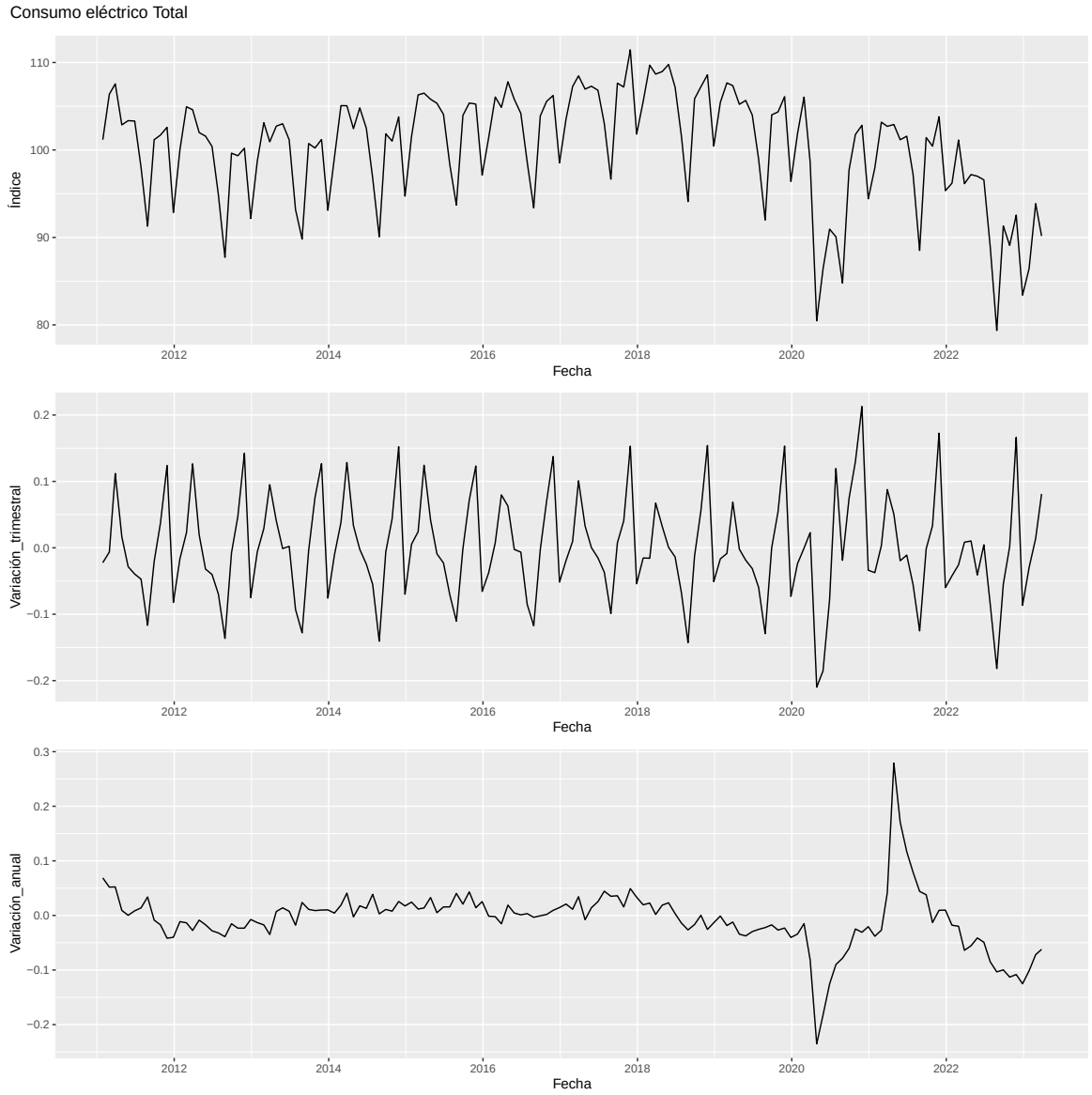

Figura 1.3: Evolución del Consumo eléctrico de España de 2010 hasta la actualidad, serie corregida de efectos de estacionalidad y calendario. Niveles, Variaciones trimestrales y Variaciones anuales.

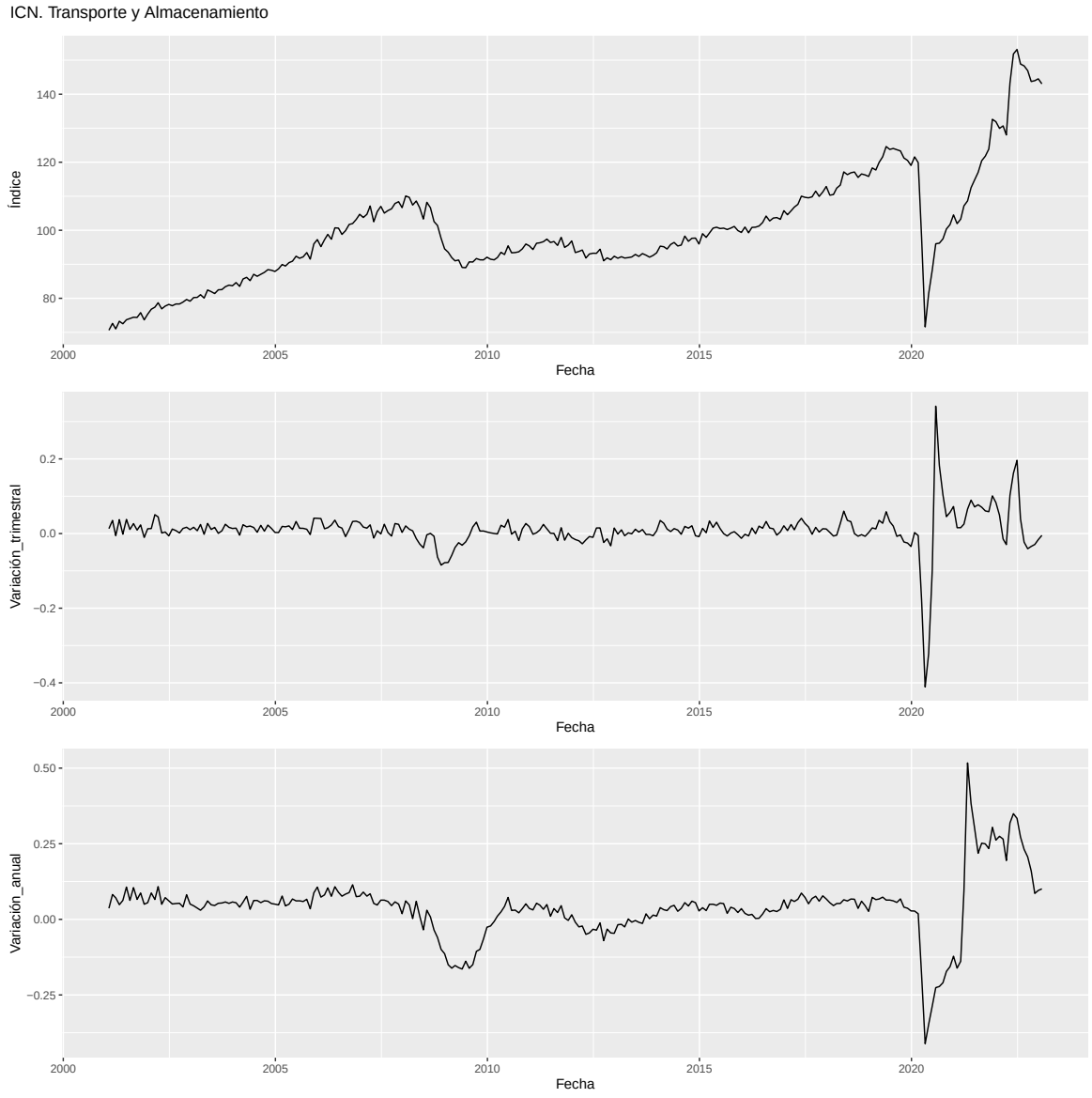

Figura 1.4: Evolución del ICN en Transporte y Almacenamiento de España de 2000 hasta la actualidad, serie corregida de efectos de estacionalidad y calendario. Niveles, Variaciones trimestrales y Variaciones anuales.

comercio internacional. Esto pudo haber impulsado el Índice de Cifra de Negocios en los sectores de Transporte y Almacenamiento, ya que se necesitaba una infraestructura logística sólida para facilitar el comercio y el transporte de mercancías.

En la década de 2000, España continuó desarrollando su infraestructura de transporte y logística, incluyendo mejoras en las carreteras, los puertos y las redes de transporte ferroviario. Esto pudo haber contribuido a un crecimiento adicional en el Índice de Cifra de Negocios en los sectores de Transporte y Almacenamiento.

Sin embargo, la crisis financiera mundial de 2008 y la posterior crisis económica tuvieron un impacto significativo en la economía española y en los sectores de Transporte y Almacenamiento. Durante estos períodos, la demanda de transporte y servicios de almacenamiento puede haber disminuido debido a la contracción económica general y la reducción del comercio.

A partir de mediados de la década de 2010, España comenzó a mostrar signos de recuperación económica gradual. A medida que la economía se estabilizaba, es posible que el Índice de Cifra de Negocios en los sectores de Transporte y Almacenamiento haya experimentado un aumento nuevamente, impulsado por la mejora de la actividad económica y la demanda de servicios logísticos.

De nuevo, esta variable, está profundamente influenciada por el crecimiento en general de la economía de un país lo que explicará la mayoría de las variaciones de esta serie.

Índice de Producción Industrial

En la Figura 1.5 puede verse la evolución de esta variable desde 1995 hasta 2023. Durante la década de 1990, España experimentó un crecimiento económico sólido, y el índice de producción industrial mostró una tendencia al alza. Esto se debió en gran medida a la expansión del sector manufacturero y a la mejora de la competitividad de las industrias españolas. Factores como la integración en la Unión Europea y las inversiones extranjeras contribuyeron al aumento de la producción industrial.

Sin embargo, a principios de la década de 2000, España enfrentó desafíos económicos, como el impacto de la crisis de las puntocom y los efectos de la desaceleración global. Estos factores llevaron a una disminución en el índice de producción industrial.

A partir de mediados de la década de 2000, España experimentó un auge económico impulsado en gran medida por el sector de la construcción y el mercado inmobiliario. Esto tuvo un impacto positivo en el índice de producción industrial, ya que aumentó la demanda de materiales de construcción y productos relacionados.

Más adelante, la crisis financiera mundial de 2008 y la posterior crisis económica tuvieron un impacto significativo en la producción industrial en España. La demanda interna disminuyó, y la contracción del sector de la construcción afectó negativamente a muchas industrias relacionadas. Esto resultó en una caída pronunciada en el índice de producción industrial durante varios años.

A partir de mediados de la década de 2010, España comenzó a mostrar signos de recuperación económica gradual. El índice de producción industrial mostró una tendencia ascendente a medida que la economía se estabilizaba y la demanda interna y externa aumentaba.

En general, estas variables demuestran la importancia de poder tener un dato actualizado del PIB, ya que, muchas series de tiempo se ven influenciadas por el crecimiento del resto de sectores y el global, por lo tanto, si se conoce la evolución del PIB se puede tener una idea fundada de como será la fluctuación de muchas otras series de tiempo.

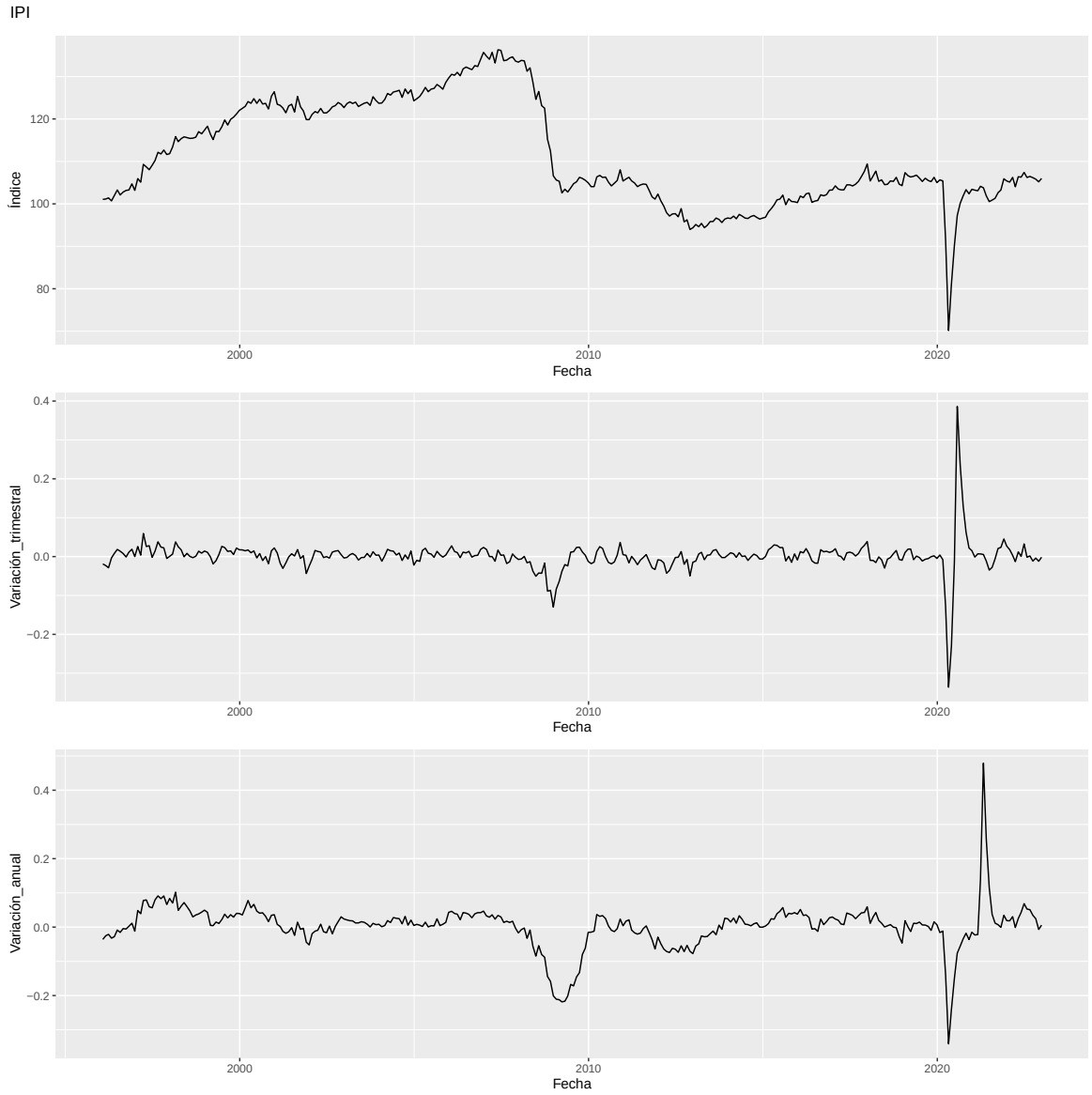

Figura 1.5: Evolución del IPI español de 1995 hasta la actualidad, serie corregida de efectos de estacionalidad y calendario. Niveles, Variaciones trimestrales y Variaciones anuales.

Nombre original	Nombre nuevo	Fuente	Unidad de medida	Corrección
SUBTOTAL GASÓLEOS AUTO	GASÓLEOS_AUTO	CORES	Miles de toneladas	Ninguna
EOH NUMERO DE PERNOCTACIONES DE NO RESIDENTES TOTAL	PERNOC_NO_RESIDENTES	INE	Unidades	Ninguna
CONSUMO DE GAS-OIL AGRICULTURA Y PESCA	C_GASOIL_AGRI	CORES	Miles de toneladas	Ninguna
ÍNDICE DE PRODUCCIÓN DE LA CONSTRUCCIÓN TOTAL ESPAÑA CCAL	I_PROD_CONST	EUROSTAT	Base 100	Calendario
ÍNDICE CIFRA NEGOCIOS SECTOR SERVICIOS GENERAL CVEC	ICN_SERV	INE	Base 100	CVEC
ÍNDICE CIFRA NEGOCIOS SECTOR SERVICIOS COMERCIO CVEC	ICN_COMERCIO	INE	Base 100	CVEC
ÍNDICE CIFRA NEGOCIOS SECTOR SERVICIOS OTROS SERVICIOS CVEC	ICN_OTROS_SERV	IN	Base 100	CVEC
ÍNDICE CIFRA NEGOCIOS SECTOR SERVICIOS TRANSPORTE Y ALMACENAMIENTO CVEC	ICN_TRANS	INE	Base 100	CVEC
ÍNDICE CIFRA NEGOCIOS SECTOR SERVICIOS HOSTELERÍA CVEC	ICN_HOST	INE	Base 100	CVEC
ÍNDICE CIFRA NEGOCIOS SECTOR SERVICIOS INFORMACIÓN Y COMUNICACIONES CVEC	ICN_INFO	INE	Base 100	CVEC
ÍNDICE DE PRODUCCIÓN INDUSTRIAL	IPI	MINECO	Base 100	CVEC
ED CRÉDITOS	ED_CRÉDITOS	BDE	Miles de euros	Ninguna
GRANDES EMPRESAS VENTAS INTERIORES DEFLACTADAS Y CVEC	VENTAS_G_EMPRESAS	AEAT	Base 100	CVEC
EDIFICIOS VISADOS TODOS LOS USOS	EDIFICIOS_VISADOS	MITMA	Unidades	Ninguna
VIVIENDAS VISADAS	VIVIENDAS_VISADAS	MITMA	Unidades	Ninguna

Cuadro 1.3: Especificaciones de las variables empleadas.

Autor	País	Empleo	Consumos	Producción	PMI	Vivienda	Precios	Ventas	Turismo
Fernández Cerezo (2023)	ESP	Sí	Sí	Sí	Sí	No	No	No	No
Babii et al. (2022)	USA	No	No	Sí	Sí	Sí	Sí	Sí	No
Hopp (2022)	USA	No	No	No	No	Sí	Sí	Sí	No
Kuzin et al. (2011)	EU	Sí	Sí	Sí	No	Sí	Sí	No	No
Zhemkov (2021)	RUS	Sí	No	Sí	Sí	No	No	Sí	No
Asimakopoulos et al. (2013)	USA	Sí	No	No	No	No	No	Sí	No
Torres et al. (2015)	EU	No	No	Sí	Sí	No	Sí	Sí	Sí

Cuadro 1.4: Variables empleadas en artículos similares.

Capítulo 2

Transformación de los datos

Las series de tiempo de carácter económico como el PIB se pueden utilizar para evaluar el rendimiento económico e identificar tendencias. Sin embargo, estos indicadores están influenciados por efectos de calendario y estacionales que pueden distorsionar la información que proporcionan estas series de tiempo. Aunque muchas de las series utilizadas en este trabajo ya han sido corregidas de estos efectos desde la fuente de los datos, algunas de ellas no.

Los métodos de ajuste estacional pueden dividirse en 3 categorias: métodos de suavizado lineal, métodos basados en modelos ARIMA y modelos de series de tiempo estructurales. A su vez, existen 3 herramientas que los implementan pero solo dos se usan de forma oficial (véase INE (2019) o Eurostat (2015)) en las agencias de estadística gubernamentales que son X12ARIMA y TRAMO-SEATS. La forma de utilizarlos por parte de estos organismos es mediante el software `JDemetra+` que permite usar cualquiera de los dos métodos de los cuales se hablará a lo largo de este capítulo con base a lo visto en Dagum and Bianconcini (2016).

En este trabajo se utiliza la función `seas()` del paquete `seasonal` (véase Sax (2022)) que emula las funciones de `JDemetra+`.

A la hora de elegir el método con el que corregir las series se ha tenido en cuenta que X12ARIMA es más adecuado para series de tiempo con patrones de estacionalidad simples y estables en el tiempo, mientras que el método TRAMO-SEATS proporciona mejores resultados con series de tiempo más inestables (véase Maravall and Planas (1998)), por lo tanto, teniendo en cuenta que las series de tiempo utilizadas en este trabajo abarcan un periodo de tiempo marcado por dos profundas crisis económicas, se ha optado por utilizar la metodología TRAMO-SEATS. En este capítulo se abordaran los métodos utilizados para la transformación de las series de tiempo tanto X12ARIMA como TRAMO-SEATS (este último se tratará con detalle debido a que es el que se ha utilizado en la parte práctica).

2.1. X12ARIMA

X12ARIMA, es uno de los ajustes estacionales más utilizados en la actualidad por las agencias de estadística. Fue desarrollado por Findley et al. (1998) y es una versión mejorada del método X11 ARIMA, ambos son métodos basados en el método II-X11 desarrollado por el Bureau of the Census Shiskin (1967).

Estos métodos se basan en el suavizado de filtros lineales o medias móviles aplicadas de forma secuencial mediante la adición (y sustracción) de una observación a la vez, además asumen que los componentes de las series de tiempo cambian a lo largo del tiempo y de forma estocástica.

En general, los métodos de la familia X-11 siguen el siguiente procedimiento:

Suponiendo que las observaciones en una serie temporal, y_t, $t = 1, ..., n$, pueden descomponerse de manera aditiva o multiplicativa (para describir la metodología se empleará la descomposición aditiva)

$$Y_t = T_t + S_t + I_t$$

Donde T_t es la componente de tendencia (o "ciclo-tendencia" porque también incluye movimientos cíclicos como los ciclos económicos), S_t es la componente estacional e I_t es la componente irregular (o aleatoria).

El objetivo es estimar cada una de las tres componentes y luego eliminar la componente estacional de la serie temporal, produciendo una serie temporal ajustada estacionalmente.

La descomposición se realiza mediante la aplicación iterativa de medias móviles simétricas cuyos coeficientes son:

- Media móvil de 13 términos (M_{2x12})

$$\frac{1}{24}\{1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1\}$$

- Media móvil de 5 términos (M_{3x3})

$$\frac{1}{9}\{1, 2, 3, 2, 1\}$$

- Media móvil de 7 términos (M_{3x5})

$$\frac{1}{15}\{1, 2, 3, 3, 3, 2, 1\}$$

- Media móvil de Henderson de 13 términos (H_{13})

$$\frac{1}{16796}\{-325, -468, 0, 1100, 2475, 3600, 4032, 3600, 2475, 1100, 0, -468, -325\}$$

Para una descomposición aditiva de una serie temporal mensual, el método realiza los siguientes pasos (para más detalles ver Dagum and Bianconcini (2016)):

1. Se obtiene una estimación inicial de la tendencia al calcular una media móvil de 13 términos ($M_{2\times12}$):
$$T_t = M_{2\times12}(y_t)$$

2. Se resta la estimación inicial de la componente tendencia de la serie original, dejando las componentes estacional e irregular:
$$(S_t + I_t) = Y_t - T_t$$

3. Se calcula una estimación inicial de la componente estacional utilizando una media móvil de 5 términos ($M_{3\times3}$) en la suma de las componentes estacional e irregular para cada mes, cuyos coeficientes se normalizan de manera que la suma de los mismos, para todo período de 12 meses, sea aproximadamente nula.
$$S_t = M_{3\times3}[(S_t + I_t)]$$
Al normalizar:
$$\tilde{S}_t = S_t - M_{2\times12}(S_t)$$

4. Se calcula una serie ajustada estacionalmente inicial (A_t) restando la componente estacional inicial de la serie original.
$$A_t = (T_t + I_t) = Y_t - \tilde{S}_t$$

5. Se calcula otra estimación de la tendencia utilizando una media móvil de Henderson de 13 términos (H_{13}) Mazzi et al. (2018) en la serie obtenida en el paso anterior.
$$T_t = H_{13}(A_t)$$

6. Se repite el paso 2 con las nuevas componentes.

7. Se estima la componente estacional con una media móvil de 7 términos ($M_{3\times5}$) en la suma de las componentes estacional e irregular para cada mes, cuyos coeficientes se normalizan de manera que la suma de los mismos, para todo período de 12 meses, sea aproximadamente nula.

$$S_t = M_{3\times5}[(S_t + I_t)]$$

Al normalizar:

$$\tilde{S}_t = S_t - M_{2\times12}(S_t)$$

8. Se repite el paso 4 con los nuevos coeficientes obteniendo así la serie corregida de variaciones estacionales.

2.2. TRAMO-SEATS

TRAMO-SEATS (véase Gómez and Maravall Herrero (1996)) es un método de ajuste estacional basado en los modelos ARIMA.

TRAMO estima a través de variables ficticias y modelos regARIMA los componentes determinísticos, días laborables, festividades móviles (Pascua) y valores atípicos, que luego se eliminan de la serie. En una segunda etapa, SEATS estima los componentes estocásticos, estacionalidad y ciclo-tendencia, a partir de un modelo ARIMA ajustado a los datos donde se han eliminado los componentes determinísticos. SEATS utiliza los filtros derivados del modelo ARIMA que describe el comportamiento de la serie temporal. Al imponer ciertas condiciones, se realiza una descomposición canónica única para obtener los modelos ARIMA para cada componente.

2.2.1. TRAMO

TRAMO (*Time Series Regression with ARIMA Noise, Missing Observations, and Outliers*) es un método de regresión que realiza la estimación, predicción e interpolación de los valores faltantes y errores ARIMA, en presencia de varios tipos de valores atípicos. Puede estimar el modelo ARIMA automáticamente o éste puede ser definido manualmente por el investigador.

Para el vector de observaciones $\mathbf{y} = (y_1, ..., y_n)'$ TRAMO ajusta el modelo de regresión

$$y_t = \mathbf{x}_t'\beta + \upsilon_t \tag{2.1}$$

donde $\beta = (\beta_1, ..., \beta_p)'$ es un vector de coeficientes, $\mathbf{x}_t = (x_{1t}, ..., x_{pt})'$ son los p regresores que definen la parte determinista del modelo. Por otra parte, υ_t representa la parte estocástica del modelo y se asume que viene determinado por un modelo ARIMA

$$\Psi(B)\upsilon_t = \pi(B)a_t, \tag{2.2}$$

donde B es el operador de retardos, $\Psi(B)$ y $\pi(B)$ son los polinomios en B, y a_t representa la innovación de ruido blanco $(0, \sigma^2)$.

El polinomio $\Psi(B)$ contiene las raíces unitarias asociadas a la diferenciación regular y estacional y el polinomio con raíces autorregresivas estacionarias. $\pi(B)$ representa el polinomio de medias móviles invertidas. En TRAMO, se asume que estos polinomios siguen la forma multiplicativa

$$\Psi(B) = \phi_p(B)\Phi_p(B^s)(1-B)^d(1-B^s)^D = (1+\phi_1 B+...+\phi_p B^p)(1+\Phi_1 B^s+...+\Phi_p B^{sP})(1-B)^d(1-B^s)^D,$$

$$\pi(B) = \theta_q(B)\Theta_Q(B^s) = (1 + \theta_1 B + ... + \theta_q B^q)(1 + \Theta_1 B^s + ... + \Theta_Q B^{sQ}),$$

donde s es el número de observaciones por año. El modelo puede contener una constante igual a la media de las series diferenciadas $(1-B)^d(1-B^s)y_t$. En la práctica se estima como uno de los parámetros de la regresión en (2.1).

Los regresores x_t se pueden especificar por el investigador o también los puede generar TRAMO de forma automática, en este último caso TRAMO tiene un procedimiento específico para generar las variables referidas a las variaciones en los días laborales de un mes, festividades de Pascua y valores atípicos.

Días laborales: La construcción de esta variable viene motivada por el hecho de que la distinta composición de los meses puede afectar a variables que cuantifiquen la producción o demanda de un bien o servicio, lo que implica que un mes podría tener una mayor producción que otro debido únicamente a que hubo más días laborables, para parametrizar este efecto se dividen los días de la semana en 2 categorías, días laborales (de lunes a viernes) y días no laborales (de sábado a domingo) y se define la variable como el número de días laborales menos 5/2 el número de días no laborales, dentro de esta variable se debe tener en cuenta el efecto que genera que un año sea bisiesto o no, para ello se incluirá dicho efecto mediante su agregación, el efecto valdrá:

$$\begin{cases} 0{,}75 \ si \ febrero \ y \ bisiesto \\ -0{,}25 \ si \ febrero \ pero \ no \ bisiesto \\ 0 \ otro \ caso \end{cases}$$

Pascua: Esta variable modela un constante cambio en el nivel de actividad diaria durante los d días anteriores a Pascua. La variable Pascua contiene ceros para todos los meses distintos de marzo y abril siendo el valor asignado a marzo la proporción de los d días que tiene el mes (p_M) menos la media de los p_M de años anteriores, el valor asignado a abril sigue la misma idea que el valor de marzo.

Valores atípicos: Esta variable refleja el efecto de algún evento inusual en una serie de tiempo, dicho efecto puede ser un pico aislado, un efecto transitorio que se diluya a lo largo de la serie o un efecto que se mantenga a lo largo del tiempo. Para saber de que tipo de valor atípico se trata y en que parte de la serie se encuentra, TRAMO, usa el procedimiento de Chen and Liu (1993) para la detección y corrección de valores atípicos. El efecto se modela de la siguiente forma

$$y_t = \omega\xi(B)I_t(t_0) + \upsilon_t$$

donde $\xi(B)$ es el cociente de los polinomios en B que modela el tipo de valor atípico e $It(t_0)$ es una variable que indica su localización.

$$I_t(t_0) = \begin{cases} 1 \ si \ t = t_0 \\ 0 \ otro \ caso \end{cases}$$

siendo ω el impacto del valor atípico en el momento t_0 y υ_t la serie libre de valores atípicos especificada en el modelo (2.1).

Por defecto, en el proceso de detección y corrección automática de TRAMO, $\xi(B)$ toma los siguientes valores en función del tipo de valor atípico detectado:

- Efecto aislado (AO): $\xi(B) = 1$;
- Efecto transitorio (TC): $\xi(B) = 1/(1-\delta B)$, por defecto $\delta = 0{,}7$;

- Efecto mantenido (LS): $\xi = 1/(1 - B)$

En la práctica la metodología TRAMO sigue los siguientes procedimientos:

1. Prueba preliminar para la especificación en log/niveles: Se realiza una prueba de regresión media de rango acortado para determinar si los datos originales deben transformarse en logaritmo o mantenerse en nivel. Si la pendiente es positiva, se elige la transformación logarítmica, si es negativa, se elige mantener el nivel. Cuando la pendiente está cerca de cero, se utiliza el Criterio de Información Bayesiano (BIC) para decidir cual de las dos especificaciones es mejor.

2. Prueba preliminar para los efectos de días laborables y festividades móviles: Una vez determinada una especificación apropiada se contrasta la presencia de efectos de calendario, esta prueba preliminar se realiza con regresiones utilizando el modelo (2.2) para el ruido, si el modelo cambiase posteriormente, se volvería a realizar la prueba.

3. Detección y corrección automática de valores atípicos: TRAMO cuenta con una función para detectar valores atípicos y eliminar sus efectos. El investigador puede introducir manualmente lo datos atípicos o pueden ser identificados automáticamente como en Tsay (1986) y Chen and Liu (1993). El procedimiento utilizado para incorporar o rechazar los valores atípicos es similar al procedimiento de regresión paso a paso para seleccionar la mejor ecuación de regresión.

4. Selección automática del modelo: TRAMO realiza una identificación automática del modelo ARIMA en dos pasos. En el primer paso se obtiene el polinomio no estacionario $(1-B)^d(1-B^s)^D$ del modelo (2.2). Esto se logra iterando en una secuencia de modelos AR y ARMA (véase Tiao and Tsay (1983) y Tsay (1984)). Se obtienen diferencias regulares y estacionales, hasta un orden máximo de $(1 - B)^2(1 - B^s)$.

   El segundo paso identifica un modelo ARMA para la serie estacionaria (modificado para incluir los efectos de valores atípicos y de regresión) siguiendo el procedimiento de Hannan-Rissanen Hannan and Rissanen (1982), con una mejora que consiste en utilizar el filtro de Kalman para calcular los primeros residuos en el cálculo del estimador de la varianza de las innovaciones del modelo (2.2).

   Finalmente, TRAMO combina las funciones para la identificación automática y la corrección de valores atípicos con las de identificación automática de modelos ARIMA.

5. Estimación del modelo regARIMA (véase Gómez and Maravall (1994)): Se estiman los parámetros de la regresión, los valores atípicos, valores faltantes y parámetros del modelo ARIMA.

6. Verificaciones de diagnóstico: Se realizan pruebas de diagnóstico en los residuos para evaluar la hipótesis de que son independientes e idénticamente distribuidos normalmente, con media cero y varianza constante. Se examinan el gráfico de los residuos y las funciones de autocorrelación. Se realizan pruebas de Ljung-Box para evaluar la falta de autocorrelación de los residuos, y se aplican pruebas de asimetría y kurtosis para evaluar la normalidad de los residuos. También se realizan test de predicciones fuera de la muestra para evaluar si las predicciones se comportan de forma coherente con el modelo.

7. Predicciones óptimas: Si se satisfacen las verificaciones anteriores, se utiliza el modelo para calcular las predicciones óptimas para la serie en función del Error Cuadrático Medio (ECM).

2.2.2. SEATS

SEATS (*Signal Extraction in ARIMA Time Series*), es un procedimiento que fue desarrollado para series económicas por Burman para el Banco de Inglaterra. Burman (1980).

Este procedimiento consiste en la descomposición de las series de tiempo en componentes no observados utilizando modelos ARIMA. En particular la metodología SEATS consiste en los siguientes pasos Dagum and Bianconcini (2016):

1. **Estimación del modelo ARIMA.** SEATS, comienza por utilizar el modelo ARIMA para ajustar los datos de la serie temporal, pero solo a la parte de los datos que no está afectada por los componentes deterministas como variaciones en días laborales, festividades móviles y valores atípicos. Sea y_t la serie linealizada resultante y se considere un modelo de descomposición aditivo (multiplicativo si se aplica a la transformación logarítmica de y_t), de modo que

$$z_t = (1-B)^d(1-B)^D y_t,$$

represente la serie diferenciada.

El modelo para la serie linealizada diferenciada z_t se puede escribir como

$$\phi_p(B)\Phi_p(B^s)(z_t - \bar{z}) = \theta_q(B)\Theta_Q(B^s)a_t,$$

donde $\bar{z}$ es la media de z_t; $t = 1, ..., n$, a_t es una serie de innovaciones, distribuidas normalmente con media cero y varianza σ_a^2, $\phi_p(B)$ y $\Phi_p(B^s)$ son polinomios autorregresivos (AR) en B y $\theta_q(B)$ y $\Theta_Q(B^s)$ son polinomios de media móvil (MA) en B también, que se expresan en forma multiplicativa como el producto de un polinomio regular en B y un polinomio estacional en B^s. El modelo completo se puede escribir en forma detallada como

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

y, en forma concisa, como

$$\Psi(B)y_t = \pi(B)a_t + c, \tag{2.3}$$

donde c es igual a $\Psi(B)\bar{y}$, siendo $\bar{y}$ la media de la serie linealizada y_t. En otras palabras, el modelo que asume SEATS, es el de una serie temporal lineal con innovaciones gaussianas. Cuando se usan SEATS y TRAMO conjuntamente, la estimación del modelo ARIMA se realiza mediante el método de máxima verosimilitud exacta descrito en Gómez and Maravall (1994). Cuando se usa SEATS de forma individual, se aplica el método de cuasi-máxima verosimilitud descrito por Burman (1980).

SEATS, comienza con la estimación del modelo ARIMA. Las raíces (inversas) de los polinomios AR y MA siempre se mantienen dentro del círculo unitario. Cuando el módulo de una raíz converge dentro de un intervalo preestablecido alrededor de 1 (por defecto (0,98, 1]), se fija automáticamente la raíz. Si es una raíz AR, el módulo se fija en 1. Si es una raíz MA, se fija en el límite inferior. Esta característica simple hace que SEATS, sea muy robusto a la sobre y subdiferenciación. Al utilizar TRAMO y SEATS de forma conjunta, SEATS controla las raíces AR y MA mencionadas anteriormente, utiliza el modelo ARIMA para filtrar la serie linealizada y obtiene así nuevos valores de los residuos.

2. **Derivación de los modelos ARIMA para cada componente.** Se descompone la serie en varios componentes siguiendo el modelo (2.3). La descomposición puede ser multiplicativa o aditiva pero se tratará solo la descomposición aditiva, ya que, la descomposición multiplicativa se puede manejar con la transformación logarítmica de los datos. Dicha descomposición se define como

$$y_t = T_t + C_t + S_t + I_t,$$

donde T_t denota el componente de tendencia, C_t el componente cíclico, S_t representa el componente estacional y I_t el componente irregular.

La descomposición se realiza en el dominio de frecuencia, esto implica analizar una serie de tiempo y descomponerla en diferentes componentes en función de su espectro de frecuencia. El espectro se divide en espectros aditivos, asociados con los diferentes componentes que se determinan, en su mayoría, a partir de las raíces AR del modelo:

- **Componente de tendencia:** Representa la evolución a largo plazo de la serie de tiempo. En el espectro, se observa un pico en la frecuencia 0, lo que indica que esta componente tiene una influencia constante en todos los períodos de la serie.
- **Componente estacional:** Captura patrones estacionales o periódicos en la serie de tiempo. En el espectro, se pueden observar picos en las frecuencias correspondientes a los ciclos estacionales. Estos picos indican que la serie muestra una variación repetitiva en esos períodos específicos.
- **Componente cíclico:** Captura fluctuaciones periódicas en un período mayor a un año. En el espectro, se observa un pico para una frecuencia entre 0 y $(2\pi/s)$. Además de estas fluctuaciones a largo plazo, el componente cíclico también captura las variaciones a corto plazo que se asocia a los componentes MA de orden bajo y a las raíces AR con módulos pequeños.
- **Componente irregular:** Recoge el comportamiento del ruido blanco, se caracteriza por tener un espectro plano, lo que significa que no hay picos destacados en ninguna frecuencia en particular.

Los componentes se determinan y se derivan completamente a partir de la estructura del modelo ARIMA (2.3) para la serie linealizada identificada directamente a partir de los datos.

Una suposición importante es la ortogonalidad entre los componentes, y cada uno se describirá mediante un modelo ARIMA. Para identificar los componentes, se requiere que (excepto el irregular) estén libres de ruido. Esto se conoce como la propiedad " canónica" e implica que no se puede extraer ruido blanco aditivo de un componente que no sea el irregular. La varianza de este último se maximiza de esta manera, mientras que, por el contrario, la tendencia, la estacionalidad y el ciclo son lo más estable posible.

La condición canónica sobre las componentes de tendencia, estacional y ciclo identifica una descomposición única, a partir de la cual se obtienen los modelos ARIMA para los componentes (incluyendo la aleatoriedad en las innovaciones). Esto se logra de la siguiente manera.

Sea $\Psi(B)$ el polinomio AR total del modelo ARIMA (2.3), que se factoriza como

$$\Psi(B) = \phi_p(B)\Phi_P(B^s)(1-B)^d(1-B^s)^D.$$

Las raíces de $\Psi(B)$ se asignan a los componentes no observadas de la siguiente manera:

- las raíces de $(1-B)^d = 0$ se asignan a el componente de tendencia;
- las raíces de $(1-B^s)^D = 0$ se factorizan de tal manera que $((1-B)(1+B+...+B^{s-1}))^D = 0$. En particular, la raíz de $(1-B) = 0$ van al componente de tendencia, mientras que las de $(1+B+...+B^{s-1}) = 0$ van al componente estacional;
- las raíces reales positivas de $\phi_p(B) = 0$ se asignan a la tendencia si están en un módulo mayor o igual que k, y se asignan al ciclo si están en un módulo menor que k;
- las raíces reales negativas de $\phi_p(B) = 0$ se asignan al componente estacional si $s \neq 1$, y al ciclo si $s = 1$;
- las raíces complejas de $\phi_p(B) = 0$ se asignan al componente estacional si $\omega \in [una frecuencia estacional \pm \epsilon]$, siendo ω la frecuencia de la raíz. De lo contrario, se asignan al ciclo; y
- las raíces reales positivas de $\Phi_p(B^s) = 0$, si son mayores o iguales que k se asignan a la tendencia. Por otro lado, cuando son menores que k se asignan al ciclo.

Los parámetros k y ϵ (cuya función es actuar como punto de corte entre la tendencia y el ciclo y la estacionalidad y el ciclo respectivamente) se establecen automáticamente en $0,5$ y $-0,4$, respectivamente, pero también pueden ser ingresados por el investigador.

La factorización de $\Psi(B)$ se puede reescribir como

$$\Psi(B) = \Psi_T(B)\Psi_S(B)\Psi_C(B),$$

donde $\Psi_T(B), \Psi_S(B)$ y $\Psi_C(B)$ son los polinomios AR con las raíces de tendencia, estacionalidad y ciclo, respectivamente. Sean P y Q los órdenes de los polinomios $\Psi(B)$ y $\pi(B)$ en (2.3).

Considerando el caso en que $P \geq Q$, una división polinómica del espectro del modelo (2.3) da una primera descomposición del tipo

$$\frac{\pi(B)}{\Psi(B)}a_t = \frac{\tilde{\pi}(B)}{\Psi(B)}a_{1t} + \upsilon_1,$$

donde el orden de $\tilde{\pi}(B)$ es el mínimo entre Q y $P-1$, y υ_1 es una constante (0 si $P > Q$).

Una expansión de las fracciones parciales del espectro de $[\tilde{\pi}(B)/\Psi(B)]a_{1t}$ da la descomposición

$$\frac{\tilde{\pi}(B)}{\Psi(B)}a1t = \frac{\tilde{\pi_T}(B)}{\Psi_T(B)}\tilde{a}_{Tt} + \frac{\tilde{\pi_S}(B)}{\Psi_S(B)}\tilde{a}_{St} + \frac{\tilde{\pi_C}(B)}{\Psi_C(B)}\tilde{a}_{Ct},$$

donde, para un $j = T, S, C$, se tiene que: orden $(\tilde{\pi}) \leq$ orden (Ψ_j). Sea $\tilde{g_j}(\omega)$ el espectro de $[\tilde{\pi_j}(B)/\Psi_j(B)]\tilde{a}_{jt}$, con $\upsilon_j = min\{\tilde{g_j}(\omega) : 0 \leq \omega \leq \pi\}$. Al imponer la condición canónica

$$g_j(\omega) = \tilde{g}_j(\omega) - \upsilon_j; j = T, S, C,$$

$$\upsilon = \upsilon_1 + \sum_j \upsilon_j,$$

se obtiene el espectro de los componentes finales, que dan los modelos para los componentes

$$\Psi_T(B)T_t = \pi_T(B)a_{T_t}$$
$$\Psi_S(B)S_t = \pi_S(B)a_{S_t}$$
$$\Psi_C(B)C_t = \pi_C(B)a_{C_t}$$
$$I_t \sim WN(0, \upsilon).$$

Todos los componentes tienen modelos equilibrados, en el sentido de que el orden del polinomio AR es igual al del MA. Por otro lado, cuando $Q > P$, la descomposición procede de la siguiente manera. Primero se realiza una descomposición, en la que

$$ARIMA(P, Q) = ARIMA(P, P-1) + MA(Q-P).$$

El primer componente se encuentra dentro del caso anterior de $P \geq Q$, y por lo tanto puede ser descompuesto de la manera anterior. En general,

$$ARIMA(P, P-1) = T_t + S_t + C_t + I_t$$

donde T_t, S_t, C_t, e I_t denotan la tendencia, la estacionalidad, el ciclo y el componente irregular. El componente $MA(Q-P)$, que representa desviaciones estacionarias a corto plazo, se agrega al componente cíclico. Entonces, la serie se descompone en un modelo de tendencia equilibrado, un modelo de estacionalidad equilibrado, un modelo de ciclo superior pesado y una irregularidad de ruido blanco. Los primeros tres componentes se suponen canónicos (es decir, libres de ruido).

3. **Estimación de los componentes.** Para una serie de tiempo $(y_1, y_2, ..., y_n)'$, se generan los estimadores MMSE (error cuadrático medio mínimo) de los componentes, calculados con un filtro de tipo Wiener-Kolmogorov aplicado a la serie finita mediante la extensión de esta última con previsiones y retrocesos (ver Burman (1980) y Haykin (2002)). Para $i = 1, ..., n$, se obtiene la estimación $\hat{y}_{t|n}$, igual a la esperanza condicional $E(y_t|y_1, ..., y_n)$, para todos los componentes. Cuando $n \longrightarrow \infty$, la estimación $\hat{y}_{t|n}$ se convierte en la estimación "final", que se denota por $\hat{y}_t$. Para $t = n$, se obtiene la estimación concurrente, $\hat{y}_{n|n}$, es decir, la estimación para la última observación de la serie. Las estimaciones finales y concurrentes son las de mayor interés. Cuando $n - k < t < n$, $\hat{y}_{t|n}$ proporciona un estimador preliminar, y para $t > n$, una predicción. Además de sus estimaciones, SEATS, produce varios años de previsiones de los componentes, así como los errores estándar correspondientes(SE) (estas previsiones deberán ser revisadas a medida que los datos futuros estén disponibles).

2.3. Log-diferenciación

Según lo visto en la asignatura de Series de Tiempo Aneiros (2016) y en Cryer and Kellet (1991), transformar las series mediante logaritmos neperianos y diferenciación regular puede mejorar el ajuste de los modelos a las series.

Como medida para corregir la posible heterocedasticidad de una serie de tiempo se puede transformar mediante logaritmos neperianos y de esta manera estabilizar su varianza.

Una vez transformada la serie se transforma mediante diferenciación regular para eliminar la tendencia y tratar de convertir la serie de tiempo en estacionaria.

En la Figura 2.1 se puede comprobar como varía la serie de tiempo del Consumo aparente de cemento a medida que se va transformando la serie con los procedimientos explicados en este capítulo.

La corrección de los efectos estacionales y de calendario achatan la serie de tiempo, los logaritmos cambian la escala y reducen la heterocedasticidad de la serie y por último la diferenciación reduce la tendencia.

2.4. Estacionariedad

La estacionariedad, se refiere a una propiedad deseable de una serie en la cual las propiedades estadísticas de la serie no cambian con el tiempo. En otras palabras, una serie de tiempo estacionaria exhibe una media y una varianza constante a lo largo del tiempo.

La estacionariedad según el artículo de Babii et al. (2022) no es una propiedad esencial para los modelos MIDAS pero si deseable.

Para contrastar la estacionariedad de la series, se realiza el test de Dickey-Fuller (Fuller (2009)) con la función `adf.test()` de la librería `tseries` Trapletti and Hornik (2022).

El objetivo del test de Dickey-Fuller es determinar si una serie de tiempo sigue un proceso estacionario o si muestra evidencia de raíces unitarias, lo que indica la presencia de tendencia o no estacionariedad.

La hipótesis nula del test de Dickey-Fuller es que la serie de tiempo tiene una raíz unitaria, lo que significa que no es estacionaria. La hipótesis alternativa es que la serie de tiempo es estacionaria.

El test de Dickey-Fuller se basa en un modelo de regresión que compara el comportamiento de la serie de tiempo con el comportamiento de su propio rezago (lag). La idea detrás del test es que si la serie de tiempo tiene una raíz unitaria, el rezago de la serie tendrá un coeficiente no nulo en el modelo de regresión.

En el Cuadro 2.1 se puede ver el p valor resultante de realizar este test a cada serie. Puede verse que excepto la serie correspondiente al número de trabajadores en el sector de la industria, todas las demás son estacionarias una vez han sido transformadas (indicar que la no estacionariedad de la serie viene dada por los datos pertenecientes al periodo COVID-19).

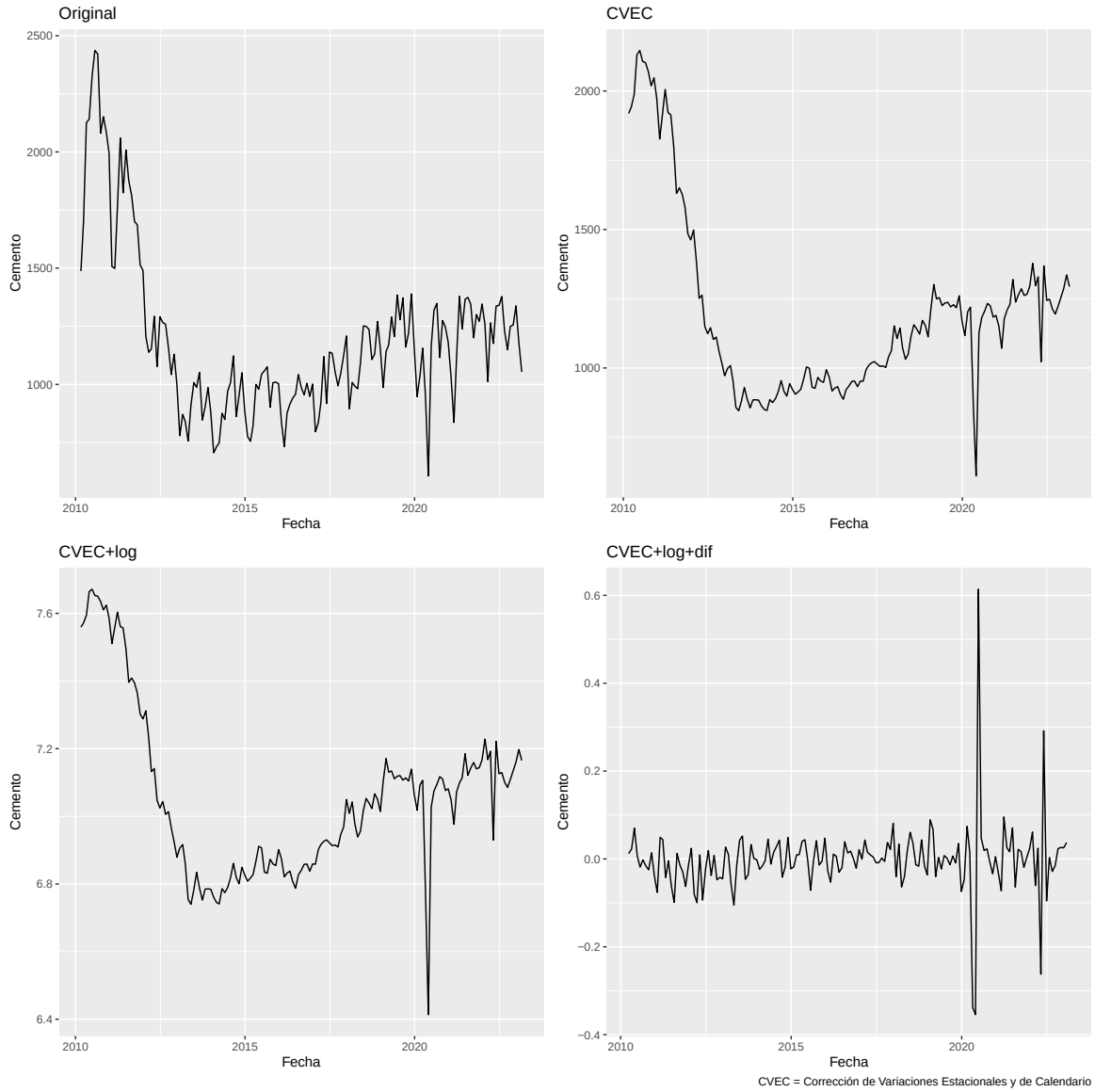

Figura 2.1: Serie del Consumo aparente de cemento a medida que se va transformando la serie.

	p-valor
PIB	0.04
TRAB_TOTAL	0.01
PMI_INDUS	0.01
PMI_SERV	0.01
ICM	0.01
ICM_SIN_ESTACIONES	0.01
TRAB_AGRI	0.01
TRAB_INDU	0.08
TRAB_CONST	0.02
TRAB_SERV	0.01
C_CEMENTO	0.01
PERNOC_TOTAL	0.01
C_GASOLINAS	0.01
C_ELECT_TOTAL	0.01
C_ELECT_INDUS	0.01
C_ELECT_SERV	0.01
C_ELECT_CONST	0.01
C_ELECT_AGRI	0.01
GASÓLEOS_AUTO	0.01
PERNOC_NO_RESIDENTES	0.01
C_GASOIL_AGRI	0.01
I_PROD_CONST	0.01
ICN_SERV	0.01
ICN_COMERCIO	0.01
ICN_OTROS_SERV	0.01
ICN_TRANS	0.01
ICN_HOST	0.01
ICN_INFO	0.01
IPI	0.01
ED_CRÉDITOS	0.01
VENTAS_G_EMPRESAS	0.01
EDIFICIOS_VISADOS	0.01
VIVIENDAS_VISADAS	0.01

Cuadro 2.1: Contraste de estacionariedad con el test de Dickey-Fuller.

Capítulo 3

Modelos

Para realizar un *nowcasting* del PIB de España se comparan cuatro modelos inspirados en los utilizados en artículos recientes con objetivos similares al de este trabajo: Bosques Aleatorios Flannery (2020), Redes Neuronales Longo et al. (2022) o Jahn (2018) y sg-MIDAS Babii et al. (2022) y MIDAS restringido Ghysels et al. (2022), los dos primeros modelo se realizarán en base a lo visto en la asignatura de Aprendizaje Estadístico Fernández et al. (2021) mientras que los dos modelos MIDAS siguiendo la literatura publicada por Eric Ghysels y otros investigadores.

3.1. Bosques aleatorios

Los bosques aleatorios son una variante de la técnica de *bagging* diseñada específicamente para trabajar con árboles de decisión, dicho de otro modo, un bosque aleatorio consiste en promediar las predicciones de múltiples árboles de decisión construidos con múltiples muestras *bootstrap* de la muestra de entrenamiento original.

En un árbol de decisión, se tiene un conjunto de variables predictoras que se utilizan para predecir una variable respuesta. El modelo divide el espacio de las variables predictoras en V regiones $R_1, R_2, R_v, ..., R_V$. Cada región representa una combinación específica de valores de las variables predictoras para cada una de las cuales se calcula una constante que representa la media de la variable respuesta para las observaciones de entrenamiento que caen dentro de esa región. Estas constantes son las que se van a utilizar para la predicción de nuevas observaciones; para ello solo hay que comprobar cuál es la región que le corresponde.

La cuestión clave es cómo se elige la partición del espacio predictor, para lo que se va a utilizar como criterio de error el RSS (suma de los residuos al cuadrado). Como se ha dicho, se modelizará la respuesta en cada región como una constante, por tanto en la región R_v nos interesa el $min_{c_v} \sum_{i \in R_v} (y_v - c_v)^2$, que se alcanza en la media de las respuestas y_v (de la muestra de entrenamiento) en la región R_v, a la que se denomina $\hat{y}_{R_v}$. Por tanto, se deben seleccionar las regiones $R_1, R_2, ..., R_V$ que minimicen

$$RSS = \sum_{v=1}^{V} \sum_{i \in R_v} (y_i - \hat{y}_{R_v})^2$$

Una forma de simplificar este problema es mediante el método CART (*Classification and Regression Trees*) se utiliza para encontrar un equilibrio entre el rendimiento del modelo y su sencillez. En lugar de considerar todas las particiones posibles, se realiza un proceso iterativo que realiza cortes binarios. En cada iteración, se selecciona una variable explicativa y un punto de corte que minimizan el error en las dos regiones resultantes.

El proceso se representa como un árbol binario, donde cada nodo representa una partición y las ramas indican las posibles respuestas a las preguntas de división. El proceso de crecimiento del árbol

continúa hasta alcanzar un criterio de parada, como una profundidad máxima o un número mínimo de observaciones en cada región.

Después de construir el árbol, se procede a podarlo para encontrar un equilibrio entre sesgo y varianza. Se utiliza un hiperparámetro para controlar el tamaño del árbol y se busca el subárbol óptimo que minimice el error en los datos de validación. La poda sucesiva de nodos internos da lugar a una sucesión finita de subárboles, de los cuales se selecciona el óptimo.

Al aplicar *bagging*, se generan muestras bootstrap que introducen aleatoriedad y hacen que cada árbol sea diferente. Sin embargo, en algunas ocasiones, estos árboles no son lo suficientemente distintos entre sí. Es decir, es común que los árboles tengan estructuras muy similares, especialmente en las primeras capas, aunque se diferencien más a medida que se desciende en ellos. Este fenómeno se conoce como correlación entre árboles y ocurre cuando un árbol es un modelo adecuado para describir la relación entre los predictores y la respuesta, especialmente cuando uno de los predictores es muy fuerte y tiene un alto grado de relevancia, lo que resulta en que este predictor esté presente en el primer nivel de corte en la mayoría de los árboles. Esta correlación entre árboles se traduce en una correlación entre sus predicciones.

El promedio de variables altamente correlacionadas produce una reducción de la varianza mucho menor que si se promedian variables no correlacionadas. La solución a este problema consiste en introducir aleatoriedad en el proceso de construcción de los árboles, de manera que se elimine la correlación entre ellos. Uno de los enfoques más destacados a la hora de llegar a esta solución es el propuesto por Breiman (2001), que consiste en que en la construcción de cada árbol que forma parte del bosque, se realizan cortes binarios y para cada corte se selecciona una variable predictora. La modificación introducida consiste en que antes de hacer cada uno de los cortes, de todas las p variables predictoras, se seleccionan al azar $k < p$ predictores que van a ser los candidatos para el corte. y representando k el valor del hiperparámetro de los bosques aleatorios que para problemas de regresión suele tomar un valor. El valor de este hiperparámetro k se suele fijar a $k = p/3$ en problemas de regresión.

3.2. Redes neuronales

Las redes neuronales artificiales han surgido como una poderosa herramienta en el campo del aprendizaje automático y la inteligencia artificial. Estas redes, inspiradas en el funcionamiento del cerebro humano, son capaces de aprender y generalizar a partir de datos, permitiendo abordar una amplia gama de problemas complejos en diferentes áreas de estudio.

Hay distintas formas de construir redes neuronales, en este trabajo se construirán redes neuronales *feedforward* de una sola capa oculta debido a su simplicidad, su eficiencia computacional, su capacidad de modelar relaciones no lineales en datos económicos y su capacidad de generalización adecuada en un contexto económico.

3.2.1. Redes neuronales *feedforward* de una sola capa oculta

Esta red se compone de una capa oculta con O variables ocultas que actúan como nodos. Cada variable oculta, denotada como h_o, es una combinación lineal de las variables predictoras X, con los pesos w_k (los parámetros w_{0o} reciben el nombre de parámetros sesgo), que se transforma por una función de activación $\phi(u)$ la cual puede variar de forma en función del objetivo de la red neuronal, para este estudio se utiliza la indicada por defecto para realizar una regresión en los artículos que usa como referencia la librería utilizada en R Kuhn (2022) y Ripley (2022).

$$h_o(X) = \phi(w_{0o} + w_{1o}x_1 + w_{2o}x_2 + ... + w_{po}x_p)$$

La forma funcional seleccionada es la logística para los nodos intermedios

$$\phi(u) = \frac{e^u}{1 + e^u}$$

y la identidad para los nodos finales.

El modelo final es una combinación lineal de las variables ocultas

$$m(x) = \gamma_0 + \gamma_1 h_1 + \gamma_2 h_2 + ... + \gamma_M h_M$$

lo que implica una gran cantidad de parámetros a estimar por lo que estos modelos se consideran hiperparametrizados y con tendencia al sobreajuste.

La estimación de los parámetros se realiza minimizando una función de pérdida, generalmente el error residual al cuadrado (RSS). La solución exacta a este problema de optimización es generalmente imposible debido a su naturaleza no convexa, por lo que se utiliza un algoritmo heurístico llamado retropropagación Werbos (1974), que se basa en el descenso de gradiente. Este algoritmo va a converger a un óptimo local pero difícilmente a un óptimo global lo que genera inestabilidad en los resultados. Este algoritmo iterativo se ejecuta en lotes de datos de la muestra de entrenamiento y se repite un número finito de veces.

Para abordar el problema del sobreajuste, se penalizan los parámetros mediante una técnica llamada reducción de pesos. Esto implica agregar un término de penalización que aumente en base a la suma de los cuadrados de los pesos, la adición de esta penalización tiene el efecto de regularizar el modelo, lo que significa que se fomenta la simplicidad del modelo al penalizar los pesos grandes. Esto ayuda a prevenir el sobreajuste, ya que los pesos grandes pueden llevar a un modelo que se ajuste demasiado a los datos de entrenamiento y no se ajuste bien a nuevos datos. Los hiperparámetros, como el término de penalización λ y el número de nodos O, se seleccionan mediante validación cruzada confiando en que el proceso de regularización forzará a que muchos pesos (parámetros) sean próximos a cero.

3.3. MIDAS

Los modelos de regresión *MIxed DAta Sampling* (MIDAS) permiten utilizar en el mismo modelo datos medidos a distintas frecuencias sin necesidad de transformarlas, la idea de estos modelos viene inspirada por los modelos con retardos distribuidos que pueden verse en Greene (2003).

La ecuación de los modelos MIDAS ha ido evolucionando con el paso de los años (Ghysels et al. (2004), Ghysels et al. (2007)), en este trabajo se utilizará el modelo visto en Babii et al. (2022) denominado *autoregressive distributed lag* MIDAS (ARDL-MIDAS).

Siendo y_t; $t = 1, ..., n$; la variable de baja frecuencia (trimestral) y $x_{mt,k}$; $k = 1, ..., p$; representando las p variables regresoras de alta frecuencia observadas m veces dentro de un intervalo de tiempo de la variable de baja frecuencia

$$\phi(B)y_t = \rho_0 + \sum_{k=1}^{p} \psi(B; \beta_k)x_{mt,k} + u_t; t = 1, ..., n,$$

donde B es el operador de retardos, $\phi(B) = I - \rho_1 B - \rho_j B^j - ... - \rho_J B^J$ es un polinomio de los retardos de baja frecuencia con $J = 0, j, ..., J$, representando el número de retardos de baja frecuencia y $\psi(B; \beta_k)x_{mt,k} = \sum_{i=0}^{I} \beta_{i,k} x_{mt-i,k}$ es un polinomio de los retardos de alta frecuencia con $I = 0, i, ..., I$, representando los retardos de alta frecuencia y u_t es el término de error aleatorio.

Este modelo permite explicar una variable de baja frecuencia utilizando como regresores variables medidas a una frecuencia más alta sin necesidad de agregar los datos.

Uno de los principales problemas de los modelos MIDAS son el número de parámetros a estimar $J + 1 + I \times p$ lo que puede ser muy grande dependiendo del valor de I que suele recomendarse que sea mayor o igual que m, debido al interés de explicar una observación de baja frecuencia con las observaciones de alta frecuencia en el mismo intervalo de tiempo como mínimo.

Para solventar la proliferación de parámetros, se parametriza el polinomio de retardos de alta frecuencia como en Ghysels et al. (2006)

$$\psi(B; \beta_k)x_{mt,k} = \sum_{i=0}^{I} \omega(i; \beta_k)x_{mt-i,k}, \tag{3.1}$$

donde β_k es un vector de coeficientes B-dimensional con $B \leq I$ y $\omega : [0,1] \times \mathbf{R}^B \longrightarrow \mathbf{R}$ es una función de ponderaciones.

Después, se aproxima la función de ponderaciones de la siguiente manera

$$\omega(u;\beta_k) \approx \sum_{l=1}^{L} \beta_{k,l} w_l(u), u \in [0,1],$$

donde $w_l : l = 1, ..., L$ representa las distintas formas en las que se puede reparametrizar $\omega(u;\beta_k)$, (en este caso los grados seleccionados del polinomio) se hace especial incapié en Babii et al. (2022), en el uso de polinomios ortogonales para reducir la multicolinearidad por ejemplo el polinomio Legendre Farouki et al. (2003), aunque una de las parametrizaciones más utilizadas es la del *exponential Almon lag polynomial* Almon (1965).

Tomando la ecuación (3.1) como base se utilizan dos modelos MIDAS distintos, uno es mediante los *Sparse Group* MIDAS vistas en Babii et al. (2022) y otra es mediante un modelo MIDAS restringido como el utilizado en Ghysels et al. (2022).

3.3.1. sg-MIDAS

El estimador LASSO, Tibshirani (1996), permite aumentar la precisión de las predicciones seleccionando un modelo basado en el principio de parsimonia, es decir, reduciendo su complejidad, lo que a su vez reduce los problemas del modelo MIDAS de sobreparametrización Marsilli (2014).

El estimador LASSO no considera la relación temporal entre las covariables en diferentes retrasos de alta frecuencia, pero el *Sparse Group* LASSO o LASSO de grupos dispersos (sg-LASSO) puede incorporar dicha estructura en el procedimiento de estimación.

Para describir el proceso de estimación, siendo $\mathbf{y} = (y_1, ..., y_n)'$, un vector de la variable dependiente y $\mathbf{X} = (\iota, y_1, ..., y_J, Z_1 W, ..., Z_p W)$, una matriz de diseño, donde $\iota = (1, 1, ...,)'$ es un vector unitario, $\mathbf{y}_j = (y_{1-j}, ..., y_{n-j})', Z_k = (x_{k,mt-i})_{t\in[n],i\in[I]}$ es una matriz de covariables $k \in [p]$, y $W = (w_l(i)_{i\in[I],l\in[L]}$ es una matriz de pesos $I \times L$. Además, $\beta = (\beta_0', \beta_1', ..., \beta_p')$ donde $\beta_0 = (\rho_0, \rho_1, ..., \rho_J)'$ es un vector de parámetros pertenecientes a un grupo compuesto por la ordenada en el origen y los coeficientes autoregresivos y $\beta_k \in \mathbf{R}^B$ denota los parámetros del polinomio de alta frecuencia que pertenecen a la covariable $k \geq 1$.

El estimador sg-LASSO resuelve el problema de mínimos cuadrados penalizados

$$min_{\beta\in\mathbf{R}^p} ||\mathbf{y} - \mathbf{X}\beta||_n^2 + 2\lambda\Omega(\beta) \tag{3.2}$$

con una función de penalizaciones que interpola entre la penalización LASSO en la norma ℓ_1 y la penalización de LASSO por agrupaciones

$$\Omega(\beta) = \alpha|\beta|_1 + (1-\alpha)||\beta||_{2,1},$$

donde $||\beta||_{2,1} = \sum_{g\in G} |\beta_g|_2$ es la norma del *group* LASSO (véase Yuan and Lin (2006)) y G (los grupos entre los que se divide p) es la estructura de grupos especificada por el investigador. Para un vector $\beta \in \mathbf{R}^p$, la estructura de grupos dispersos se describe como el par (S_0, G_0), donde $S_0 = \{k \in p : \beta_k \neq 0\}$ y $G_0 = \{g \in G : \beta_g \neq 0\}$ representan respectivamente, el conjunto de índices k distintos de cero en β_p y el conjunto de grupos g con al menos un coeficiente distinto de cero en β_p.

El valor de la penalización de la ecuación (3.2) esta controlada por el parametro $\lambda > 0$, mientras que $\alpha \in [0,1]$ es un parámetro indica la dispersión de los grupos. Cuando $\alpha = 1$ se uliza la penalización LASSO, cuando $\alpha = 0$ se utiliza una penalización similar a la *elastic net* y para $\alpha = 0{,}5$ se utiliza la penalización *sparse group* LASSO.

3.3.2. MIDAS restringido

Este modelo utiliza el método de mínimos cuadrados no lineales (NLS) y series de tiempo cortas para estimar el modelo porque, como se indica en Ghysels et al. (2022), tratar de estimar demasiados

parametros (que pueden estar altamente correlacionados) puede tener como consecuencia que el modelo resultante sea inestable y tenga unos malos resultados en sus predicciones. Además, para tratar de obtener mejores resultados se utiliza la técnica de combinación de predicciones (vease Timmermann (2006)) que consiste en realizar un modelo para cada uno de los regresores de forma individual combinando esas múltiples predicciones en una única predicción lo que producirá predicciones más robustas a modelos mal especificados.

Esta combinación se hace en función de 4 esquemas de ponderaciones:

- **EW (*Equal Weighting*):** En este esquema de ponderación, todas las predicciones reciben el mismo peso. Es la forma más básica de ponderación y se utiliza cuando no hay información adicional disponible para asignar pesos diferentes.
- **BICW (*Bayesian Information Criterion Weighting*):** El esquema de ponderación BICW utiliza el criterio de información bayesiano (BIC) para asignar pesos a las predicciones. El BIC es un criterio utilizado para seleccionar el mejor modelo entre varios modelos candidatos. Las predicciones del modelo que se ajustan mejor a la realidad reciben mayores pesos, mientras que las predicciones que no se ajustan tan bien reciben pesos más bajos.
- **MSFE (*Mean Squared Forecast Error Weighting*):** En este esquema de ponderación, se asignan los pesos a las predicciones en función de su error cuadrado medio. Las predicciones con errores más bajos reciben mayores pesos, lo que significa que tienen una mayor influencia.
- **DMSFE (*Dynamic Mean Squared Forecast Error Weighting*):** El esquema de ponderación DMSFE es similar al MSFE, la diferencia es que los pesos se recalculan a medida que se van realizando predicciones.

Se calculan las medidas de precisión MSE, MAPE y MASE para cada una de las combinaciones, y se selecciona como predicciones finales las de la combinación que minimicen esas medidas.

3.3.3. Medidas de Precisión

Para terminar el capítulo se definen las medidas de precisión o error (ME, RMSE, MAE, MAPE y R^2) que se tendrán en cuenta para evaluar los modelos construidos para la predicción del PIB..

- **ME (*Mean Error*):** El error medio, es una métrica utilizada en la evaluación de modelos de regresión para medir el sesgo promedio de las predicciones. El ME se calcula como la media aritmética de los errores, que es la diferencia entre los valores predichos y los valores observados.

$$\text{ME} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)$$

- **RMSE (*Root Mean Squared Error*):** Es una medida del error promedio entre las predicciones y los valores observados en un modelo de regresión. Se calcula como la raíz cuadrada de la media de los errores al cuadrado. El RMSE penaliza los errores grandes de manera más significativa que los errores pequeños, lo que lo hace especialmente útil cuando se desea minimizar la magnitud de los errores. Un valor de RMSE más bajo indica un mejor ajuste del modelo a los datos.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- **MAE (*Mean Absolute Error*):** Es otra medida del error promedio entre las predicciones y los valores observados. Se calcula como la media de los valores absolutos de los errores. El MAE

es menos sensible a los errores grandes que el RMSE, ya que no involucra elevar al cuadrado los errores. Al igual que el RMSE, un valor de MAE más bajo indica un mejor ajuste del modelo.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

- **MPE (*Mean Percentage Error*):** Es una medida del error porcentual promedio entre las predicciones y los valores observados. Se calcula como el promedio de los errores porcentuales, es decir, la diferencia porcentual entre los valores predichos y los valores observados. El MPE puede ser útil para evaluar el rendimiento del modelo en términos de precisión porcentual. Sin embargo, es importante tener en cuenta que el MPE puede tener problemas cuando los valores observados son cercanos a cero.

$$\text{MPE} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{y_i}\right) \times 100$$

- **MAPE (*Mean Absolute Percentage Error*):** Es similar al MPE, pero utiliza los valores absolutos de los errores porcentuales en lugar de los errores porcentuales sin signo. El MAPE se calcula como el promedio de los valores absolutos de los errores porcentuales. Al igual que el MPE, el MAPE es útil para evaluar la precisión porcentual del modelo. También puede presentar problemas cuando los valores observados son cercanos a cero.

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100$$

- $\tilde{R}^2$ **(*pseudo R-squared*):** También conocido como coeficiente de determinación, es una medida de cuánta variabilidad en los valores observados es explicada por el modelo. $\tilde{R}^2$ varía entre 0 y 1, donde 0 indica que el modelo no explica la variabilidad de los datos y 1 indica una explicación perfecta. $\tilde{R}^2$ se calcula como 1 menos la proporción de la suma de cuadrados residual dividida por la suma de cuadrados total. Un valor de $\tilde{R}^2$ más alto indica un mejor ajuste del modelo a los datos.

$$\tilde{\text{R}}^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Capítulo 4

Implementación en R

Una vez importados los datos, las series de tiempo se corrigen de estacionalidad y calendario con la metodología TRAMO-SEATS utilizando la librería de R `R-seasonal`. Estas correcciones provocan que durante el periodo afectado por el COVID-19 algunas de las series de tiempo relacionadas con el turismo (como el número de pernoctaciones), tuviesen valores negativos, estos valores negativos se sustituyen por ceros. A continuación, se transforma la serie por Log-diferencias.

Se realiza el test de Dickey-Fuller a cada una de las series de tiempo para contrastar su estacionariedad. Como no todas las series mensuales tienen la misma fecha de publicación, es decir, algunas series tienen su dato más actual en Diciembre del 2022 y otras en Enero de 2023 o Febrero de 2023, se desplazan 3 meses, 2 meses y 1 mes respectivamente hacia delante para realizar una predicción con los datos lo más actualizados posible de cada serie. Se construyen los modelos y se validan sus predicciones en 2 periodos distintos para ver como se comportan los modelos en diferentes escenarios:

- Un periodo preCOVID, entre el cuarto trimestre de 2017 y el cuarto trimestre de 2019, entorno poco volátil y sin grandes variaciones.

- Un periodo postCOVID, entre el cuarto trimestre de 2019 y el primer trimestre de 2023, entorno volátil marcado el conflicto bélico en Ucrania y el aumento notable de las tasas de inflación.

4.1. Preparación de datos (alineado de frecuencias)

Para poder utilizar los modelos con variables medidas en distintas frecuencias, lo primero es transformar las variables de alta frecuencia mediante la transformación de alineado de frecuencias (veáse Ghysels et al. (2016)), esta transformación consiste en reescribir los datos de las series de tiempo en forma matricial. Se utilizará el esquema de alineado de frecuencias a la hora de reescribir los datos empleados para todos los modelos construidos en este trabajo.

Para este estudio, se trabaja con datos trimestrales y mensuales, dado que cada trimestre consta de tres meses, en este caso la frecuencia m (número de observaciones de alta frecuencia en un periodo de baja frecuencia) es 3. Recordando la notación de la ecuación (3.3) y poniendo como un ejemplo ajeno a lo que se hace en este trabajo que se quiera estimar el valor de un trimestre con los datos mensuales de ese trimestre y el último dato del trimestre anterior, se querría modelar cada valor de y_t con los los valores mensuales de $x_{3t-0}, x_{3t-1}, x_{3t-2}, x_{3t-3}$; por ejemplo, para $t = 2$ se modelaría y_2 con los datos de x_6, x_5, x_4, x_3.

Para un caso general, una variable de alta frecuencia en formato matricial se representa como en la matriz (4.1)

$$\mathbf{X} := \begin{bmatrix} x_{3(t+0)-0} & x_{3(t+0)-1} & x_{3(t+0)-i} & \cdots & x_{3(t+0)-I} \\ x_{3(t+1)-0} & x_{3(t+1)-1} & x_{3(t+1)-i} & \cdots & x_{3(t+1)-I} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{3n-0} & x_{3n-1} & x_{3n-i} & \cdots & x_{3n-I} \end{bmatrix} \tag{4.1}$$

donde $i = (0, 1, ..., I)$ representa el número de retardos de la variable de alta frecuencia.

Una vez definida la matriz 4.1, ésta se parametriza utilizando polinomios de Almon y Legendre, para así reducir el número de parámetros a estimar. Sean $\mathbf{X}$ una matriz $n \times I$ y $\mathbf{W}$ una matriz $I \times L = 1, l, ..., L$ y siendo L el grado del polinomio (o dimensión de la base del polinomio) se construye una matriz de pesos $\mathbf{W}$ para obtener las nuevas variables $\mathbf{WX}$

$$\mathbf{WX} := \begin{bmatrix} x_{3(t+0)-0} & x_{3(t+0)-1} & x_{3(t+0)-i} & \cdots & x_{3(t+0)-I} \\ x_{3(t+1)-0} & x_{3(t+1)-1} & x_{3(t+1)-i} & \cdots & x_{3(t+1)-I} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{3n-0} & x_{3n-1} & x_{3n-i} & \cdots & x_{3n-I} \end{bmatrix} \begin{bmatrix} w_{1,1} & w_{1,l} & \cdots & w_{1,L} \\ w_{i,1} & w_{i,l} & \cdots & w_{i,L} \\ \vdots & \vdots & \ddots & \vdots \\ w_{I,1} & w_{I,l} & \cdots & w_{I,L} \end{bmatrix} \tag{4.2}$$

tras la reparametrización, el nuevo número de coeficientes a estimar pasa a ser de L para cada una de las p variables predictoras.

Cabe destacar que es necesario que el número de observaciones de la variable de alta frecuencia sea exactamente 3 veces el número de observaciones de la variable de baja frecuencia para que funcione correctamente el alineado de frecuencias.

Aunque el método de alineado es el mismo tanto en sg-MIDAS como en MIDAS restringido, se emplean distintas librerías de R para el alineado y posterior desarrollo de los modelos: `midasml` en Striaukas et al. (2022) y `midas.r` en Ghysels et al. (2016) respectivamente.

4.2. División de la muestra

Con el objetivo de evaluar los modelos construidos se dividen las observaciones en dos grupos, muestra de entrenamiento y muestra de test, lo habitual es dividir la muestra en dos segmentos fijos pero en este trabajo además de dividirlas de ese modo también se utilizan los esquemas de *Recursive Forecasting* y *Rolling Forecasting* para los que Armesto et al. (2010) comprueba que obtienen un menor RMSE que con el esquema tradicional.

4.2.1. Segmentos fijos

Esta división de la muestra viene motivada por ser un procedimiento tradicional en el campo del machine learning.

1. Muestra de entrenamiento $\approx 80\,\%$ de los datos: comprendida entre Julio de 2011 y Septiembre de 2019.

2. Muestra de test $\approx 20\,\%$ de los datos: comprendida entre Octubre de 2019 y Marzo de 2023.

4.2.2. *Recursive Forecasting*

Este método consiste en que a medida que se van realizando predicciones va aumentando la muestra de entrenamiento de modo que todas las predicciones que se realizan sean en el horizonte uno.

En esta división, la muestra de entrenamiento comienza conteniendo las mismas observaciones que en los segmentos fijos, pero a medida que se va aumentando el horizonte de predicción también lo hará la muestra de entrenamiento.

4.2.3. *Rolling Forecasting*

Este método es similar al método anterior, la única diferencia es que el tamaño de la muestra de entrenamiento es fijo, es decir, a medida que se aumenta la muestra de entrenamiento con datos más recientes se suprimen de la muestra de entrenamiento el mismo número de datos antiguos.

4.3. Bosque Aleatorio

En esta sección, se utiliza la función `randomForest` de la librería homónima Liaw et al. (2002) en R para construir un modelo de Bosques Aleatorios. Para la configuración de los hiperparámetros, se siguen las recomendaciones por defecto para regresión según la documentación de la función `randomForest()`.

En particular, se genera un conjunto de 500 árboles, cada vez que se realiza una división en un árbol, se seleccionan aleatoriamente el valor entero de un tercio de las variables predictores disponibles, 10 en este caso (se seleccionan 10 aunque según Liaw et al. (2002) el valor de este hiperparámetro no provoca cambios drásticos en los resultados), además, se establece una restricción en los nodos terminales, donde se exige que no puedan contener menos de 5 observaciones. Una vez construidos los 500 árboles se promedian sus predicciones para dar lugar al bosque aleatorio.

Una vez se ha construido el modelo, se puede obtener una estimación de su tasa su error cuadrático medio en base al cálculo del error de *"out of bag"* ("OOB") que son las observaciones que no se incluyen en la muestra bootstrap para la construcción de cada árbol individual.

El error "OOB" se calcula promediando las predicciones de las observaciones "OOB" de todos los árboles en el Random Forest y comparándolas con los valores reales correspondientes. En la Figura 4.1 se puede examinar la convergencia del error en las muestras "OOB", de donde se deduce que 500 árboles parecen ser suficientes (de lo contrario, la predicción "OOB" podría tener un sesgo positivo, ver Bylander (2002)).

Además, al calcular el MSE de las predicciones"OOB" se pueden analizar las variables predictoras más importantes (ver Figura 4.2), para realizar esta evaluación, se repite el cálculo del MSE después de permutar cada variable predictora, es decir, aleatoriamente se altera el orden de los valores de una variable, la diferencia entre los errores antes y después de la permutación se promedia en todos los árboles del modelo y se normaliza dividiéndola por la desviación estándar de esas diferencias. Esta medida de importancia relativa permite identificar las variables que tienen un mayor impacto en la precisión de las predicciones. Es importante destacar que si la desviación estándar de las diferencias es igual a cero para una variable, no se realiza la división. Sin embargo, en ese caso, el promedio de las diferencias es casi siempre igual a cero, lo que indica que esa variable no tiene un efecto significativo en la capacidad predictiva del modelo.

Las variables referidas al mercado laboral y a los Índices de Cifra de Negocios parecen ser las que mejor explican las variaciones del PIB, información que se puede ser útil a la hora de seleccionar variables para la construcción de modelos supervisados como el MIDAS restringido que se verá más adelante.

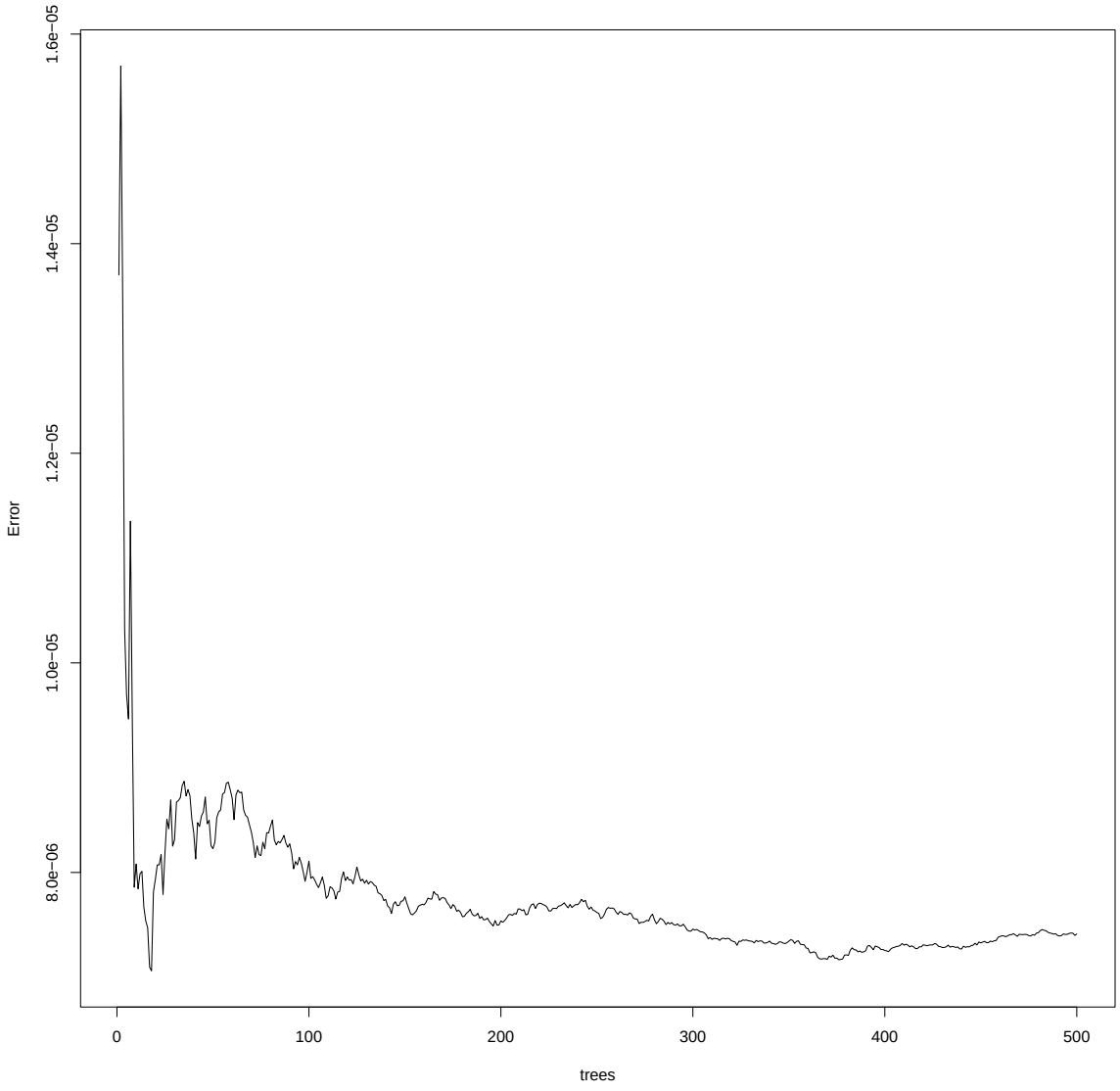

Figura 4.1: Evolución del error a medida que aumenta el número de árboles.

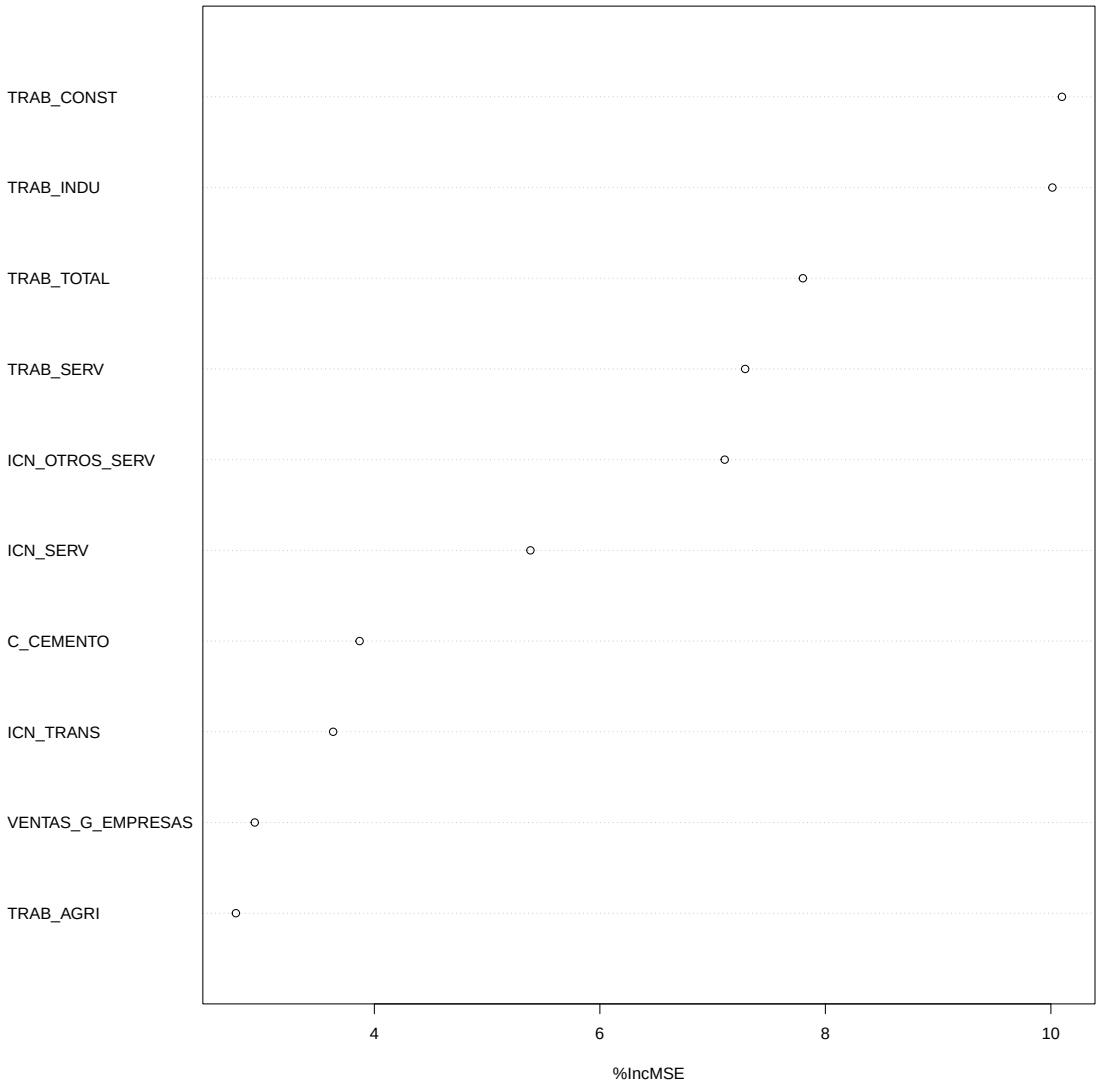

Figura 4.2: Importancia de las variables en función del incremento del MSE del modelo cuando se elimina o añade una variable.

4.4. Redes neuronales *feedforward* de una sola capa oculta

Para desarrollar y entrenar este modelo de red neuronal, se utiliza la librería `caret` Kuhn (2022) con una extensión de la librería `nnet` Ripley (2022).

En términos de preprocesamiento de datos, se aplica la técnica de reescalado utilizando la función `preProc` con el argumento `range`. Esto permite transformar las variables predictoras para que estén en el rango [0,1]. El reescalado es una práctica común en redes neuronales, ya que puede ayudar a evitar que las variables con diferentes escalas dominen el proceso de aprendizaje. Al limitar el rango de las variables a un intervalo común, se facilita la convergencia del modelo y se mejora su capacidad para capturar patrones significativos en los datos.

En cuanto a los hiperparámetros, se utiliza la función el argumento `tuneGrid` para definir una rejilla que considere todas las combinaciones entre diferentes tamaños de capa oculta (2,4,6,8,10) y valores para el parámetro de regularización de los pesos (0, 0.001, 0.01).

Se utiliza validación cruzada con 10 grupos para evaluar el rendimiento del modelo en diferentes subconjuntos de datos y seleccionar así la combinación de hiperparámetros óptima.

El modelo resultante tiene 32 variables de entrada, 2 nodos en la capa oculta y un nodo en la capa de salida, en total 69 parámetros, que sería con el que se obtiene un menor RMSE como se puede observar en la Figura 4.3.

Este modelo también permite observar cuales son las variables que tienen una mayor importancia en el modelo en base a los pesos asignados a cada variable en la red neuronal. En el Cuadro 4.1 pueden verse algunas de las variables más importantes. Al comparar las variables con mayor relevancia de este modelo con las del modelo anterior, siendo las más importantes las referidas al mercado laboral, al Índice de Cifra de Negocio de actividades relacionadas con el sector servicios y a la producción.

4.5. sg-MIDAS

Una vez preparados los datos, la matriz (4.1) se parametriza utilizando una función de pesos, que agrega los 12 retardos que se seleccionan para modelar la relación entre los regresores y el regresando. Indicar que si bien todos los predictores tienen 12 retardos, estos no corresponden al mismo intervalo de tiempo para todos ellos. Esto se debe a que las variables cuentan con diferentes fechas de publicación. Por lo tanto, los regresores se dividen en 3 conjuntos de regresores en función de la disponibilidad del dato más reciente (febrero, enero o diciembre).

En función de cada conjunto de regresores, se desplazan los datos mensuales uno, dos o tres periodos hacia adelante. Por ejemplo, al predecir el PIB para el primer trimestre de 2023, se considerarían observaciones desde febrero de 2023 hasta marzo de 2022 para algunos regresores, desde enero de 2023 hasta febrero de 2022 para otros regresores, y desde diciembre de 2022 hasta enero de 2022 para otros regresores.

Finalmente, se asigna un grupo a cada regresor y se procede con el ajuste del modelo de regresión sg-LASSO, utilizando unos parámetros de regularización λ y α, seleccionados mediante validación cruzada con 5 grupos, aplicando el criterio de minimización del error de predicción.

4.6. MIDAS restringido

4.6.1. Selección de variables

A diferencia de en los otros tres modelos donde se utilizaban todas las variables disponibles, para el MIDAS restringido la recomendación (como se comentaba en el capítulo anterior) es no utilizar un número de variables (que podrían estar altamente correlacionadas) demasiado grande, es por ello, que con el fin de conseguir un ajuste lo mejor posible se siguen los siguientes criterios de selección:

- La correlación de los regresores con el regresando.

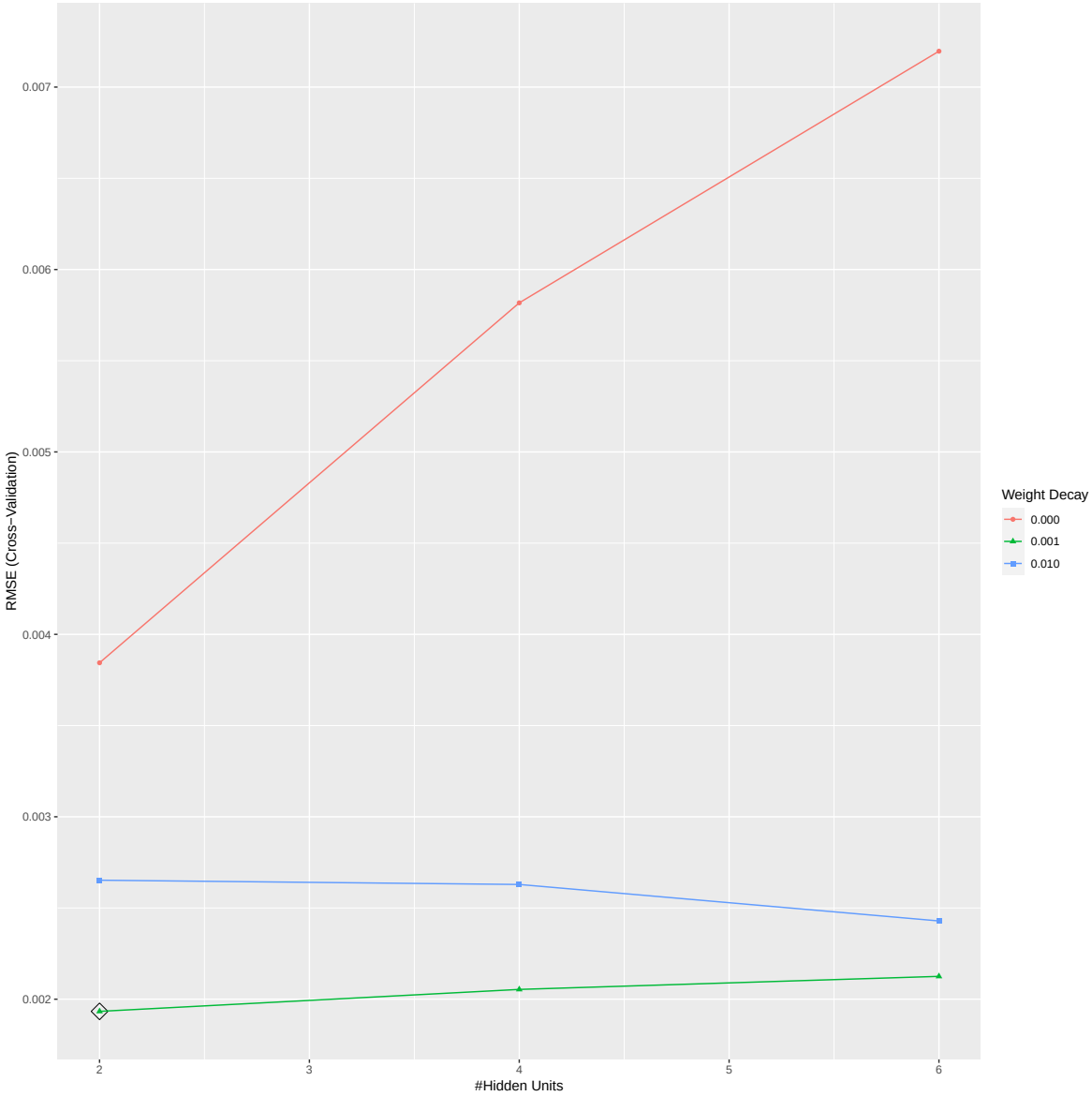

Figura 4.3: Evolución del RMSE a medida que aumenta el número de nodos de la capa oculta.

	Importancia
TRAB_CONST	100.00
ICN_TRANS.1	89.10
TRAB_INDU	81.65
VIVIENDAS_VISADAS.1	81.07
C_GASOLINAS	77.53
TRAB_SERV	68.61
TRAB_TOTAL	67.98
EDIFICIOS_VISADOS	64.00
GASÓLEOS_AUTO	63.28
IPI	60.47
C_ELECT_SERV.1	54.49
ICM	54.10
C_ELECT_AGRI.1	51.40
C_ELECT_CONST	51.24
C_ELECT_INDUS	50.96
C_GASOIL_AGRI.1	49.56
EDIFICIOS_VISADOS.1	48.83
VENTAS_G_EMPRESAS	48.15
ICN_OTROS_SERV.1	45.53
PERNOC_NO_RESIDENTES.1	44.22

Cuadro 4.1: Importancia de las variables en el modelo de Redes Neuronales.

- El juicio de profesionales en el campo de la economía.
- Las variables consideradas por los modelos anteriores como las más importantes.

Siendo las variables seleccionadas:

- Trabajadores de alta afiliados a la SS. Sector de la construcción.
- Consumo eléctrico. Total.
- Índice de cifra de negocios. Transporte y almacenamiento.
- Índice de producción industrial.

4.6.2. Correlación

Se estudia la correlación entre el PIB y las variables mensuales como un criterio a tener en cuenta en la selección de variables debido a su simplicidad y a los buenos resultados que suele producir este criterio en la práctica como se indica en Guyon and Elisseeff (2003).

Para ello, las correlaciones se hacen en términos de las Log-diferencias de los datos. Las variables mensuales trimestralizan tomando la media aritmética como en Fernández Cerezo (2023) para poder realizar este estudio.

Se estudia la correlación para la totalidad de los datos y a mayores se estudia la correlación entre los datos excluyendo el periodo influído por los efectos del COVID-19, el cual distorsiona las correlaciones de forma artificial. Los datos que se omiten en el estudio de las correlaciones sin el COVID-19 son los correspondientes a los tres primeros trimestres del año 2020.

Los resultados de este estudio pueden verse en el Cuadro 4.2. Para las series de tiempo que mostraron una mayor correlación con el PIB sin tener en cuenta el periodo de tiempo marcado por los efectos del COVID-19, se estudia la relación lineal entre todas ellas, sus correlaciones y la densidad de los datos de cada una de las series, lo que puede verse en la Figura (4.4).

En la Figura (4.4) se puede observar la distorsión que provocó el periodo del COVID-19, puede verse claramente en la nube de puntos PIB $\sim$ GASOLINAS como hay algunos puntos que podrían ser candidatos a considerarse atípicos, por ello se considera tener más en cuenta la correlación entre las series sin tener en cuenta el periodo afectado por la pandemia. Además, se obtiene la lectura de que algunas de estas variables presentan una alta correlación entre ellas y no solo con el PIB.

4.6.3. Modelo

Igual que para el modelo sg-MIDAS también se parametriza la matriz (4.1) en función de una función de pesos.

Como ya se comentaba en el modelo anterior, no todas las variables están actualizadas a la misma fecha, por lo tanto, no se utilizará el mismo rango de fechas para seleccionar los 12 retardos de los regresandos que se utilizan para modelar el PIB.

Seleccionadas las variables se procede con el ajuste del PIB con cada uno de los regresores de forma individual, además, se ajusta un modelo para todos los retardos seleccionados $1, 2, ..., 12$ y se selecciona el que minimiza el criterio de información de Akaike.

Por último, se realiza la combinación de predicciones como se comentaba en secciones anteriores.

4.7. Mejores Resultados

En esta sección se recogen los resultados en los dos escenarios estudiados (preCOVID y post COVID) de las versiones de cada uno de los modelos que han tenido un mejor desempeño en función de su RMSE, el resto de resultados puede verse en el Apéndice. Cabe destacar que se contemplo la posibilidad de

	CON COVID-19	SIN COVID-19
PIB	1.00	1.00
TRAB_TOTAL	0.48	0.50
PMI_INDUS	-0.10	0.09
PMI_SERV	-0.39	-0.34
ICM	-0.19	0.17
ICM_SIN_ESTACIONES	-0.14	0.28
TRAB_AGRI	0.16	0.22
TRAB_INDU	0.41	0.41
TRAB_CONST	0.03	0.41
TRAB_SERV	0.51	0.32
C_CEMENTO	-0.14	0.26
PERNOC_TOTAL	0.55	-0.02
C_GASOLINAS	0.39	0.72
C_ELECT_TOTAL	0.62	0.55
C_ELECT_INDUS	0.44	0.41
C_ELECT_SERV	0.70	0.66
C_ELECT_CONST	0.44	0.29
C_ELECT_AGRI	0.22	0.26
GASÓLEOS_AUTO	0.37	0.65
PERNOC_NO_RESIDENTES	0.54	-0.02
C_GASOIL_AGRI	0.09	0.00
I_PROD_CONST	-0.01	-0.00
ICN_SERV	0.43	0.61
ICN_COMERCIO	0.23	0.49
ICN_OTROS_SERV	0.71	0.71
ICN_TRANS	0.73	0.60
ICN_HOST	0.55	0.75
ICN_INFO	0.62	0.62
IPI	0.57	0.80
ED_CRÉDITOS	-0.27	0.18
VENTAS_G_EMPRESAS	0.38	0.58
EDIFICIOS_VISADOS	0.30	0.53
VIVIENDAS_VISADAS	0.27	0.46

Cuadro 4.2: Correlación de las variables con el PIB.

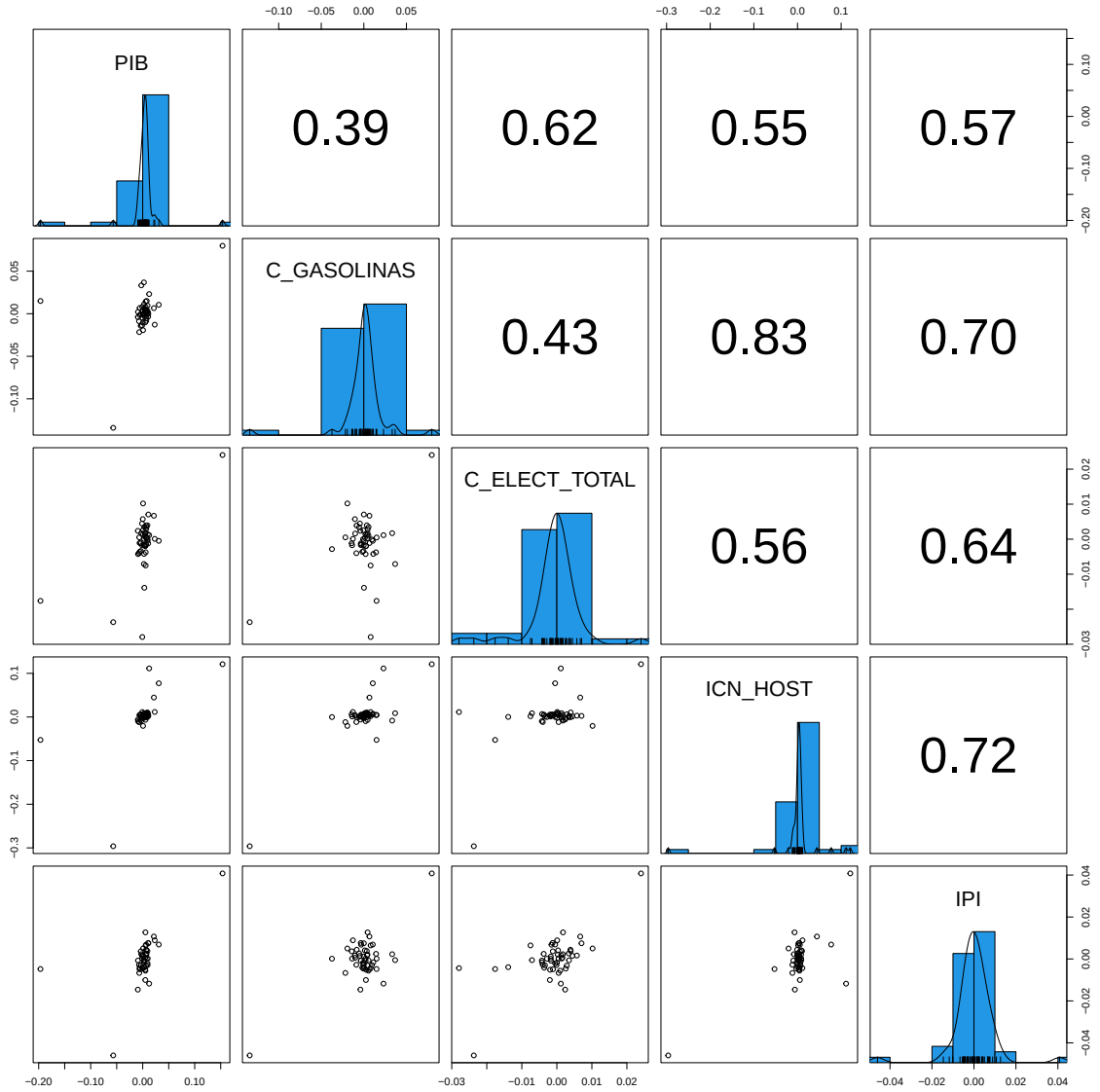

Figura 4.4: Correlaciones y densidades de las variables con mayor correlación con el PIB.

incluir en esta comparación los modelos construidos sin haber diferenciado los datos para intentar mantener la información de las series de tiempo sobre la tendencia y así obtener un mejor ajuste, esta consideración fue descartada al obtener resultados similares que con los modelos construidos de la forma actual.

Para presentar los resultados se separan las predicciones entre las divisiones de muestra en segmentos fijos y *recursive forecasting* y *rolling forecasting*, esto se debe a que no sería una comparación justa, a pesar de que se testea la precisión de las predicciones en el mismo periodo solo se tendría la misma muestra de entrenamiento en la primera predicción, ya que, como se comentaba anteriormente para la división de muestra en segmentos fijos la muestra de entrenamiento no aumenta, para las otras dos divisiones si que aumenta a medida que avanza el horizonte de predicción. Dicho esto, resulta interesante la comparación entre *recursive forecasting* y *rolling forecasting* para estudiar si se los resultados son

similares entre ellas y utilizar en ese caso preferiblemente *rolling forecasting* utilizando así una muestra de entrenamiento más pequeña que en el *recursive forecasting* reduciendo de este modo los tiempos de aprendizaje.

4.7.1. Escenario preCOVID

En la Figura 4.5 pueden verse los mejores modelos en la división segmentos fijos. Ninguno de los modelos parece ajustarse demasiado bien a las observaciones, este mal ajuste de los 4 modelos puede observarse numéricamente viendo las medidas de precisión del Cuadro 4.3, dentro de estos resultados, el polinomio de Nealmon ha sido el que mejores resultados obtuvo para todos los modelos excepto para el sg-MIDAS.

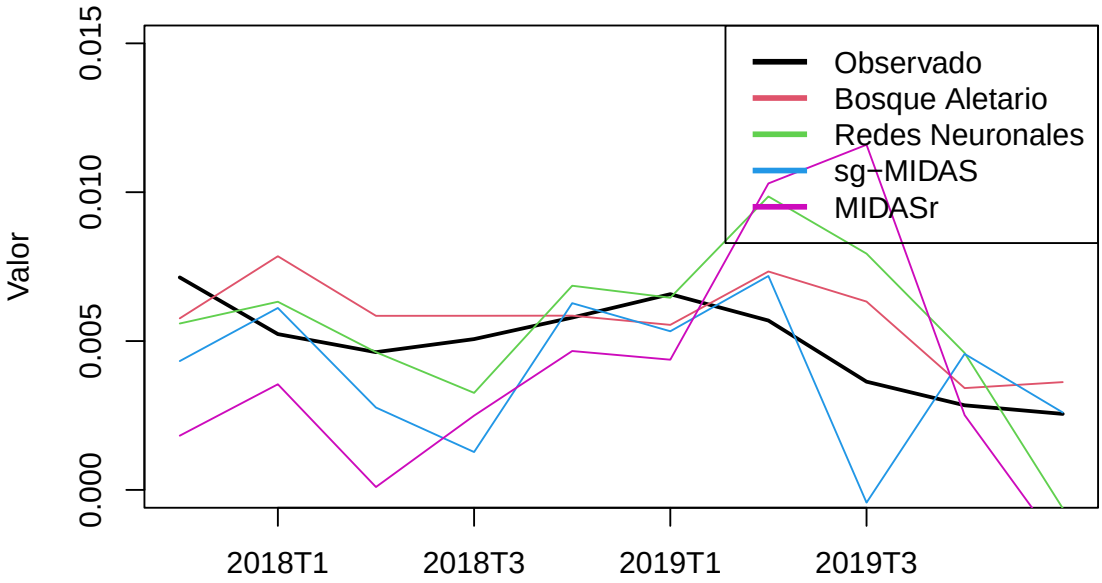

Figura 4.5: Predicciones con las mejores configuraciones de cada modelo para el periodo preCOVID con las divisiones de segmentos fijos.

En la Figura 4.6 y el Cuadro 4.4 se ven los resultados con las divisiones *recursive forecasting* y *rolling forecasting*, se aprecia una reducción del RMSE con respecto a la división de muestra anterior pero no lo suficiente como para proporcionar buenas predicciones, en este caso los modelos que utilizan polinomios de Legendre son los que obtienen mejores resultados y solamente para los modelos MIDASr se puede reducir el tiempo de aprendizaje y obtener mejores resultados utilizando *rolling forecasting*

4.7.2. Escenario postCOVID

En la Figura 4.7 y en el Cuadro 4.5 se observan los resultados obtenidos con la división segmentos fijos para el escenario postCOVID, aunque los resultados no son muy buenos, son bastante mejores que para el escenario anterior, destaca el modelo sg-MIDAS es capaz de detectar casi por completo la

	ME	RMSE	MAE	MPE	MAPE	pseudo R^2
Bosques Aleatorios	-0.00	0.0015	0.00	-20.15	27.91	-0.52
Redes Neuronales	-0.00	0.0022	0.00	-25.95	39.10	-2.22
sg-MIDAS	0.00	0.0023	0.00	19.22	44.15	-2.43
MIDASr	0.00	0.0040	0.00	2.22	68.86	-9.22

Cuadro 4.3: Medidas de precisión de las predicciones en la muestra de test para el periodo preCOVID con las divisiones de segmentos fijos.

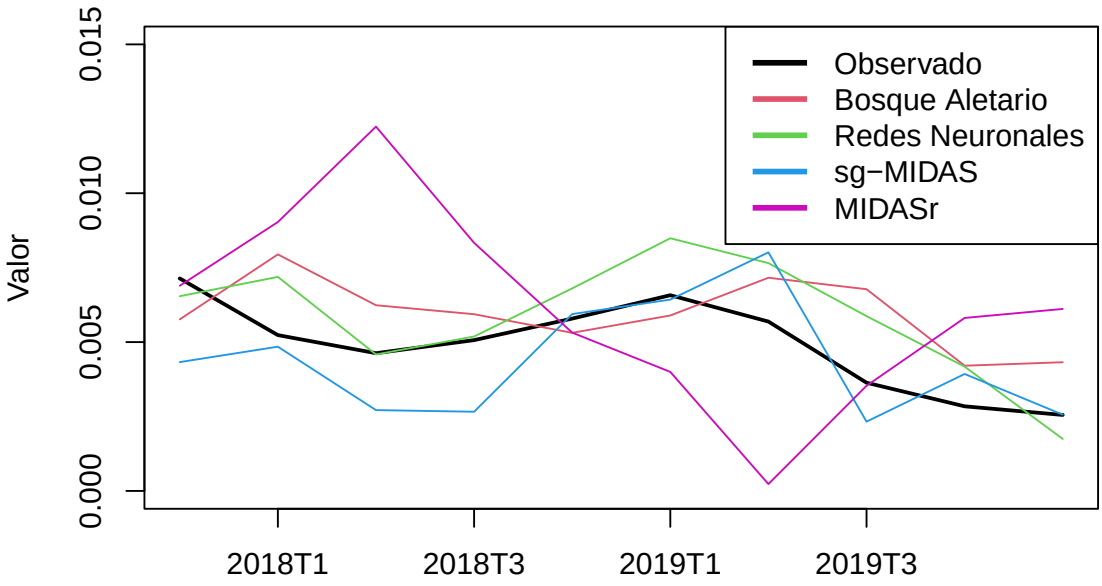

Figura 4.6: Predicciones con las mejores configuraciones de cada modelo para el periodo preCOVID con las divisiones de *recursive*recursive y *rolling forecasting*.

variación negativa del segundo trimestre de 2020. Indicar que en este caso los mejores resultados se han obtenido al utilizar el polinomio de Nealmon.

Por último, en la Figura 4.8 y el Cuadro 4.6 se observan los resultados en las divisiones de muestra *recursive forecasting* y *rolling forecasting* para el escenario postCOVID, esta vez los modelos no son capaces de reducir el RMSE con respecto a la división segmentos fijos, es posible que esto se deba a la inclusión en la muestra de entrenamiento de los efectos del COVID. Los mejores resultados se obtienen al utilizar polinomios de Legendre y para Redes Neuronales y MIDASr se podría utilizar *rolling forecasting* y se seguiría obteniendo el mejor resultado.

Teniendo en cuenta el desempeño de los modelos en este escenario se decide obtener la predicción para el primer trimestre de 2023 con los modelos construidos con polinomios Nealmon y con un tamaño

	ME	RMSE	MAE	MPE	MAPE	pseudo R^2
Bosques Aleatorios	-0.00	0.0017	0.00	-25.14	33.56	-0.86
Redes Neuronales	-0.00	0.0014	0.00	-24.46	26.52	-0.34
sg-MIDAS	0.00	0.0016	0.00	10.23	28.40	-0.77
MIDASr	-0.00	0.0034	0.00	-67.99	67.99	-6.49

Cuadro 4.4: Medidas de precisión de las predicciones en la muestra de test para el periodo preCOVID con las divisiones de *recursive*recursive y *rolling forecasting*.

	ME	RMSE	MAE	MPE	MAPE	pseudo R^2
Bosques Aleatorios	-0.01	0.0698	0.04	-137321817.22	137321871.30	0.06
Redes Neuronales	0.01	0.0622	0.05	-32631571.49	148363460.33	0.25
sg-MIDAS	0.00	0.0378	0.02	-51505377.82	51505471.46	0.72
MIDASr	-0.00	0.0491	0.03	-84924521.40	84924643.44	0.53

Cuadro 4.5: Medidas de precisión de las predicciones en la muestra de test para el periodo postCOVID con las divisiones de segmentos fijos.

de muestra de entrenamiento fijo hasta antes del COVID. La predicción obtenida es muy similar para los 4 modelos y ronda entre una variación trimestral del cuarto trimestre del 2022 al primer trimestre de 2023 de entre el 0.90 % y el 0,91 %.

	ME	RMSE	MAE	MPE	MAPE	pseudo R^2
Bosques Aleatorios	-0.00	0.0768	0.04	-133978286.64	133978365.68	-0.14
Redes Neuronales	0.01	0.0683	0.05	-60068889.90	139789805.50	0.10
sg-MIDAS	-0.06	0.2356	0.12	-88243071.32	176957297.22	-9.69
MIDASr	-0.02	0.0608	0.04	-132253491.64	136868071.53	0.29

Cuadro 4.6: Medidas de precisión de las predicciones en la muestra de test para el periodo postCOVID con las divisiones de *recursive*recursive y *rolling forecasting*.

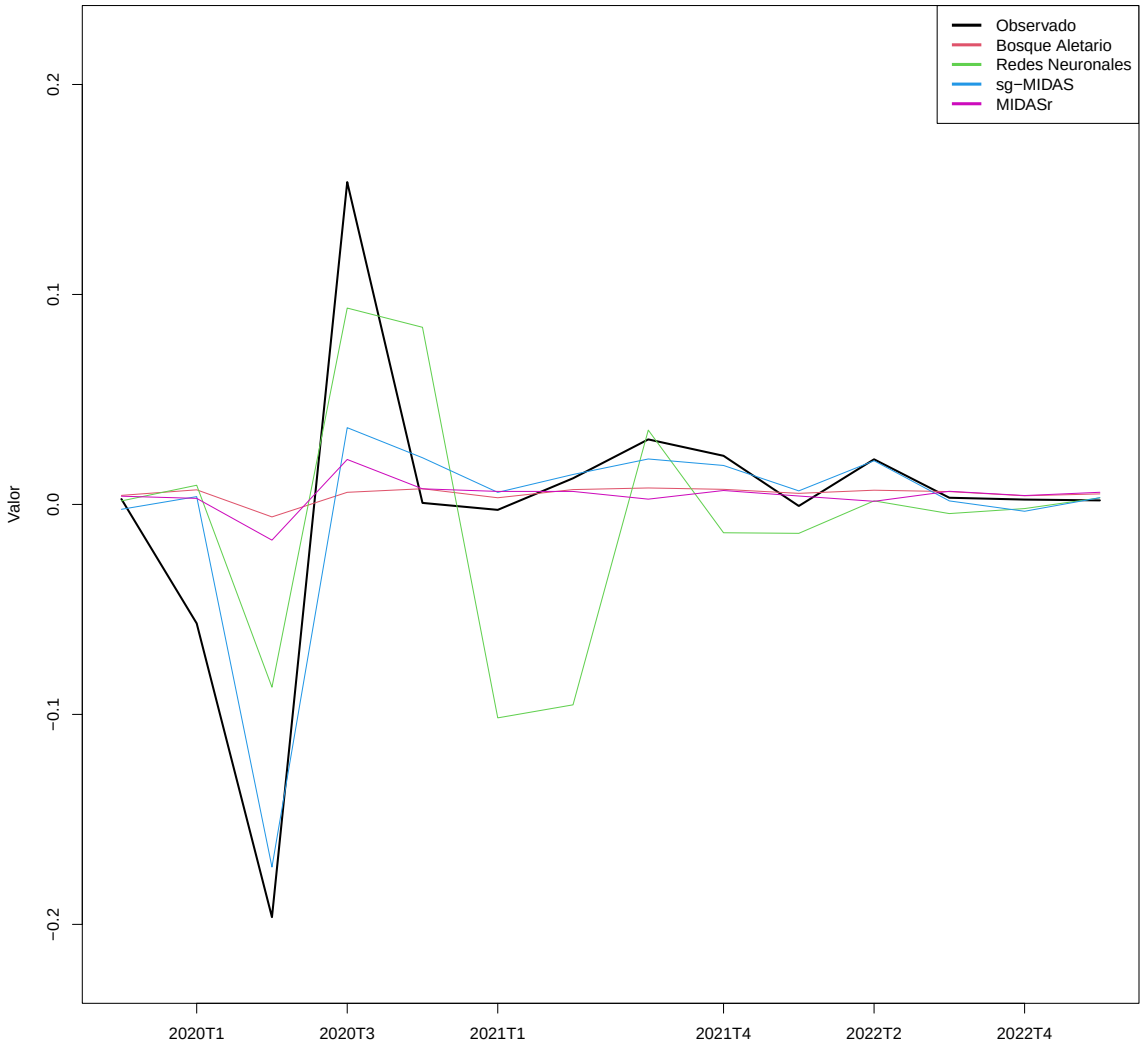

Figura 4.7: Predicciones con las mejores configuraciones de cada modelo para el periodo postCOVID con las divisiones de segmentos fijos.

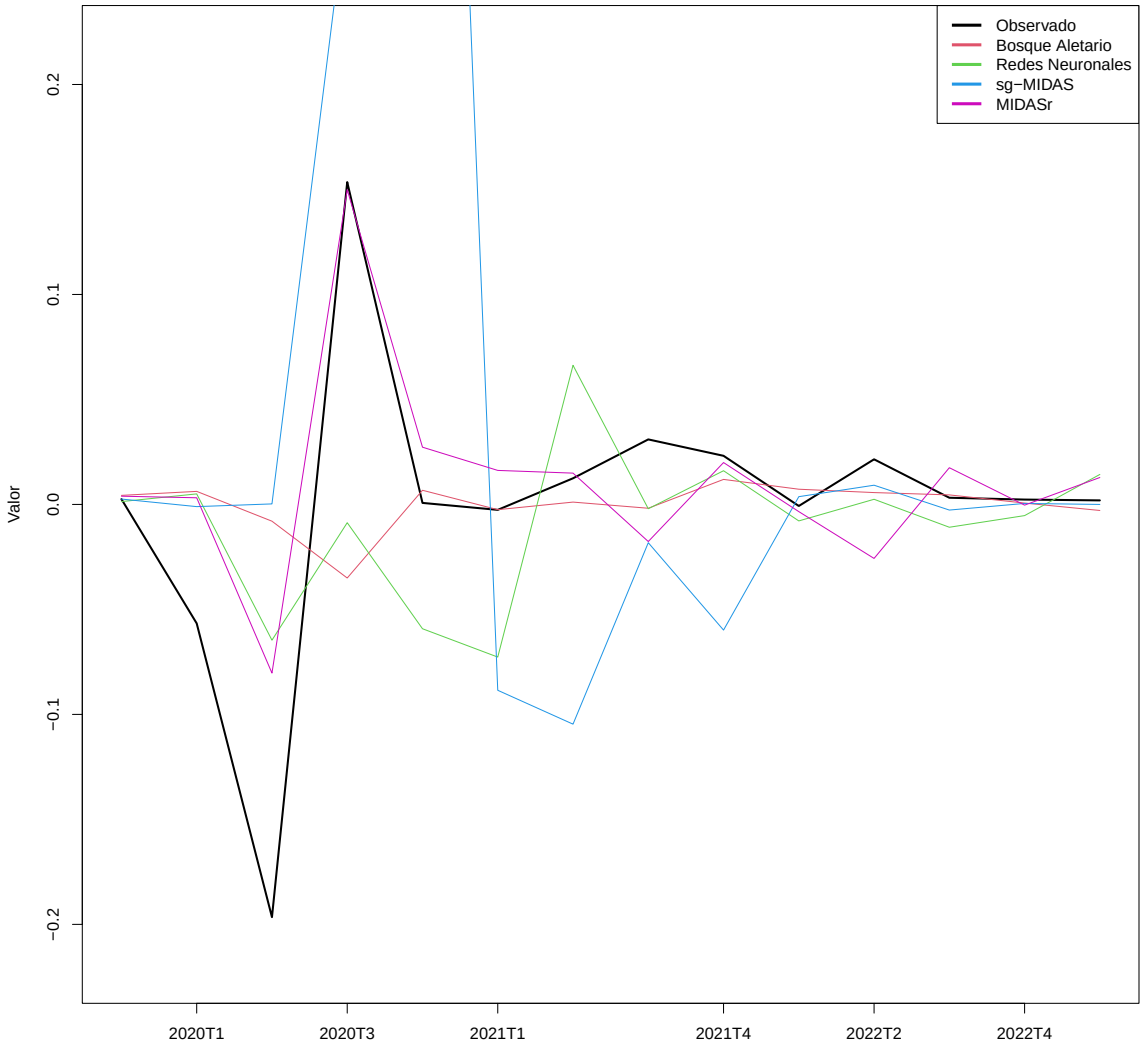

Figura 4.8: Predicciones con las mejores configuraciones de cada modelo para el periodo postCOVID con las divisiones de *recursive*recursive y *rolling forecasting*.

Capítulo 5

Conclusiones y líneas futuras

5.1. Conclusiones

El objetivo del TFM es disponer de un mecanismo de predicción anticipado y rápido, ejecutable en tiempo real, de evolución del PIB. Para ello se han desarrollado cuatro modelos que proporcionan una estimación con los datos más recientes de una serie de variables. Ello permite su cómputo de manera inmediata, en tiempo real, una vez conocidas las variables e ir actualizándolas a medida que se van publicando nuevos datos. En concreto, se extraen 33 variables de 10 fuentes de datos distintas y con fechas de publicación distintas (diciembre, enero y febrero). Este mecanismo se ha validado con la predicción de la evolución del PIB en el primer trimestre de 2023, el dato observado es de una variación del 0,47 % mientras que la predicción obtenida es de cerca del 0,90 % lo que no se aleja demasiado de la realidad.

5.2. Líneas futuras

Al margen del resultado obtenido para el primer trimestre de 2023 y teniendo en cuenta las medidas de precisión calculadas para los dos escenarios de validación de los modelos, no se han obtenido unos resultados deseables, por ello se citan algunas nuevas líneas de estudio que podrían mejorar el mecanismo construido:

1. Existen dos enfoques distintos para predecir el PIB. El enfoque utilizado como base en este trabajo es el enfoque directo, que consiste en la predicción del PIB agregado. Sin embargo, también existe otro enfoque denominado enfoque indirecto, descrito en el estudio de Fernández Cerezo (2023). Este enfoque se basa en la predicción de los componentes que conforman el PIB y, posteriormente, agregar estos valores. Por ejemplo, si consideramos el cálculo del PIB desde el lado de la oferta, los componentes principales serían los Valores Añadidos Brutos (VAB) por sector de actividad, que se definen como el valor total de los bienes y servicios de un sector de actividad específico producidos dentro de un área geográfica concreta.

   En futuras investigaciones, sería interesante explorar y comparar en mayor profundidad estos dos enfoques. También se podría considerar la inclusión de variables adicionales o añadir variables medidas a una frecuencia más alta que la frecuencia mensual, para tratar de capturar mejor las complejidades de la serie del PIB.

2. Uno de los motivos para el mal desempeño de los modelos puede a ver sido influenciado por la no inclusión de algún efecto no lineal de los predictores y a posibles sus interacciones. Por este motivo se podrían considerar en el futuro modelos aditivos con dependencia Wood (2017) que permitan capturar patrones más complejos en los datos.

Apéndice A

Resultados Bosques Aleatorios

En las Figuras A.1, A.2, A.3 y A.4 pueden verse los resultados de este modelo para los dos periodos de testeo.

En los Cuadros A.1 y A.2 se detallan algunas medidas de bondad del ajuste para el modelo en los dos rangos de fechas indicados anteriormente.

	ME	RMSE	MAE	MPE	MAPE	pseudo R^2
Fixed Nealmon	0.01	0.02	0.01	-2491265.89	4080163.76	-0.75
Fixed Legendre	0.00	0.01	0.00	-3587183.81	3587269.56	0.76
Recursive Nealmon	0.01	0.01	0.01	-8041383.66	10713293.41	-0.59
Recursive Legendre	0.02	0.02	0.02	27507061.10	45720808.81	-2.78
Rolling Nealmon	0.01	0.01	0.01	-8041383.66	10713293.41	-0.59
Rolling Legendre	0.01	0.02	0.02	7679126.46	18028486.56	-1.30

Cuadro A.1: Medidas de precisión de las predicciones en la muestra de test. Bosque Aleatorio.

	ME	RMSE	MAE	MPE	MAPE	pseudo R^2
Fixed Nealmon	-0.00	0.00	0.00	-32.34	32.48	-0.53
Fixed Legendre	-0.00	0.00	0.00	-37.86	40.61	-1.10
Recursive Nealmon	-0.00	0.00	0.00	-37.43	37.83	-0.91
Recursive Legendre	-0.00	0.00	0.00	-35.20	39.01	-0.93
Rolling Nealmon	-0.00	0.00	0.00	-25.25	34.58	-0.30
Rolling Legendre	-0.00	0.00	0.00	-24.14	34.11	-0.27

Cuadro A.2: Medidas de precisión de las predicciones en la muestra de test. Bosque Aleatorio preCOVID.

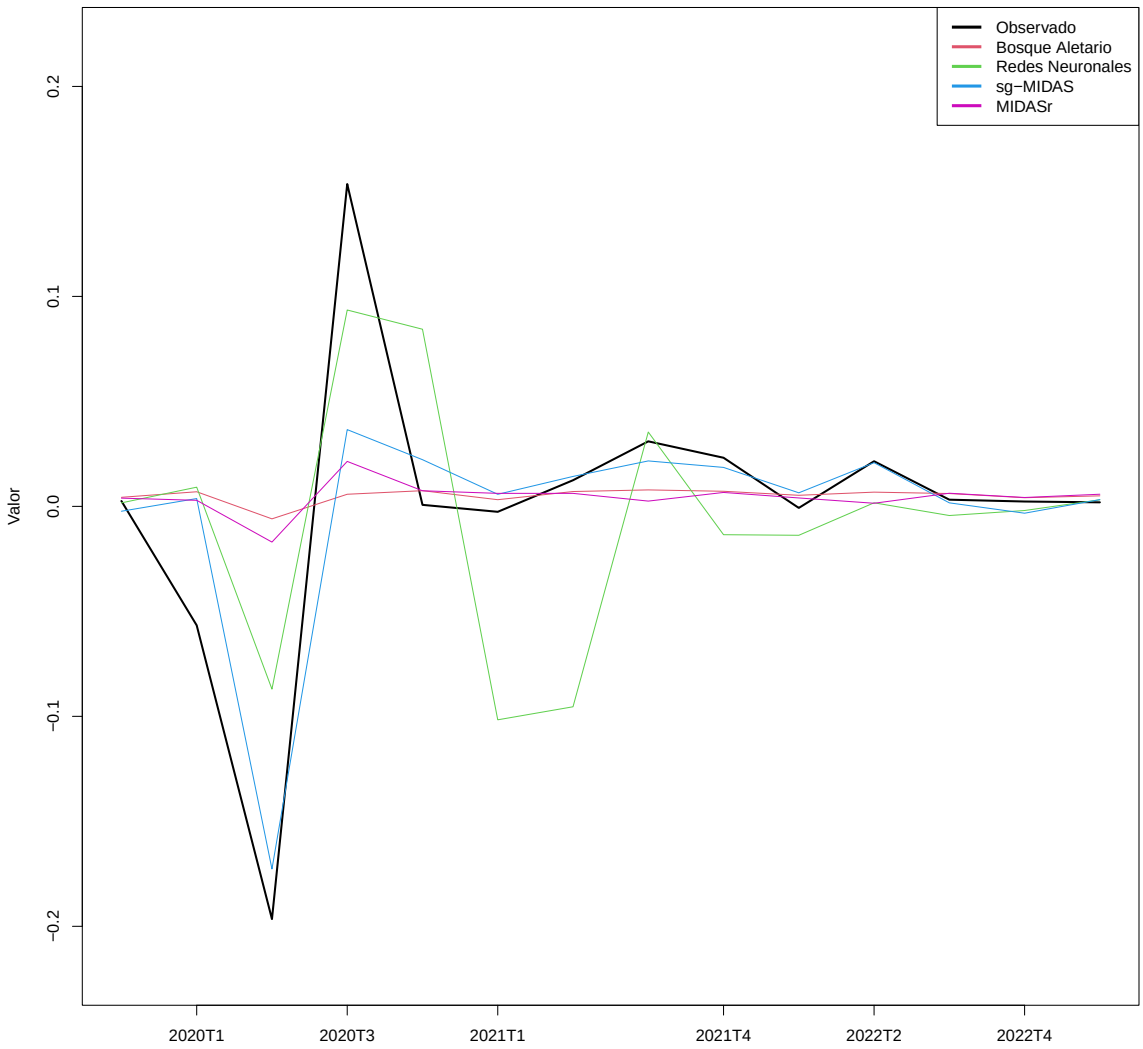

Figura A.1: Resultados Bosque Aleatorio.

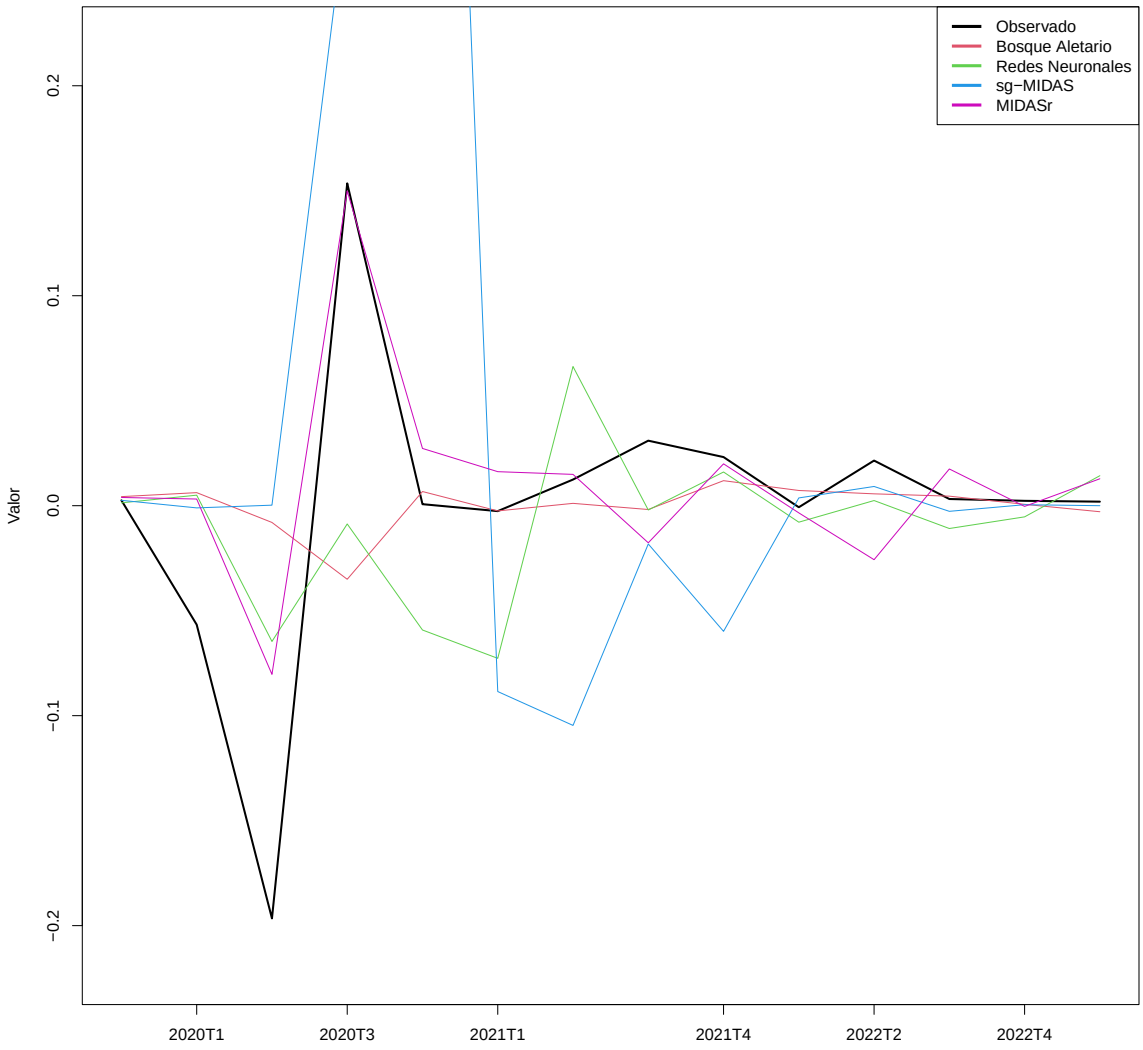

Figura A.2: Gráfico de dispersión de observaciones frente a predicciones (incluyendo la identidad, línea continua, y el ajuste lineal, línea discontinua). Bosque Aleatorio.

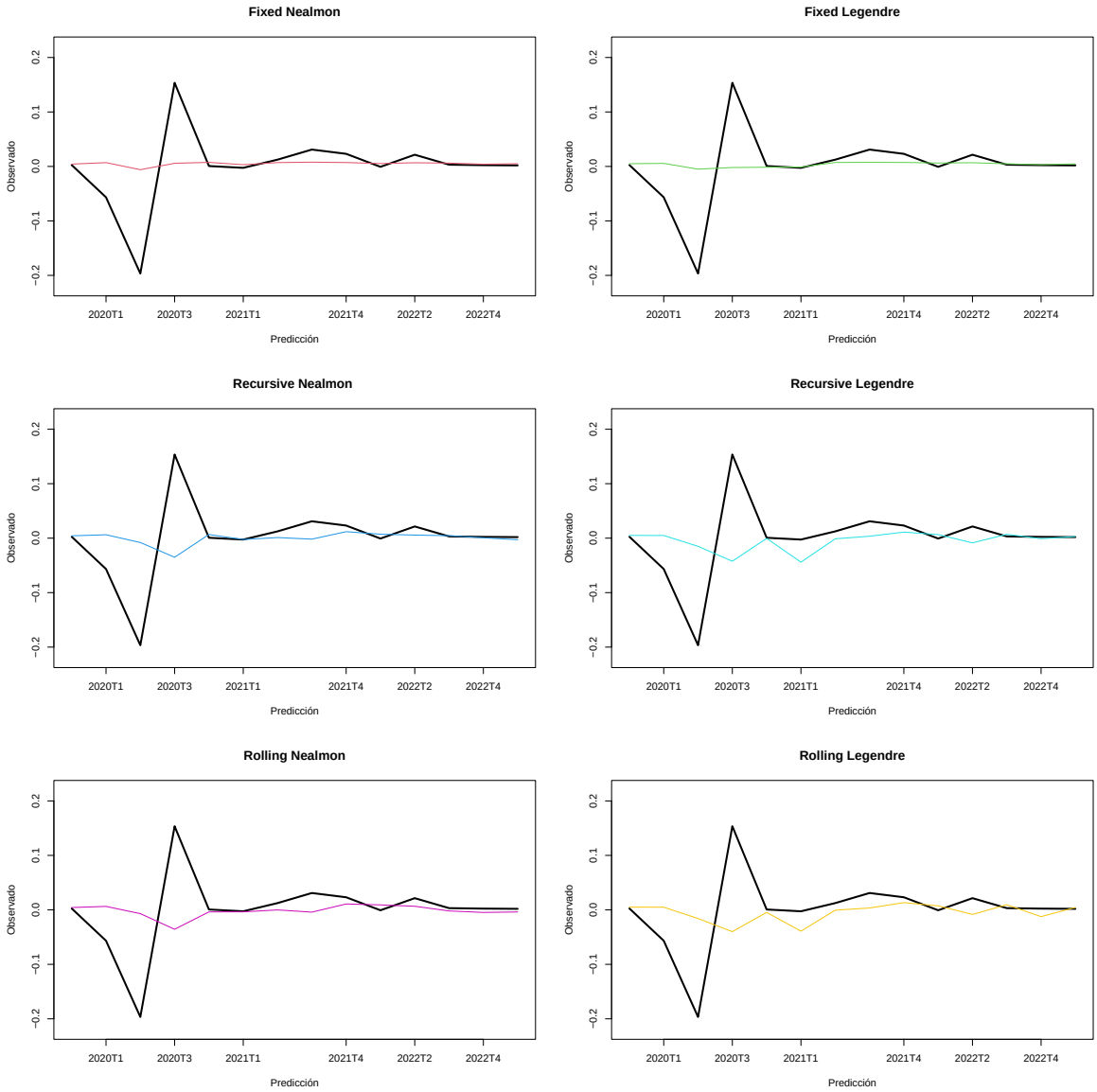

Figura A.3: Resultados Bosque Aleatorio preCOVID.

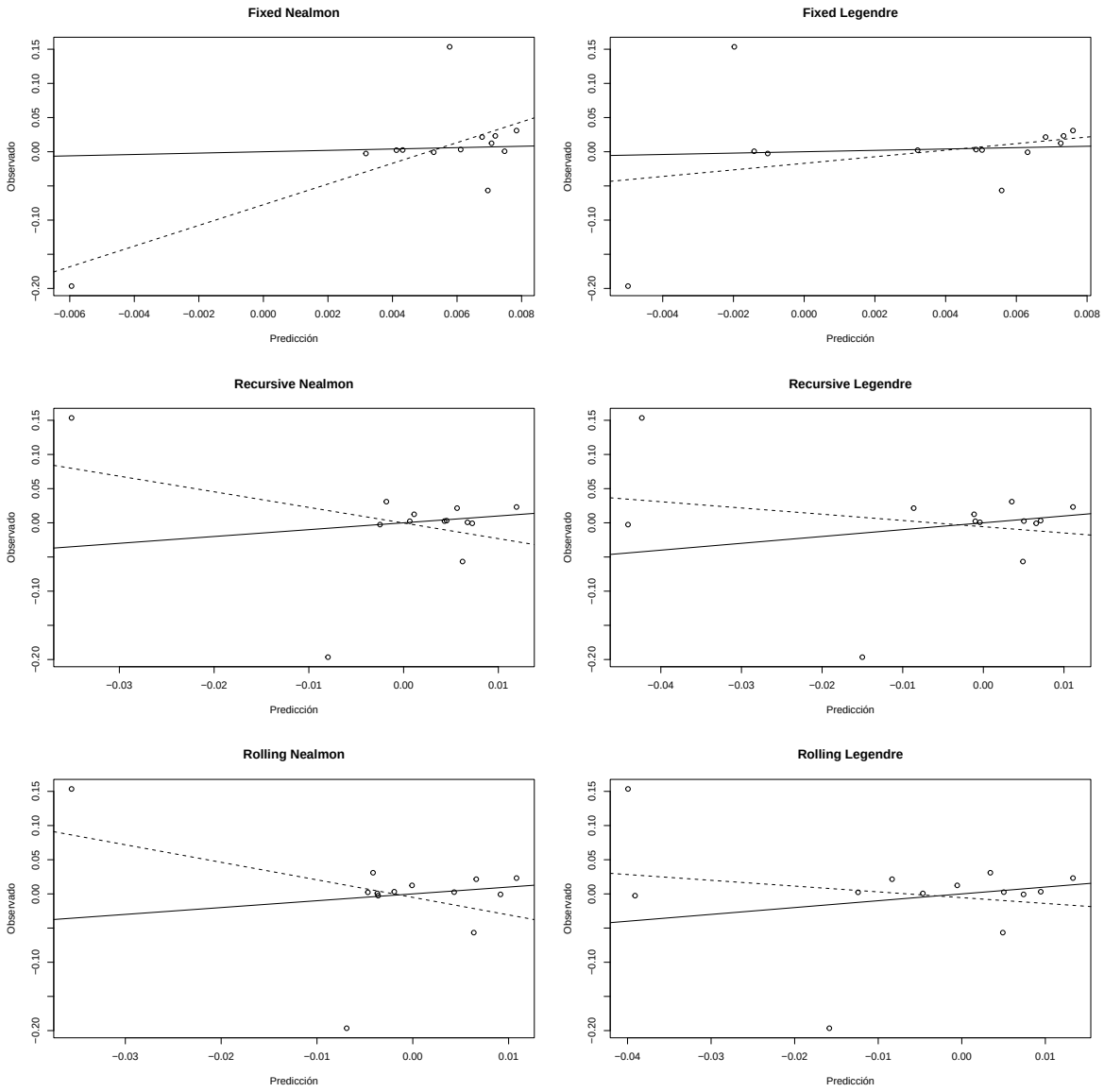

Figura A.4: Gráfico de dispersión de observaciones frente a predicciones (incluyendo la identidad, línea continua, y el ajuste lineal, línea discontinua). Bosque Aleatorio preCOVID.

Apéndice B

Resultados Redes Neuronales

En las Figuras B.1, B.2, B.3 y B.4 pueden verse los resultados de este modelo para los dos periodos de testeo.

En los Cuadros B.1 y B.2 se detallan algunas medidas de bondad del ajuste para el modelo en los dos rangos de fechas indicados anteriormente.

	ME	RMSE	MAE	MPE	MAPE	pseudo R^2
Fixed Nealmon	0.01	0.02	0.01	29985654.27	29985654.27	-0.95
Fixed Legendre	0.00	0.01	0.01	-1925140.05	8751124.80	0.67
Recursive Nealmon	0.01	0.02	0.01	29340188.66	41758448.10	-1.85
Recursive Legendre	0.00	0.03	0.02	-4624187.70	13791823.46	-4.08
Rolling Nealmon	0.01	0.02	0.01	29340188.66	41758448.10	-1.85
Rolling Legendre	0.02	0.04	0.03	87450596.52	87450673.48	-9.47

Cuadro B.1: Medidas de precisión de las predicciones en la muestra de test. Redes Neuronales.

	ME	RMSE	MAE	MPE	MAPE	pseudo R^2
Fixed Nealmon	-0.00	0.00	0.00	-43.38	54.98	-3.41
Fixed Legendre	-0.00	0.00	0.00	-20.08	31.19	-1.09
Recursive Nealmon	-0.00	0.00	0.00	-40.68	48.22	-2.09
Recursive Legendre	-0.00	0.00	0.00	-13.95	22.81	-0.17
Rolling Nealmon	-0.00	0.00	0.00	-30.82	40.28	-0.83
Rolling Legendre	-0.00	0.00	0.00	-22.09	31.86	-0.12

Cuadro B.2: Medidas de precisión de las predicciones en la muestra de test. Redes Neuronales preCOVID.

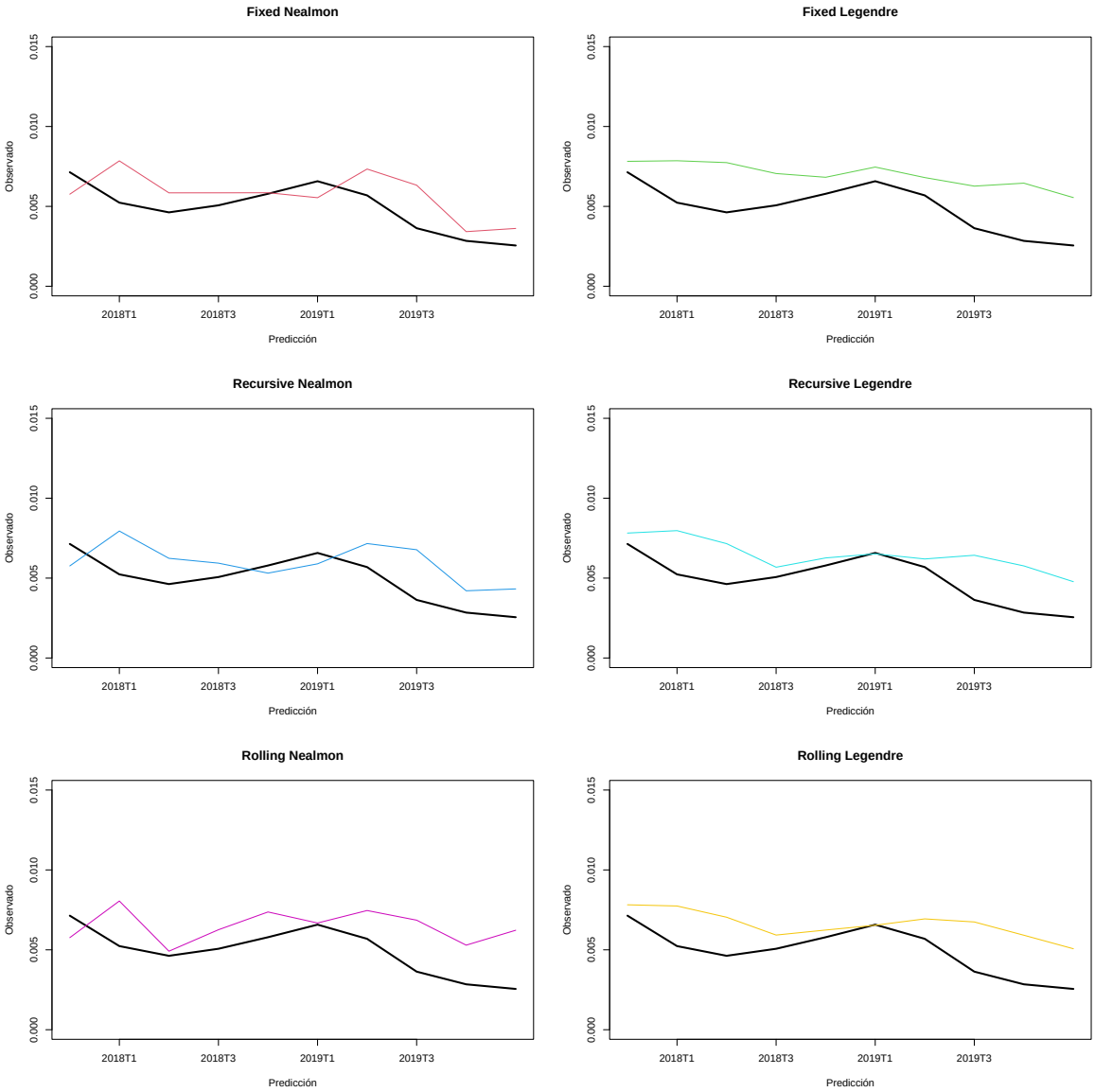

Figura B.1: Resultados Redes Neuronales.

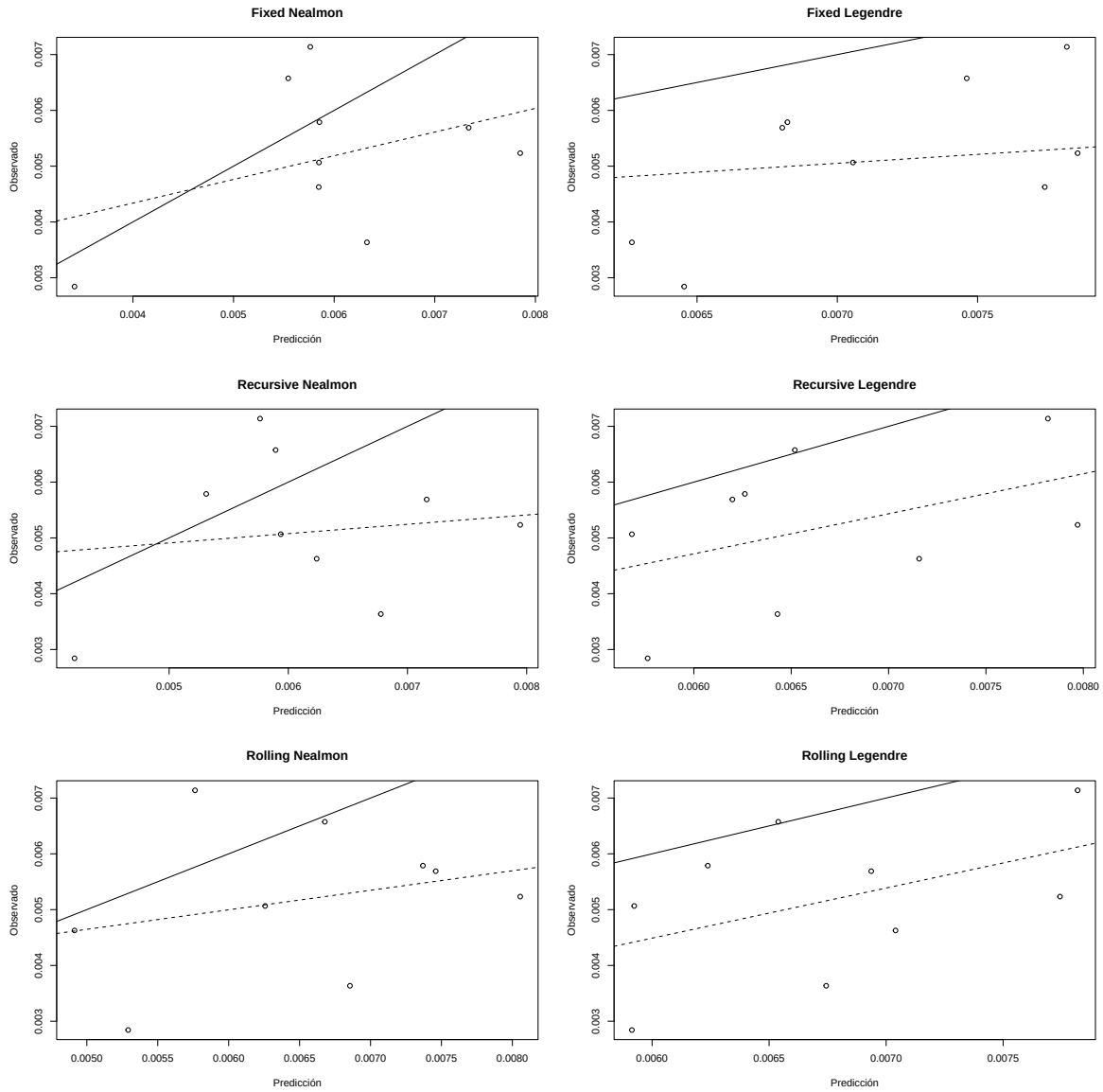

Figura B.2: Gráfico de dispersión de observaciones frente a predicciones (incluyendo la identidad, línea continua, y el ajuste lineal, línea discontinua).Redes Neuronales.

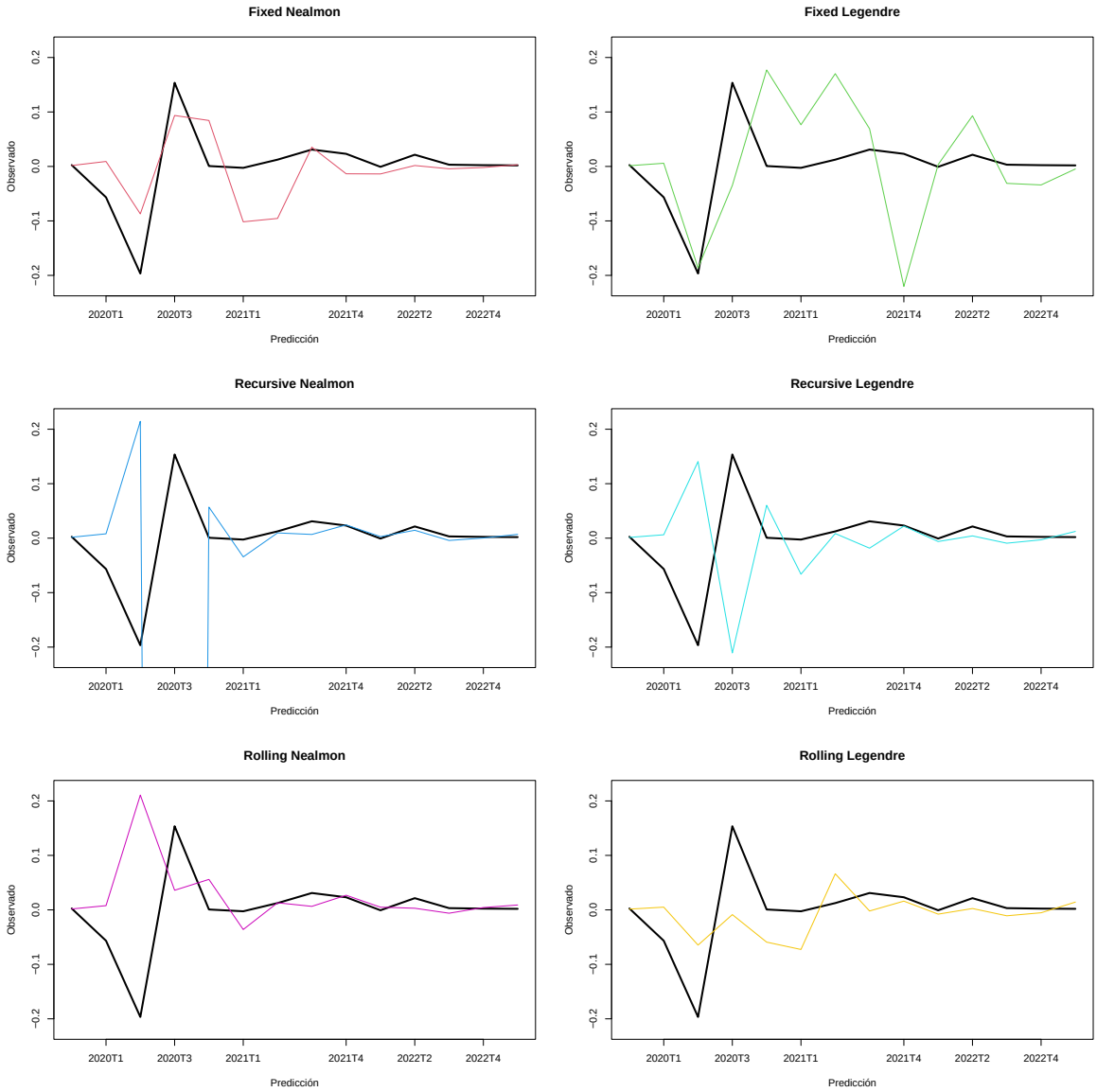

Figura B.3: Resultados Redes Neuronales preCOVID.

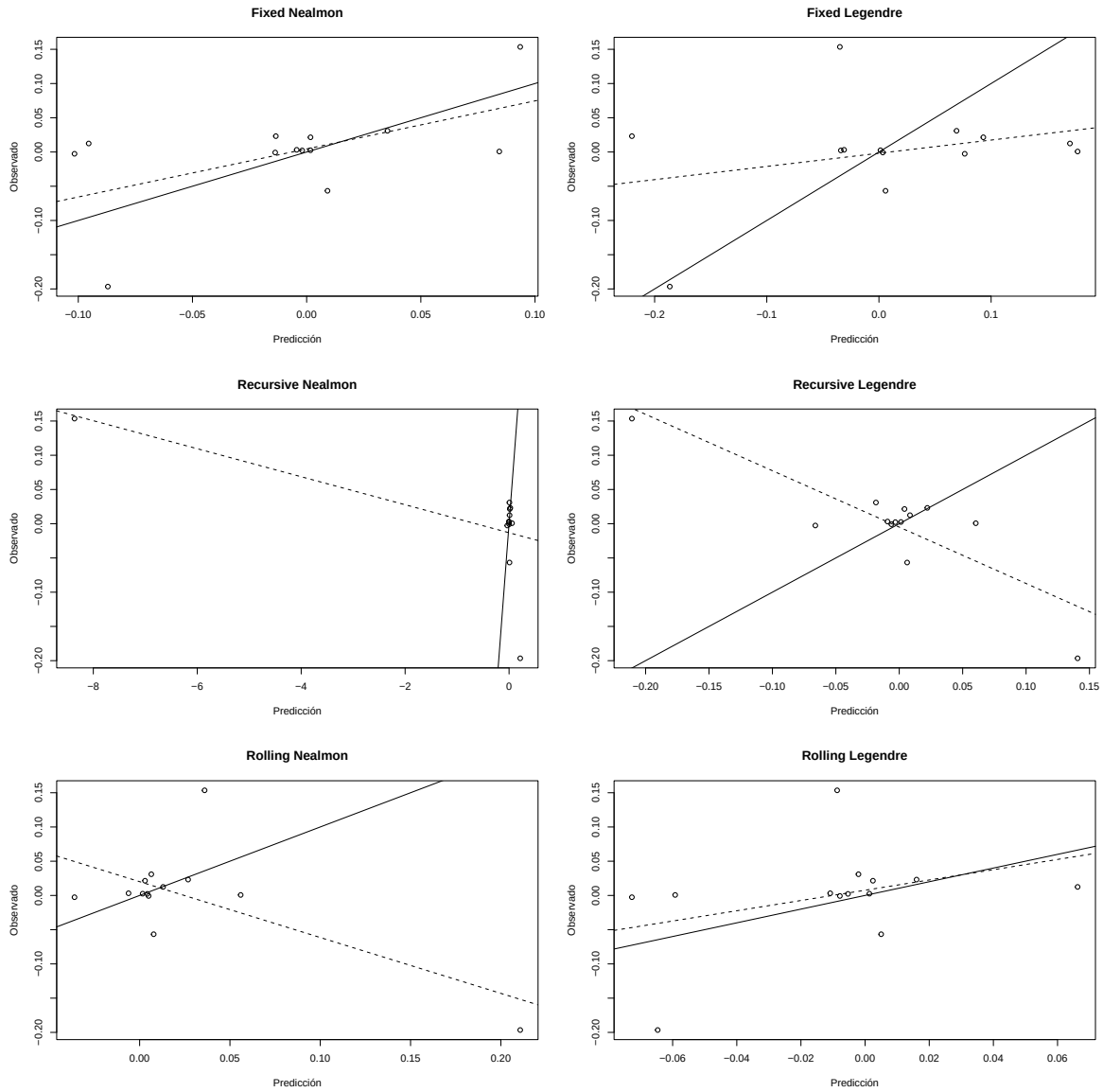

Figura B.4: Gráfico de dispersión de observaciones frente a predicciones (incluyendo la identidad, línea continua, y el ajuste lineal, línea discontinua). Redes Neuronales preCOVID.

Apéndice C

Resultados sg-MIDAS

En las Figuras C.1, C.2, C.3 y C.4 pueden verse los resultados de este modelo para los dos periodos de testeo.

En los Cuadros C.1 y C.2 se detallan algunas medidas de bondad del ajuste para el modelo en los dos rangos de fechas indicados anteriormente.

	ME	RMSE	MAE	MPE	MAPE	pseudo R^2
Fixed Nealmon	0.01	0.02	0.01	-2797156.60	2797410.26	-1.25
Fixed Legendre	0.01	0.01	0.01	-8144005.29	8144153.16	-0.21
Recursive Nealmon	0.01	0.02	0.01	-2797156.60	2797410.26	-1.25
Recursive Legendre	0.05	0.06	0.05	68492545.35	75583142.78	-26.74
Rolling Nealmon	0.01	0.02	0.01	-2797156.60	2797410.26	-1.25
Rolling Legendre	0.02	0.03	0.02	65039781.60	66285488.44	-6.68

Cuadro C.1: Medidas de precisión de las predicciones en la muestra de test. sg-MIDAS.

	ME	RMSE	MAE	MPE	MAPE	pseudo R^2
Fixed Nealmon	0.00	0.00	0.00	54.99	58.23	-4.82
Fixed Legendre	0.00	0.00	0.00	14.97	29.92	-0.25
Recursive Nealmon	0.00	0.00	0.00	53.15	59.10	-5.02
Recursive Legendre	0.00	0.00	0.00	13.21	28.25	-0.15
Rolling Nealmon	0.01	0.01	0.01	137.28	137.28	-37.56
Rolling Legendre	0.00	0.00	0.00	58.70	61.70	-4.95

Cuadro C.2: Medidas de precisión de las predicciones en la muestra de test. sg-MIDAS preCOVID.

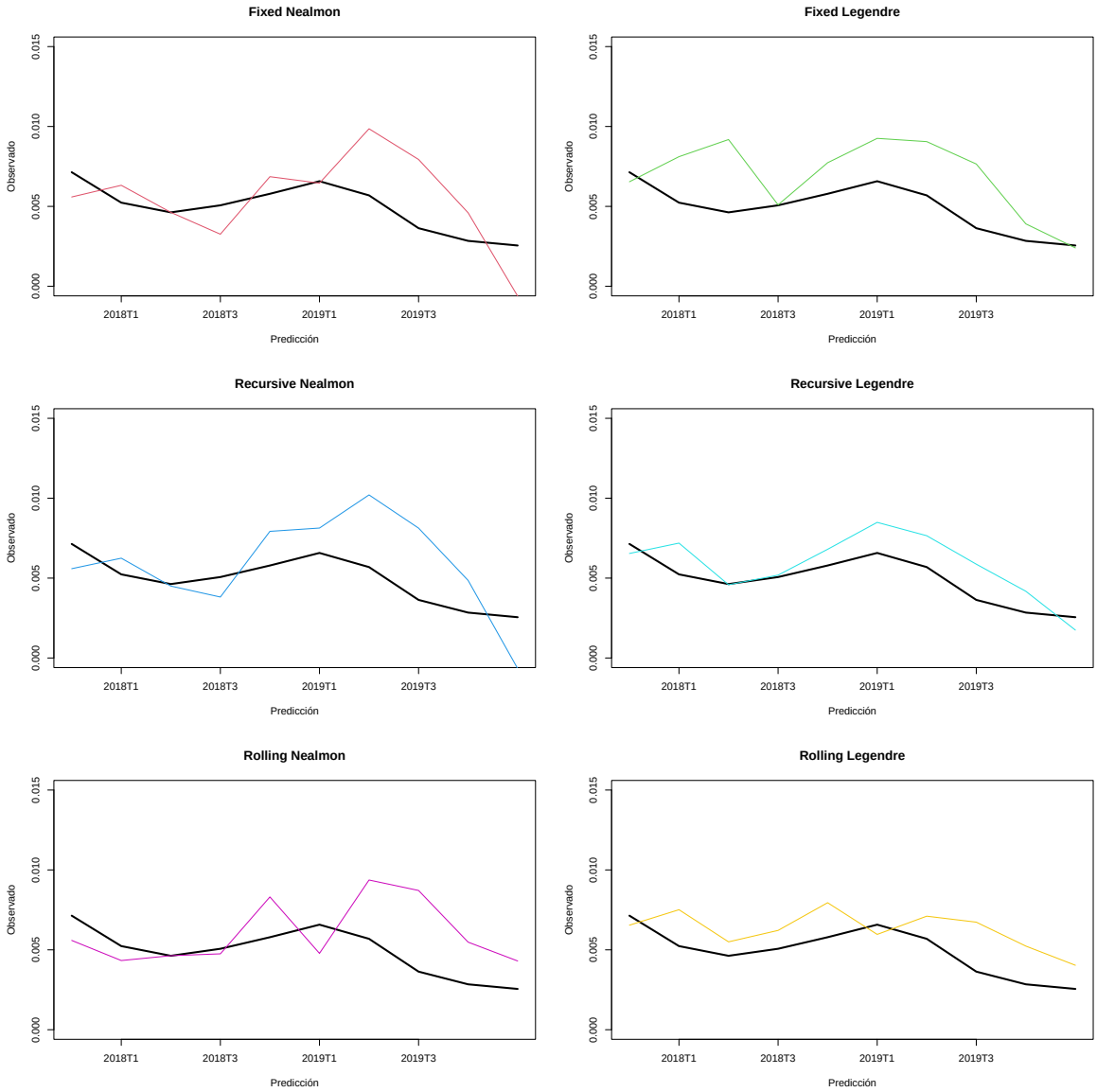

Figura C.1: Resultados sg-MIDAS.

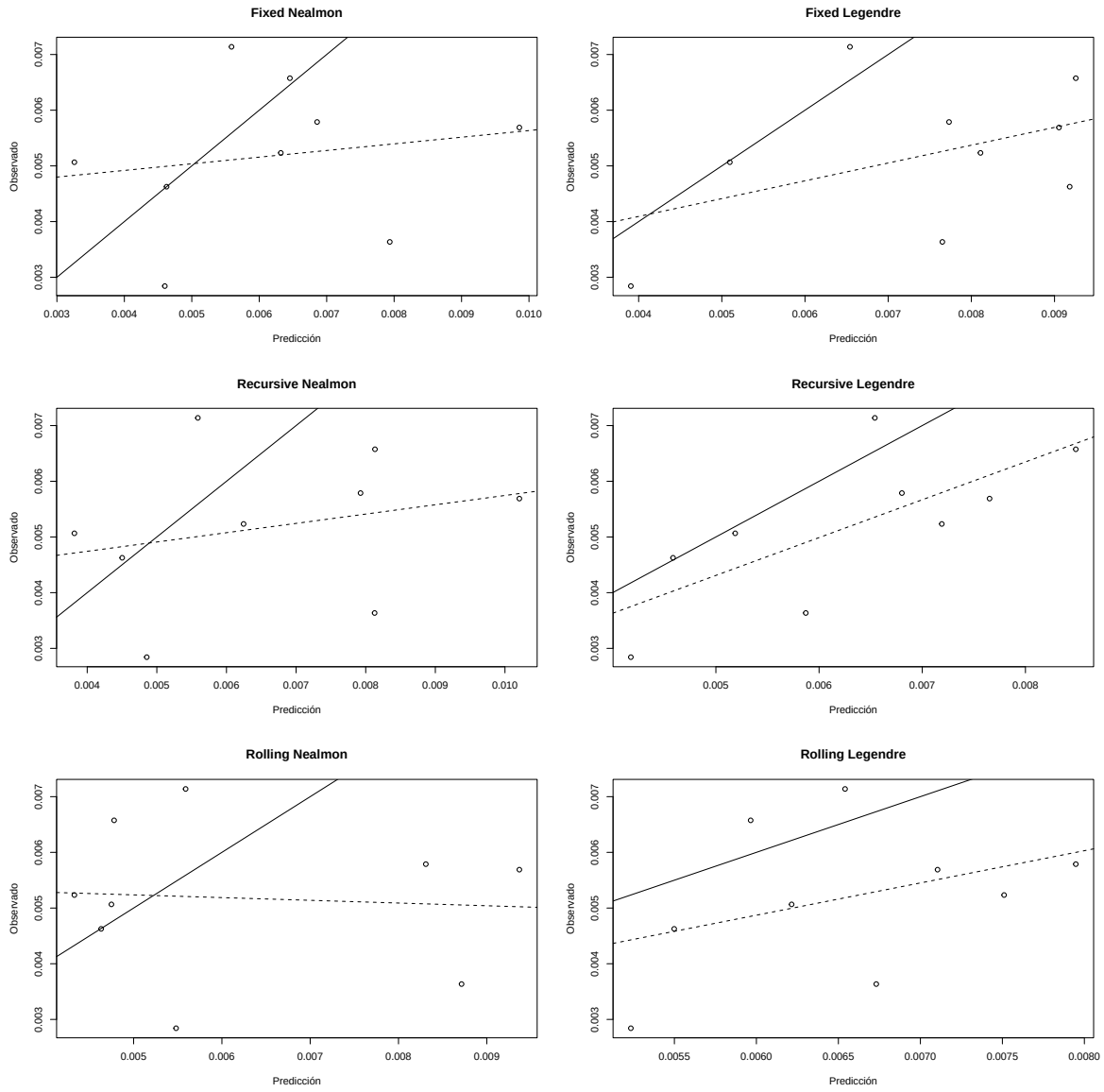

Figura C.2: Gráfico de dispersión de observaciones frente a predicciones (incluyendo la identidad, línea continua, y el ajuste lineal, línea discontinua). sg-MIDAS.

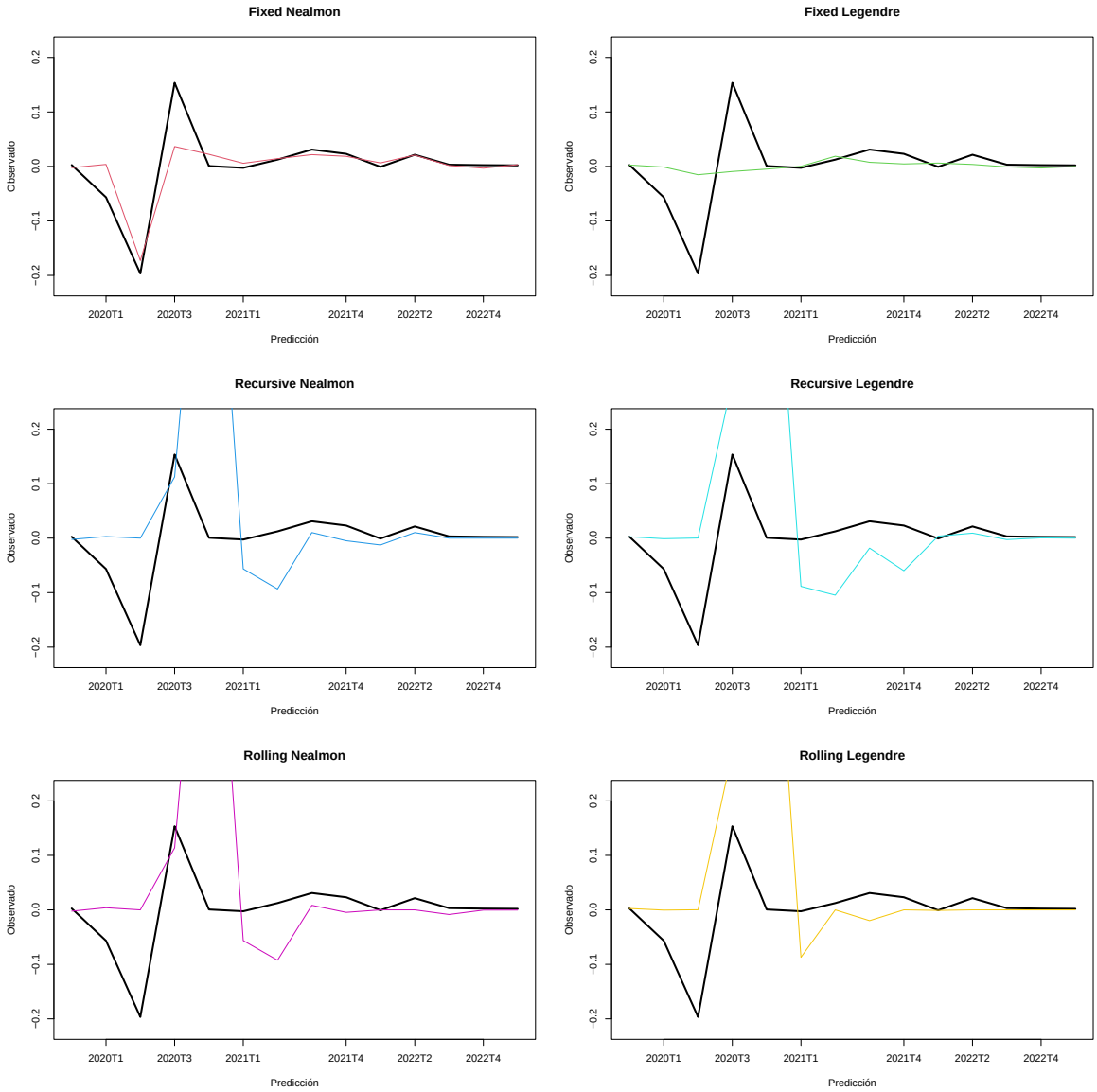

Figura C.3: Resultados sg-MIDAS preCOVID.

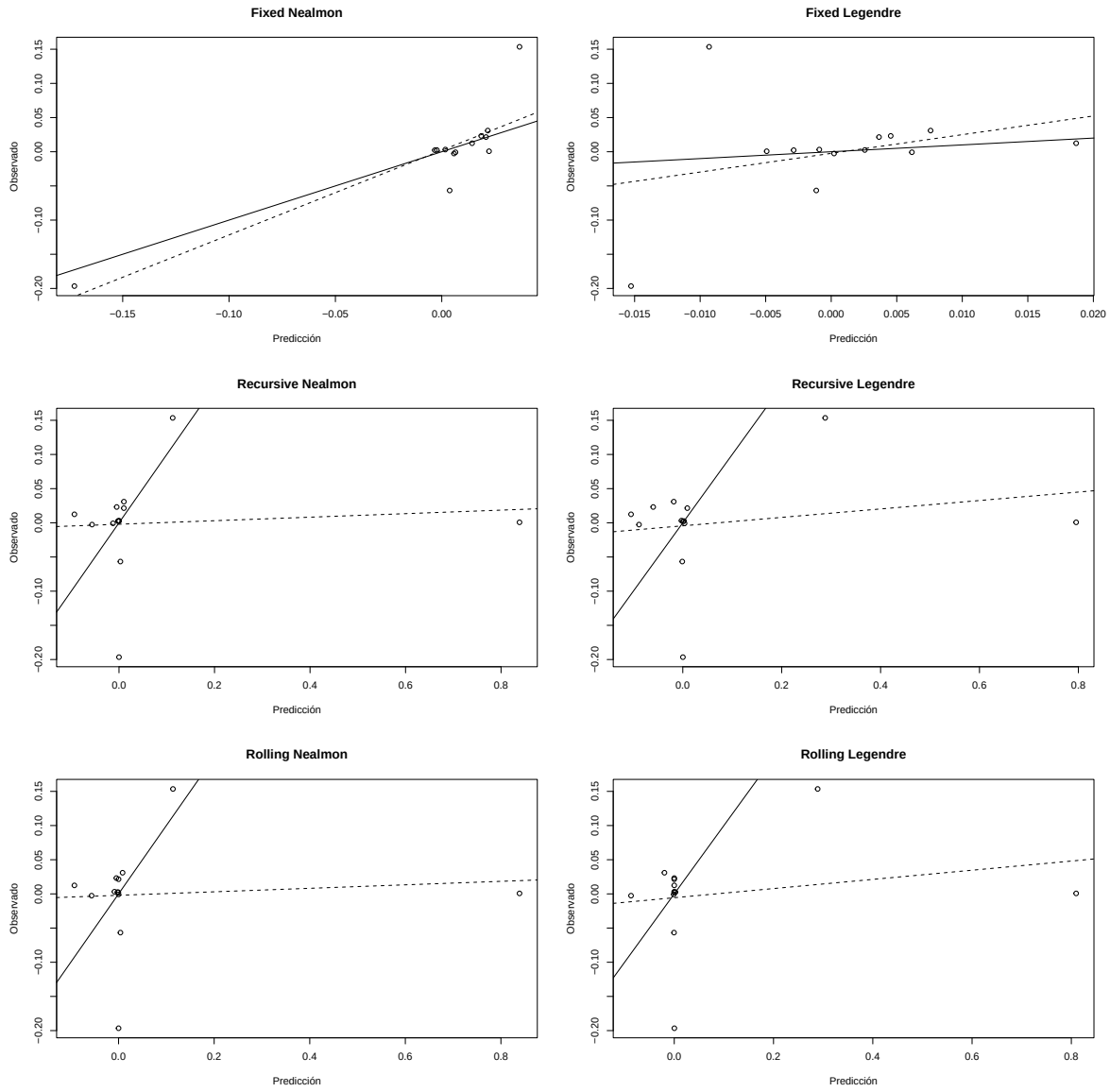

Figura C.4: Gráfico de dispersión de observaciones frente a predicciones (incluyendo la identidad, línea continua, y el ajuste lineal, línea discontinua). sg-MIDAS preCOVID.

Apéndice D

Resultados MIDASr

En las Figuras D.1, D.2, D.3 y D.4 pueden verse los resultados de este modelo para los dos periodos de testeo.

En los Cuadros D.1 y D.2 se detallan algunas medidas de bondad del ajuste para el modelo en los dos rangos de fechas indicados anteriormente.

	ME	RMSE	MAE	MPE	MAPE	pseudo R^2
Fixed Nealmon	0.01	0.02	0.01	-6855880.41	6855977.82	-0.75
Fixed Legendre	-0.02	0.02	0.02	-44219300.09	44219300.09	-3.46
Recursive Nealmon	0.01	0.03	0.02	-12794893.05	16573521.43	-3.48
Recursive Legendre	0.01	0.02	0.01	10068871.12	25535943.60	-0.78
Rolling Nealmon	0.01	0.02	0.01	-21527123.05	21527272.65	-1.51
Rolling Legendre	0.01	0.03	0.02	19725309.30	57388764.89	-3.61

Cuadro D.1: Medidas de precisión de las predicciones en la muestra de test. MIDASr.

	ME	RMSE	MAE	MPE	MAPE	pseudo R^2
Fixed Nealmon	0.00	0.00	0.00	-4.36	20.90	0.21
Fixed Legendre	-0.00	0.00	0.00	-24.20	29.58	0.02
Recursive Nealmon	-0.00	0.00	0.00	-8.19	22.83	0.06
Recursive Legendre	-0.00	0.00	0.00	-18.17	23.86	0.36
Rolling Nealmon	-0.00	0.00	0.00	-34.96	38.74	-0.66
Rolling Legendre	-0.00	0.00	0.00	-15.90	32.29	-0.63

Cuadro D.2: Medidas de precisión de las predicciones en la muestra de test. MIDASr preCOVID.

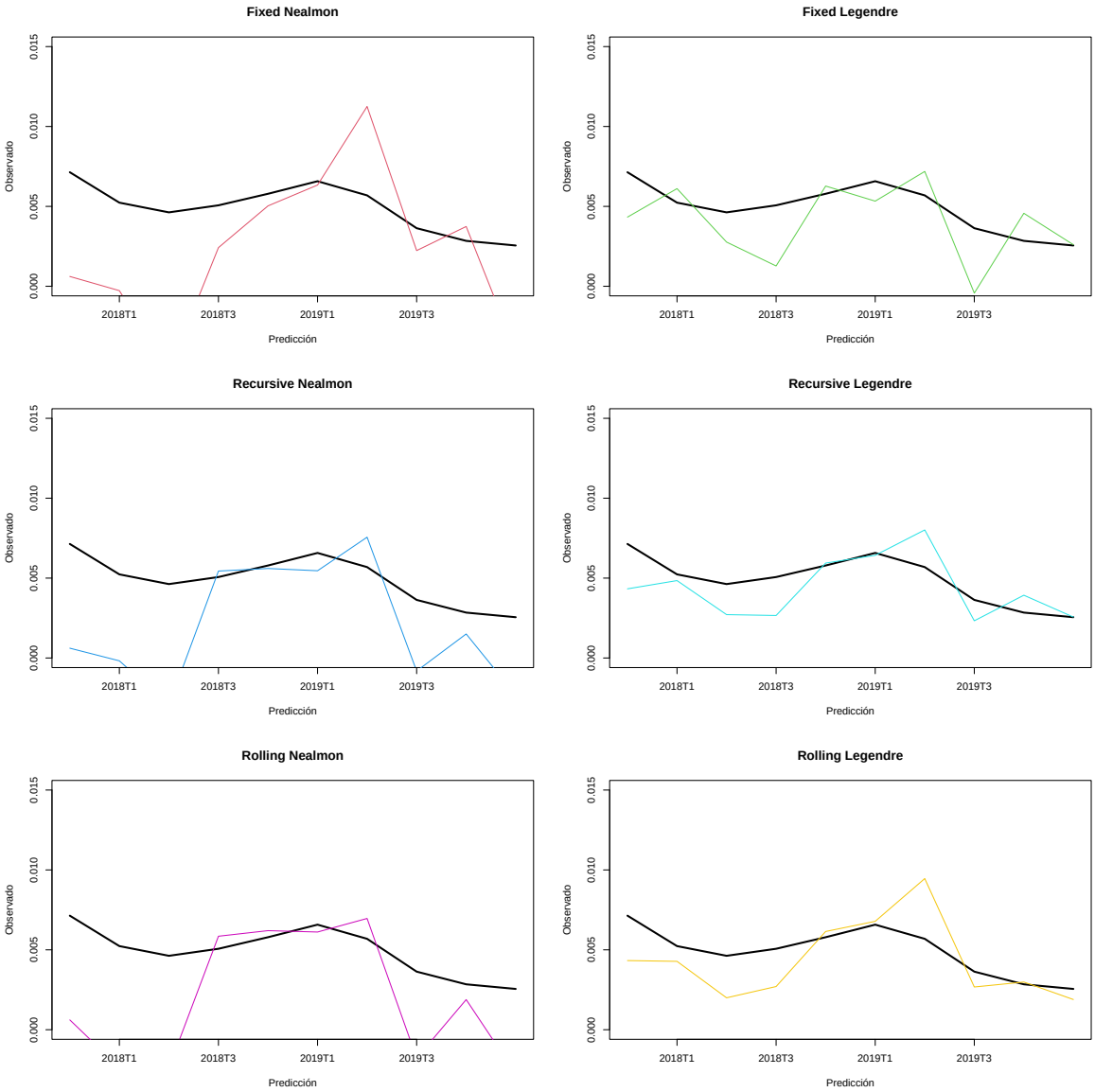

Figura D.1: Resultados MIDASr.

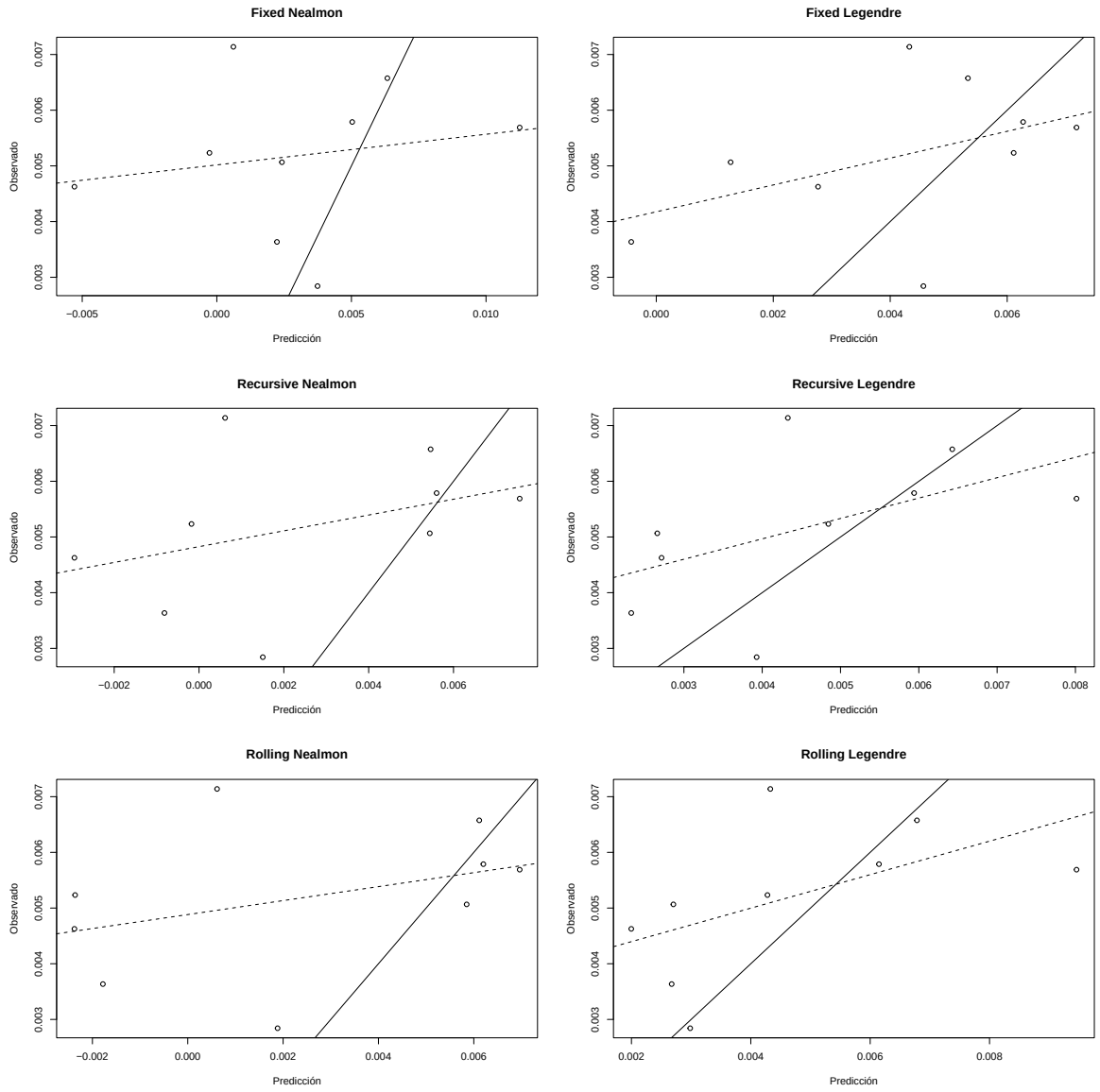

Figura D.2: Gráfico de dispersión de observaciones frente a predicciones (incluyendo la identidad, línea continua, y el ajuste lineal, línea discontinua). MIDASr.

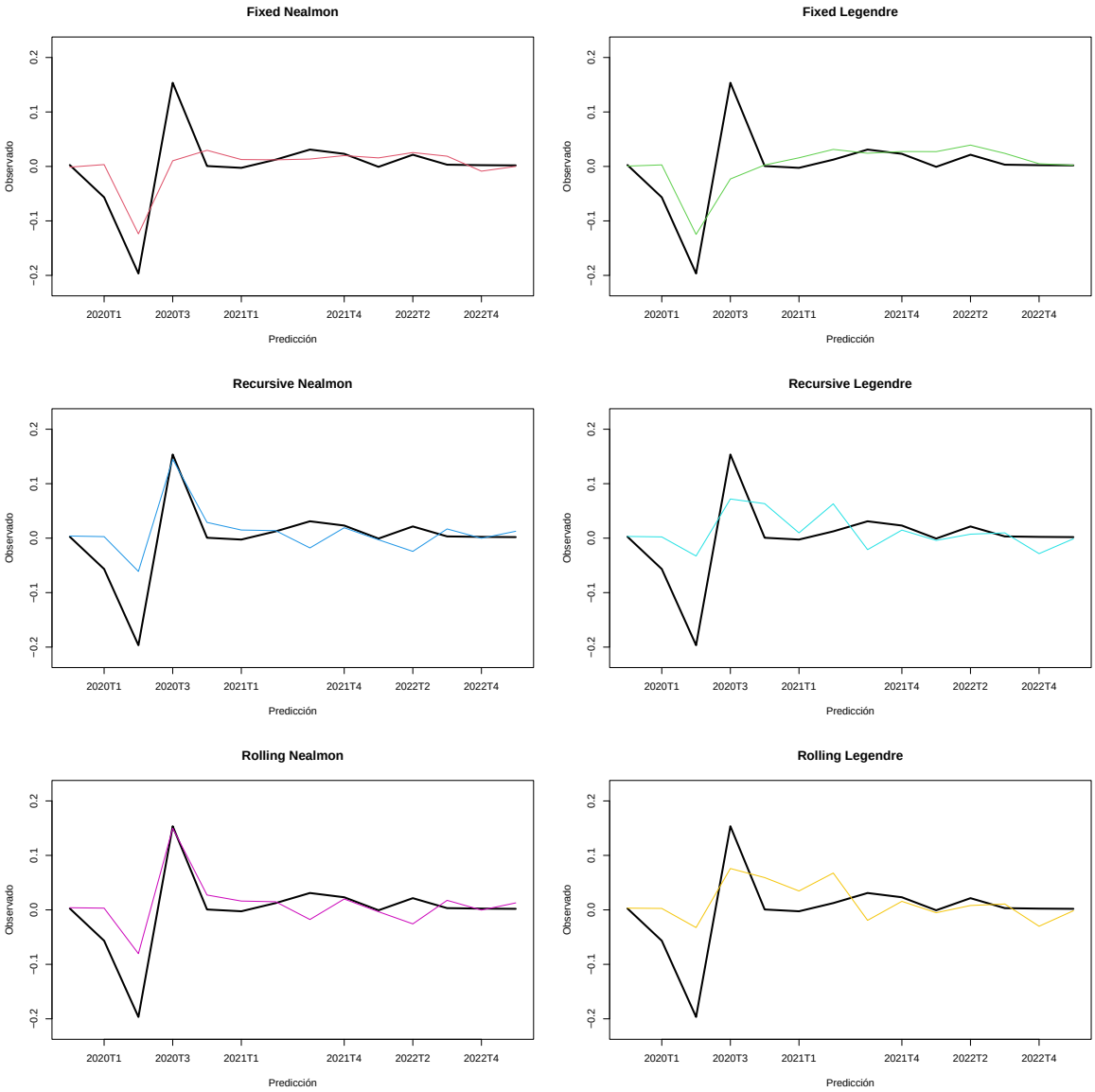

Figura D.3: Resultados MIDASr preCOVID.

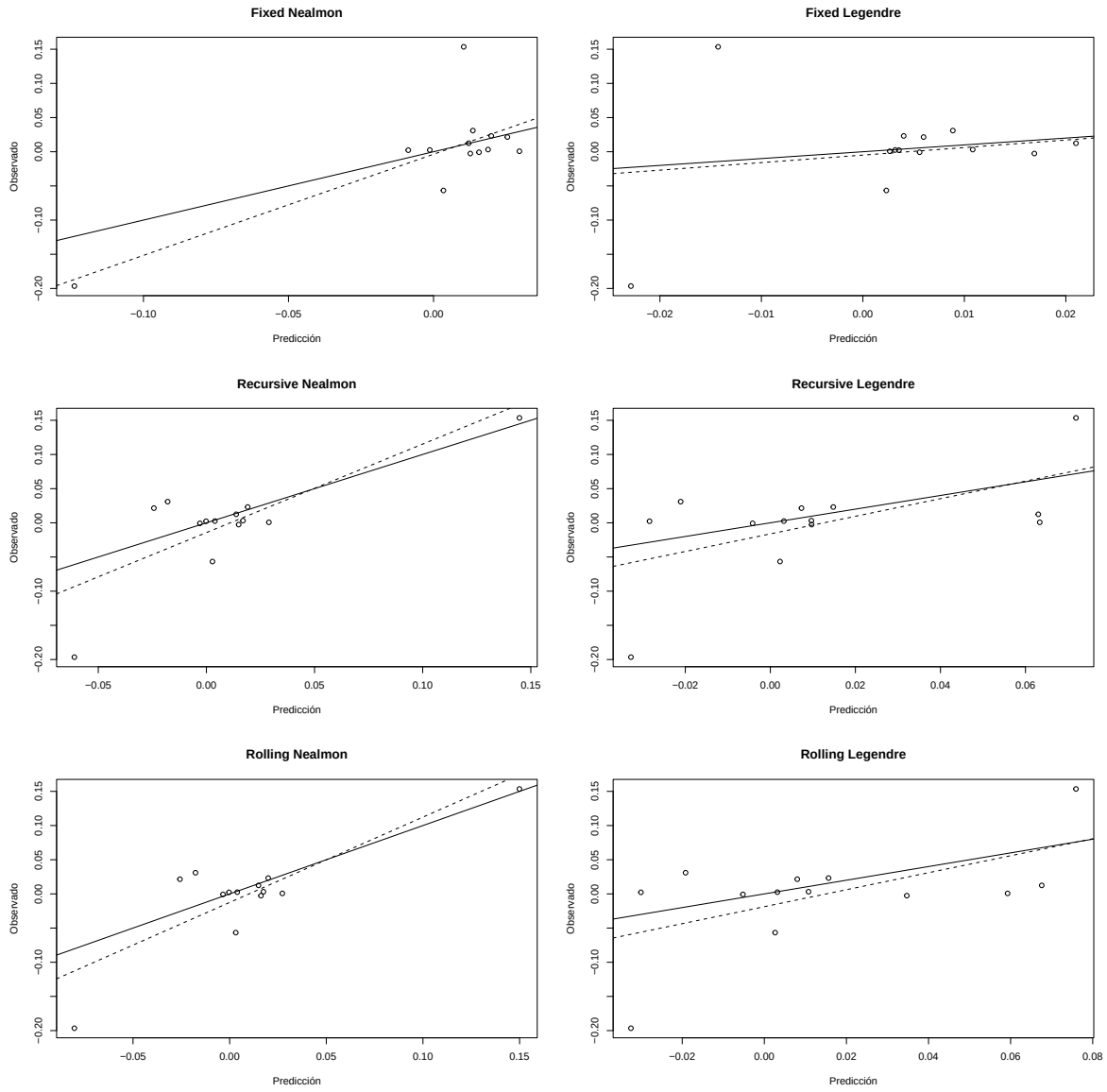

Figura D.4: Gráfico de dispersión de observaciones frente a predicciones (incluyendo la identidad, línea continua, y el ajuste lineal, línea discontinua). MIDASr preCOVID.

Bibliografía

Almon, S. (1965). The distributed lag between capital appropriations and expenditures. *Econometrica: Journal of the Econometric Society*, pages 178–196.

Aneiros, G. (2016). Series de tiempo. apuntes de la asignatura.

Armesto, M. T., Engemann, K. M., Owyang, M. T., et al. (2010). Forecasting with mixed frequencies. *Federal Reserve Bank of St. Louis Review*, 92(6):521–36.

Asimakopoulos, S., Paredes, J., and Warmedinger, T. (2013). Forecasting fiscal time series using mixed frequency data.

Babii, A., Ghysels, E., and Striaukas, J. (2022). Machine learning time series regressions with an application to nowcasting. *Journal of Business & Economic Statistics*, 40(3):1094–1106.

Bańbura, M., Giannone, D., Modugno, M., and Reichlin, L. (2013). Now-casting and the real-time data flow. In *Handbook of economic forecasting*, volume 2, pages 195–237. Elsevier.

Breiman, L. (2001). Random forests. *Machine learning*, 45:5–32.

Burman, J. P. (1980). Seasonal adjustment by signal extraction. *Journal of the Royal Statistical Society: Series A (General)*, 143(3):321–337.

Bylander, T. (2002). Estimating generalization error on two-class datasets using out-of-bag estimates. *Machine learning*, 48(1-3):287–297.

Callen, T. (2008). What is gross domestic product. *Finance & Development*, 45(4):48–49.

Chen, C. and Liu, L.-M. (1993). Joint estimation of model parameters and outlier effects in time series. *Journal of the American Statistical Association*, 88(421):284–297.

Cryer, J. D. and Kellet, N. (1991). *Time series analysis*. Springer.

Dagum, E. B. and Bianconcini, S. (2016). *Seasonal adjustment methods and real time trend-cycle estimation*. Springer.

Eurostat (2015). Ess guidelines on the seasonal adjustment. 2023.

Farouki, R. T., Goodman, T. N., and Sauer, T. (2003). Construction of orthogonal bases for polynomials in bernstein form on triangular and simplex domains. *Computer Aided Geometric Design*, 20(4):209–230.

Fernández, R., Costa, J., and Oviedo, M. (2021). Aprendizaje estadístico. apuntes de la asignatura.

Fernández Cerezo, A. (2023). Un procedimiento para la predicción a corto plazo del pib por el lado de la oferta. *Boletín económico/Banco de España [Artículos], 2023/T1, 18*.

Findley, D. F., Monsell, B. C., Bell, W. R., Otto, M. C., and Chen, B.-C. (1998). New capabilities and methods of the x-12-arima seasonal-adjustment program. *Journal of Business & Economic Statistics*, 16(2):127–152.

Flannery, R. (2020). *A Machine Learning Approach to Predicting Gross Domestic Product*. PhD thesis, Dublin, National College of Ireland.

Fuller, W. A. (2009). *Introduction to statistical time series*. John Wiley & Sons.

Ghysels, E., Grigoris, F., and Özkan, N. (2022). Real-time forecasts of state and local government budgets with an application to the covid-19 pandemic. *National Tax Journal*, 75(4):731–763.

Ghysels, E., Kvedaras, V., and Zemlys, V. (2016). Mixed frequency data sampling regression models: the r package midasr. *Journal of statistical software*, 72:1–35.

Ghysels, E., Santa-Clara, P., and Valkanov, R. (2004). The midas touch: Mixed data sampling regression models.

Ghysels, E., Santa-Clara, P., and Valkanov, R. (2006). Predicting volatility: getting the most out of return data sampled at different frequencies. *Journal of Econometrics*, 131(1-2):59–95.

Ghysels, E., Sinko, A., and Valkanov, R. (2007). Midas regressions: Further results and new directions. *Econometric reviews*, 26(1):53–90.

Gómez, V. and Maravall, A. (1994). Estimation, prediction, and interpolation for nonstationary series with the kalman filter. *Journal of the American Statistical Association*, 89(426):611–624.

Gómez, V. and Maravall Herrero, A. (1996). *Programs TRAMO and SEATS: instructions for the user (beta version: September 1996)*. Banco de España. Servicio de Estudios.

Greene, W. H. (2003). *Econometric analysis*. Pearson Education India.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182.

Hannan, E. J. and Rissanen, J. (1982). Recursive estimation of mixed autoregressive-moving average order. *Biometrika*, 69(1):81–94.

Haykin, S. S. (2002). *Adaptive filter theory*. Pearson Education India.

Hopp, D. (2022). Benchmarking econometric and machine learning methodologies in nowcasting. *arXiv preprint arXiv:2205.03318*.

INE (2019). Estándar del ine para la corrección de efectos estacionales y efectos de calendario en las series coyunturales. 2023.

Jahn, M. (2018). Artificial neural network regression models: Predicting gdp growth. Technical report, HWWI Research Paper.

Kuhn, M. (2022). *caret: Classification and Regression Training*. R package version 6.0-93.

Kuzin, V., Marcellino, M., and Schumacher, C. (2011). Midas vs. mixed-frequency var: Nowcasting gdp in the euro area. *International Journal of Forecasting*, 27(2):529–542.

Liaw, A., Wiener, M., et al. (2002). Classification and regression by randomforest. *R news*, 2(3):18–22.

Longo, L., Riccaboni, M., and Rungi, A. (2022). A neural network ensemble approach for gdp forecasting. *Journal of Economic Dynamics and Control*, 134:104278.

Maravall, A. and Planas, C. (1998). Signal extraction in non-invertible models. *Documento del Grupo de Trabajo en Ajuste Estacional, Eurostat*.

Marsilli, C. (2014). Variable selection in predictive midas models.

Mazzi, G. L., Ladiray, D., and Rieser, D. A. (2018). Handbook on seasonal adjustment.

Ripley, B. (2022). *nnet: Feed-Forward Neural Networks and Multinomial Log-Linear Models*. R package version 7.3-17.

Sax, C. (2022). seasonal: R interface to x-13-arima-seats. R package version 1.9.0.

Shiskin, J. (1967). *The X-11 variant of the census method II seasonal adjustment program*. Number 15. US Department of Commerce, Bureau of the Census.

Striaukas, J., Babii, A., and Eric Ghysels (2022). *midasml: Estimation and Prediction Methods for High-Dimensional Mixed Frequency Time Series Data*. R package version 0.1.10.

Tiao, G. C. and Tsay, R. S. (1983). Consistency properties of least squares estimates of autoregressive parameters in arma models. *The Annals of Statistics*, pages 856–871.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Timmermann, A. (2006). Forecast combinations. *Handbook of economic forecasting*, 1:135–196.

Torres, D. T. et al. (2015). Eurozona— evaluando la capacidad predictiva del midas. Technical report.

Trapletti, A. and Hornik, K. (2022). *tseries: Time Series Analysis and Computational Finance*. R package version 0.10-52.

Tsay, R. S. (1984). Regression models with time series errors. *Journal of the American Statistical Association*, 79(385):118–124.

Tsay, R. S. (1986). Time series model specification in the presence of outliers. *Journal of the American Statistical Association*, 81(393):132–141.

Werbos, P. (1974). New tools for prediction and analysis in the behavioral science. *Ph. D. dissertation, Harvard University*.

Wood, S. N. (2017). *Generalized additive models: an introduction with R*. CRC press.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.

Zhemkov, M. (2021). Nowcasting russian gdp using forecast combination approach. *International Economics*, 168:10–24.