Trabajo Fin de Máster

Construcción de un indicador adelantado de la tasa de paro

Laura Martínez García

Máster en Técnicas Estadísticas

Curso 2024-2025

Propuesta de Trabajo Fin de Máster

Título en galego: Construción dun indicador adiantado da taxa de paro
Título en español: Construcción de un indicador adelantado de la tasa de paro
English title: Construction of a Leading Indicator of the Unemployment Rate
Modalidad: Modalidad B
Autora: Laura Martínez García, Universidad de Santiago de Compostela
Directores: José Antonio Vilar Fernández, Universidad de La Coruña
Tutores: Teresa Veiga Rodríguez, ABANCA; Sergio Díaz Canosa, ABANCA
Breve resumen del trabajo: Este Trabajo Fin de Máster se realizó en colaboración con el departamento de Planificación Estratégica de ABANCA y aborda la construcción de un indicador adelantado de la tasa de paro mediante el análisis exhaustivo de las series del empleo. El estudio se centra en la desagregación temporal de estas series a partir de series relacionadas de mayor frecuencia, la corrección de los efectos de estacionalidad y calendario, con el objetivo de prever la tasa de desempleo. Los resultados obtenidos buscan proporcionar información valiosa para la toma de decisiones estratégicas en ABANCA, al anticipar datos de la evolución del empleo antes de su publicación, permitiendo una mejor comprensión del entorno económico.

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

Construcción de un indicador adelantado de la tasa de paro

fue realizado bajo su dirección por doña Laura Martínez García para el Máster en Técnicas Estadísticas. Estimando que el trabajo está terminado, dan su conformidad para su presentación y defensa ante un tribunal. Además, Don José Antonio Vilar Fernández y doña Laura Martínez García

√ sí □ no

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

En A Coruña, a 2 de junio de 2025.

El director:

VILAR FERNANDEZ JOSE ANTONIO - 32639957J

Firmado digitalmente por VILAR FERNANDEZ JOSE ANTONIO - 32639957J
Fecha: 2025.06.02 16:56:52 +02'00'

Don José Antonio Vilar Fernández

La tutora:

Doña Teresa Veiga Rodríguez

El tutor:

Don Sergio Díaz Canosa

La autora:

Doña Laura Martínez 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), **la autora declara** que el Trabajo Fin de Máster presentado es un documento original en el que se han tenido en cuenta las siguientes consideraciones relativas al uso de material de apoyo desarrollado por otros/as autores/as:

- Todas las fuentes usadas para la elaboración de este trabajo han sido citadas convenientemente (libros, artículos, apuntes de profesorado, páginas web, programas,...).
- Cualquier contenido copiado o traducido textualmente se ha puesto entre comillas, citando su procedencia.

- Se ha hecho constar explícitamente cuando un capítulo, sección, demostración,... sea una adaptación casi literal de alguna fuente existente.

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

Agradecimientos

Antes que nada, quiero agradecer a ABANCA por brindarme la oportunidad de realizar mi Trabajo de Fin de Máster en la empresa. Ha sido una experiencia muy enriquecedora que me ha permitido poner en práctica lo aprendido y, además, conocer de cerca cómo es el día a día en el mundo laboral.

También quiero dar las gracias de manera especial a mis tutores, Teresa Veiga Rodríguez y Sergio Díaz Canosa, por el apoyo y la confianza que me han brindado durante estos meses. Igualmente, agradezco a Ana Blanco Bugueiro quien ha estado guiándome y apoyándome en cada paso del proceso.

En el ámbito académico, estoy muy agradecida a todo el profesorado del Máster en Técnicas Estadísticas por compartir sus conocimientos y motivarme a seguir aprendiendo. Quiero hacer una mención especial a mi director académico, José Antonio Vilar Fernández, y también a Cheyenne Amoroso Sanmiguel, a quienes agradezco su apoyo y ayuda.

Por último, no puedo dejar de agradecer a mi familia y amigos, especialmente a mis padres, mis abuelos, Vanesa y a Yannick, por su apoyo incondicional y por estar siempre a mi lado.

Índice general

Resumen

Resumen en español

La tasa de paro es una variable clave para analizar la evolución del mercado laboral en España, pero se publica trimestralmente con un desfase de mes y medio tras el fin del trimestre en cuestión. Por ello, desde el Departamento de Planificación Estratégica y PMO de ABANCA se plantea la construcción de un indicador adelantado de la tasa de paro, estimado mensualmente a partir de los datos de afiliación a la Seguridad Social y de paro registrado, disponibles a principios de cada mes.

Para ello, mediante modelos de desagregación temporal, se mensualizan los datos de ocupados y parados usando las series mensuales de las afiliaciones y del paro registrado. A partir de estas estimaciones, se calcula una tasa de paro mensual que se convierte luego a frecuencia trimestral, permitiendo así su comparación con el dato oficial de la Encuesta de Población Activa (EPA).

Además, se incorpora una metodología de corrección de estacionalidad y efectos de calendario, esencial para el análisis de series temporales, ya que las series oficiales no están corregidas del impacto de estos factores. Una vez cancelados estos efectos, es factible un análisis más detallado y comparativas que, sin este ajuste, serían inadecuadas.

Resumo en galego

A taxa de desemprego é unha variable clave para analizar a evolución do mercado laboral en España, pero publícase trimestralmente cun desfasamento de mes e medio tras o fin do trimestre en cuestión. Por iso, desde o Departamento de Planificación Estratéxica e PMO de ABANCA plantéxase a construción dun indicador adiantado da taxa de desemprego, estimado mensualmente a partir dos datos de afiliación á Seguridade Social e de paro rexistrado, dispoñibles a principios de cada mes.

Para iso, mediante modelos de desagregación temporal, mensualízanse os datos de ocupados e parados usando as series mensuais das afiliacións e do paro rexistrado. A partir destas estimacións, calcúlase unha taxa de desemprego mensual que se converte logo a frecuencia trimestral, permitindo así a súa comparación co dato oficial da Enquisa de Poboación Activa (EPA).

Ademais, incorpórase unha metodoloxía de corrección de estacionalidade e efectos de calendario, esencial para a análise de series temporais, xa que as series oficiais non están corrixidas do impacto destes factores. Unha vez cancelados estes efectos, é factible unha análise máis detallada e comparativas que, sen este axuste, serían inadecuadas.

English abstract

The unemployment rate is a key variable for analyzing the evolution of the labor market in Spain, but it is published quarterly with a lag of one and a half months after the end of the quarter in question. Therefore, the Strategic Planning and PMO Department of ABANCA proposes the construction of a leading indicator of the unemployment rate, estimated monthly from the data on Social Security affiliation and registered unemployment, available at the beginning of each month.

To this end, using temporal disaggregation models, the data on employed and unemployed people are monthlyzed using the monthly series of affiliations and registered unemployment. From these estimates, a monthly unemployment rate is calculated, which is then converted to quarterly frequency, allowing its comparison with the official data from the Labor Force Survey (LFS).

In addition, a methodology for correcting seasonality and calendar effects is incorporated, which is essential for the analysis of time series, since the official series are not corrected for the impact of these factors. Once these effects have been canceled, a more detailed analysis and comparisons are feasible, which, without this adjustment, would be inadequate.

Capítulo 1

Introducción

El objetivo principal de este trabajo es desarrollar un indicador adelantado de la tasa de paro. El presente capítulo se estructura como sigue. La Sección 1.1 expone en detalle el problema a abordar, mientras que la Sección 1.2 introduce los principales conceptos y fuentes de datos que sustentan el análisis. A continuación, la Sección 1.3 revisa estudios previos que emplean metodologías similares. La Sección 1.4 presenta la estrategia seguida y los objetivos específicos del estudio. Las herramientas de software utilizadas en el análisis de datos se detallan en la Sección 1.5. Finalmente, la Sección 1.6 ofrece una visión general de la organización del trabajo, proporcionando una hoja de ruta para el lector.

1.1. Descripción del problema

El análisis del entorno macroeconómico constituye una función estratégica para cualquier entidad financiera, dado el impacto directo que la evolución de la economía tiene sobre el negocio bancario. Variables como el crecimiento del PIB, la inflación, los tipos de interés o el empleo influyen de forma determinante en la demanda de productos financieros, en la rentabilidad de las operaciones y en la exposición a distintos tipos de riesgo.

En ABANCA, esta labor de seguimiento económico y elaboración de previsiones recae en la Dirección General de Planificación Estratégica y PMO (Project Management Office). Desde este departamento se realiza un análisis continuo de un amplio conjunto de indicadores macroeconómicos, con especial atención a la evolución del empleo, tanto en el conjunto de España como en las regiones de Galicia y Portugal, áreas prioritarias para la actividad de la entidad. Entre estos indicadores, la tasa de paro calculada a partir de la Encuesta de Población Activa, denotada como EPA de ahora en adelante, destaca por su rigor metodológico y su papel como referencia para organismos nacionales e internacionales.

Sin embargo, la utilidad de la EPA se ve limitada por su naturaleza trimestral y el desfase con que se publica. En un entorno económico globalizado y cambiante, donde la toma de decisiones requiere información actualizada con mayor frecuencia, resulta necesario desarrollar instrumentos que permitan anticipar tendencias con más rapidez. Es aquí donde surge la motivación principal de este trabajo: diseñar un indicador adelantado que permita estimar mensualmente la tasa de paro en España, ofreciendo así una visión más frecuente, detallada y oportuna de la evolución del mercado laboral.

En este contexto, ABANCA busca un indicador que, conciliando con la tasa de paro, ofrezca una frecuencia mayor y permita anticipar cambios de patrones en el mercado laboral. Idealmente, este

indicador debería proporcionar información no solo sobre la tasa de paro, sino también sobre las estadísticas de empleo y desempleo, y estar corregido por efectos de calendario y estacionalidad para enriquecer los análisis. Por lo tanto, el problema se centra en implementar herramientas para el análisis de los datos del mercado laboral español y sirve de ayuda para permitir una mejor comprensión del mismo y facilitar la toma de decisiones estratégicas del banco.

1.2. Contexto y variables empleadas

En esta sección se presentan los conceptos y fuentes de datos fundamentales que se mencionarán a lo largo del presente trabajo. En cuanto a los datos, destacan dos conjuntos de variables. El primero, procedente de la EPA, incluye las variables de activos, ocupados, parados, inactivos y la tasa de paro. El segundo conjunto está compuesto por la serie temporal de paro registrado, publicada por el Servicio Público de Empleo Estatal (SEPE), y la de afiliados a la Seguridad Social, publicada por el Ministerio de Inclusión, Seguridad Social y Migraciones. Estas series están disponibles en la página web del Ministerio de Economía, Comercio y Empresa (MINECO).

La EPA es un estudio continuo y trimestral, iniciado en 1964, que proporciona información crucial sobre el mercado laboral español.

Con una muestra representativa de aproximadamente 65.000 familias (alrededor de 200.000 individuos) por trimestre, la EPA recopila datos sobre la población ocupada, activa, parada e inactiva. El Instituto Nacional de Estadística, de ahora en adelante INE, realiza la recogida de información mediante entrevistas presenciales en la primera toma de contacto con cada familia y posteriormente a través de entrevistas telefónicas o presenciales. Los datos, referidos a la semana anterior a la entrevista, se publican aproximadamente mes y medio después de la finalización del trabajo de campo.

A continuación, se introducen conceptos clave, definidos conforme a como se pueden encontrar en EPA (2008):

Definición 1.1. *La **población activa** es el conjunto de personas de 16 años o más que, durante la semana de referencia, se encuentran disponibles para trabajar o ya están participando en la producción de bienes y servicios. Esta población se divide en dos grupos: ocupados y parados.*

Definición 1.2. *Los **ocupados** son las personas de 16 años o más que, durante la semana de referencia, han trabajado al menos una hora a cambio de una remuneración o beneficio, ya sea monetario o en especie. Se incluyen también aquellos que, teniendo un empleo, se encuentran ausentes temporalmente (por enfermedad, vacaciones, etc.). Se clasifican en asalariados (públicos o privados) y por cuenta propia (empleadores y trabajadores independientes), y según su jornada laboral en tiempo completo (más de 30 horas semanales) y tiempo parcial (menos de 35 horas semanales).*

Definición 1.3. *Los **parados** son las personas de 16 años o más que, durante la semana de referencia, no tienen empleo, están disponibles para trabajar y buscan activamente empleo. La búsqueda activa se define por la realización de al menos una de las siguientes acciones: contacto con oficinas de empleo (públicas o privadas), envío de solicitudes a empleadores, consultas a través de redes personales o sindicatos, respuesta a anuncios, revisión de ofertas de empleo, asistencia a entrevistas o pruebas de selección, búsqueda de locales o terrenos, o gestión de permisos y recursos financieros. También se incluyen quienes tienen un empleo pero están a la espera de incorporarse, siempre que cumplan las condiciones de no tener empleo en la semana de referencia y estar disponibles para trabajar.*

Definición 1.4. *Los **inactivos** son las personas de 16 años o más que no se clasifican como activas, es decir, no están ocupadas ni paradas.*

Definición 1.5. *La **tasa de paro** es el indicador que mide el porcentaje de la población activa que*

se encuentra en situación de paro. Se calcula a través de la siguiente fórmula:

$$Tasa\ de\ paro = \left(\frac{Número\ de\ parados}{Número\ de\ activos} \right) \times 100 \tag{1.1}$$

A modo de resumen, en la Figura 1.1 se muestra como se relacionan los conceptos anteriormente introducidos de manera visual.

Figura 1.1: Esquema de la relación entre los conceptos de la EPA. Fuente: Vázquez (2018).

Para complementar el análisis basado en la EPA, se incorporan dos indicadores mensuales adicionales que ofrecen información relevante sobre la dinámica del mercado laboral: el paro registrado, que refleja el número de personas inscritas como demandantes de empleo, y el número de afiliaciones a la Seguridad Social, contabilizando las personas que cotizan al sistema. Estos datos mensuales se publican el segundo día del mes siguiente. A continuación, se definen dichos conceptos, que se pueden encontrar en Social (2025) y SEPE (2022).

Definición 1.6. *El **paro registrado** es el número total de solicitudes de empleo en alta, registradas por el Servicio Público de Empleo Estatal (SEPE) al final de cada mes, excluyendo las situaciones laborales especificadas en la Orden Ministerial de 11 de marzo de 1985 (B.O.E. de 14/03/85), que establece los criterios estadísticos para la medición del Paro Registrado.*

Definición 1.7. *El número de **afiliados a la Seguridad Social** es el total de personas incorporadas al Sistema anteriormente mencionado, cuya inclusión ha sido validada por la Tesorería General de la Seguridad Social. Representa el número de individuos que han iniciado una actividad laboral que les permite formar parte de dicho sistema.*

Como paso previo al análisis y tratamiento de los datos de este estudio, se llevó a cabo una labor de revisión de las metodologías que otros autores o instituciones han utilizado con el mismo fin. En la siguiente sección, se incluyen algunos ejemplos relevantes.

1.3. Algunos antecedentes

En el estudio de series temporales económicas, la corrección de estacionalidad y efectos de calendario constituye un paso fundamental, ya que permite identificar patrones estructurales y facilita la comparación entre distintos periodos. A lo largo del tiempo, diversas instituciones y expertos han

propuesto metodologías específicas para llevar a cabo este proceso, considerando las características propias de cada conjunto de datos.

El Instituto Nacional de Estadística de Chile (INE (2023)) utiliza la metodología X13-ARIMA-SEATS para ajustar estacionalmente las series de empleo, desagregadas por sexo y tramo etario. Este enfoque, que se basa en promedios móviles y desestacionalización indirecta, busca eliminar las fluctuaciones estacionales para obtener una visión más clara de la dinámica del mercado laboral chileno.

Para modelar el desempleo en países europeos, Eurostat (2024) utiliza el método de Denton proporcional para desagregar datos trimestrales de la EPA en series mensuales. Este método minimiza las diferencias entre puntos de datos consecutivos, sujeto a la condición de que el promedio trimestral coincida con los valores de la EPA. Posteriormente, se aplican modelos $ARIMA$ estacionales para pronosticar los factores de ajuste y obtener series de desempleo ajustadas estacionalmente.

En el contexto de la afiliación a la Seguridad Social en España, el Ministerio de Inclusión, Seguridad Social y Migraciones (Ministerio de Inclusión (2025)) emplea TRAMO-SEATS para ajustar las series mensuales y, de forma innovadora, TBATS para las series diarias. TBATS, un modelo estructural de componentes no observables, permite capturar patrones estacionales complejos, incluyendo ciclos semanales, intra-mensuales y anuales, mejorando así el análisis de datos de alta frecuencia.

1.4. Metodología y objetivos

Una vez establecidos el contexto y los conceptos clave (Sección 1.2) y enumerado metodologías específicas de interés empleadas por algunas instituciones (Sección 1.3), se procede ahora a describir el camino que se seguirá en este trabajo para abordar el problema planteado. Además, se establecen de forma precisa los objetivos que se pretenden alcanzar.

Como ya se ha indicado, la finalidad de este trabajo es construir un indicador adelantado y de frecuencia mensual de la tasa de paro en España. Para dar solución al problema planteado, la propuesta metodológica se centra en generar un indicador mensual mediante técnicas de desagregación temporal. En lugar de desagregar directamente la tasa de paro trimestral, se desagregan sus componentes: el número de parados y el número de ocupados. Esta aproximación, además de proporcionar una estimación mensual de la tasa de paro, enriquece el indicador al ofrecer información detallada sobre la evolución mensual del empleo y el desempleo por separado. Para ello, es conveniente contar con series de más alta frecuencia que sirvan de apoyo a la hora de realizar la desagregación. En este caso, se utilizarán el paro registrado y las afiliaciones a la Seguridad Social, disponibles en formato mensual.

Tras realizar la desagregación de las series trimestrales a series mensuales, se procede a la corrección de los efectos de estacionalidad y calendario, con el objetivo de depurar las series y evitar que estos componentes distorsionen la interpretación de la evolución del entorno económico.

Con el marco metodológico descrito, se desean alcanzar los objetivos que se enumeran a continuación.

- Implementar un procedimiento para la desagregación temporal de series trimestrales de empleo utilizando información auxiliar mensual.
- Corregir estacionalmente y de efectos de calendario las series temporales para disponer de series más limpias.
- Estimar una tasa de paro mensual coherente con los datos trimestrales, que actúe como indicador adelantado del mercado laboral.

- Contribuir a la mejora de los sistemas de seguimiento del empleo con métodos accesibles y reproducibles en otros casos.

- Emplear software libre (R) para la implementación de las técnicas que se requieran.

De esta manera, se busca no solo enriquecer el conocimiento sobre la evolución del empleo en España, sino también ofrecer una metodología de gran utilidad para los analistas de ABANCA para la desagregación temporal y desestacionalización de las series temporales.

Con el fin de tener una visión general de la solución propuesta en este trabajo, se presenta en la Figura 1.2 un esquema de la misma.

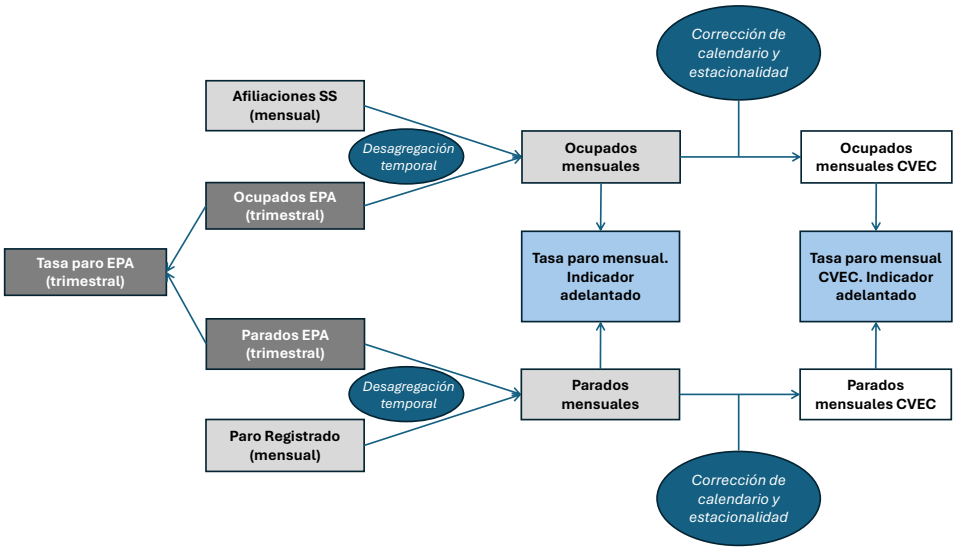

Figura 1.2: Esquema de la solución propuesta para dar solución al problema planteado por ABANCA.

1.5. Aspectos computacionales

Como se estableció entre los objetivos, el software que se va a emplear en este trabajo es R. Ahora bien, para desarrollar las metodologías antes expuestas, se requerirá de librerías específicas de desagregación y corrección de series temporales.

JDemetra+ es una herramienta de ajuste estacional y de análisis de series temporales desarrollada por el Banco Nacional de Bélgica en colaboración con el Detsche Bundesbank, Insee y Eurostat, de acuerdo con las directrices del Sistema Estadístico Europeo (Eurostat (2024)). Desde 2015, JDemetra+ ha sido oficialmente recomendado a los miembros del ESS (European Statistical System) y del Sistema Europeo de Bancos Centrales como el software para el ajuste estacional y de calendario de las estadísticas oficiales. Además del ajuste estacional, esta herramienta incluye otros modelos de series temporales útiles para el análisis de estadísticas económicas, como es la desagregación temporal.

Este programa se integra en el software R a través de la organización `rjdverse` que crea paquetes que dan acceso a los algoritmos de JDemetra+. Los paquetes que se han empleado son:

- `rjd3toolkit` (Palate et al. (2025a)), que es un paquete capa base para todos los demás paquetes, pruebas de estacionalidad, generación de variables regresoras...
- `rjd3tramoseats` (Palate et al. (2025b)), que es un paquete que se usa para ajustar series temporales eliminando efectos estacionales y de calendario.
- `rjd3bench` (Palate (2025)), el cual contiene funciones necesarias para la desagregación temporal.

La adopción de estos paquetes ofrece a ABANCA una valiosa capacidad para llevar a cabo desagregaciones temporales y correcciones de estacionalidad y efectos de calendario sobre una amplia gama de series temporales. Esta funcionalidad resulta especialmente relevante dado que la mayoría de las fuentes de datos no proporcionan series corregidas, lo que limita la capacidad de análisis. En efecto, por ejemplo, la tasa de paro no se publica corregida de estos efectos. Al implementar estas herramientas, ABANCA podrá obtener datos más limpios y enriquecidos para optimizar sus procesos.

1.6. Estructura

El trabajo se estructura en dos partes claramente diferenciadas. En la primera parte, se desarrolla el marco teórico y se expone la metodología empleada para abordar la desagregación temporal de las series de empleo, así como la corrección de los efectos de calendario y estacionalidad. La segunda parte adopta un enfoque práctico, en el que se describen las variables utilizadas y se presentan los resultados obtenidos para calcular la tasa temprana de paro. Finalmente, se recogen las conclusiones del estudio, destacando las principales aportaciones del trabajo y posibles líneas de investigación futura.

La primera parte consta de tres capítulos. En el Capítulo 2 se revisa la metodología Box-Jenkins, que incluye un amplio abanico de procesos bajo los cuales gran parte de las series temporales pueden ser modelizadas. A continuación, en el Capítulo 3 se explican los diferentes métodos de desagregación temporal empleados en el trabajo, y en el Capítulo 4 se aborda la metodología para corregir las series de efectos de calendario y estacionalidad, detallando el funcionamiento de la herramienta TRAMO-SEATS para llevar a cabo dicha corrección.

La segunda parte se estructura en cinco capítulos. En el Capítulo 5 se describen las variables empleadas, detallando su fuente de obtención y el método de automatización utilizado para la descarga de los datos. Además, se realiza un análisis exploratorio de las variables. En los Capítulos 6 y 7 se ponen en práctica las técnicas desarrolladas en la primera parte, explicando en detalle la aplicación de cada una de ellas. En el Capítulo 8 se construye el indicador adelantado de la tasa de paro y se analiza su comportamiento. Finalmente, en el Capítulo 9 se enumeran las conclusiones del trabajo así como las líneas futuras del mismo.

Parte I

Marco teórico y metodología

Capítulo 2

Tratamiento univariante de las series temporales

Este capítulo está dedicado a la presentación de las series de tiempo y a la metodología Box-Jenkins. Para la elaboración del mismo se emplearon como referencias principales Aneiros (2022), Chan and Cryer (2008), Box et al. (2015) y Peña (2005).

El capítulo se organiza como sigue. En la Sección 2.1 se introducen los conceptos elementales en el análisis de series temporales. En la Sección 2.2 se presentan los modelos clásicos tanto para procesos estacionarios como no estacionarios, que incluyen los modelos $AR(p), MA(q)$, $ARMA(p,q)$ y $ARIMA(p,d,q)$. La Sección 2.3 se amplían los modelos de la sección anterior incluyendo variables regresoras. Finalmente, la Sección 2.4 está dedicada a revisar procedimientos para las etapas de identificación, estimación y diagnosis del modelo, así como para la predicción de valores futuros en base al modelo seleccionado. Etapas todas ellas que conforman los ejes primordiales del análisis de una serie de tiempo.

2.1. Conceptos previos

Una serie de tiempo está formada por realizaciones aleatorias de una variable de interés a lo largo del tiempo. Por consiguiente, el primer paso es introducir el concepto de proceso estocástico, que proporciona el marco matemático apropiado para describir con rigor el comportamiento de variables aleatorias a lo largo del tiempo (y de ahí, de una serie de tiempo).

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

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

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

De esta forma, para un proceso estocástico como el definido en (2.1), $y(\cdot,t)$ es una variable aleatoria en el espacio muestral S, y para cada $s \in S$, $y(s,\cdot)$ corresponde a una realización del proceso estocástico a lo largo del intervalo de tiempo $\mathcal{I}$.

Una serie temporal puede entenderse como una realización de un proceso estocástico, denotada por $\{y_t\}_{t=1}^{T} = \{y_1, y_2, \ldots, y_T\}$, donde $t = 1, 2, \ldots, T$ representa los diferentes instantes de tiempo,

con T siendo el número total de observaciones. En otras palabras, se trata de una secuencia de datos organizados de manera cronológica y con intervalos de tiempo constantes entre ellos.

Dada una serie de tiempo observada, resulta de gran interés para su análisis obtener conocimiento del proceso estocástico que la ha generado. A continuación, se definen algunas funciones asociadas a un proceso estocástico, que se pueden consultar en el Capítulo 2 de Aneiros (2022).

Definición 2.2. *Sea $\{Y_t\}_{t\in\mathbb{Z}}$ un proceso estocástico.*

1. *La función de medias se define como*
$$\mu_t = \mathbb{E}(Y_t),$$
la cual es una medida de posición de carácter central de Y_t.

2. *La función de varianzas es*
$$\sigma_t^2 = Var(Y_t) = \mathbb{E}[(Y_t - \mu_t)^2],$$
que mide el grado de dispersión en torno a la función de medias de la variable Y_t.

3. *La función de autocovarianzas se define como*
$$\gamma(s,t) = Cov(Y_s, Y_t) = \mathbb{E}[(Y_s - \mu_s)(Y_t - \mu_t)] = \mathbb{E}(Y_s Y_t) - \mu_s\mu_t.$$
Esta función proporciona información sobre el grado de dependencia lineal entre las variables Y_s e Y_t, con $s,t \in \mathbb{Z}$. Las autocovarianzas tienen dimensión (la de las observaciones de la serie al cuadrado), por lo que no son adecuadas para comparar series registradas en unidades diferentes.

4. *La función de autocorrelaciones simples (ACF) evalúa el grado de dependencia lineal existente entre Y_s e e Y_t. Toma valores en $[-1,1]$ y se define formalmente como sigue:*
$$\rho(s,t) = \frac{Cov(Y_s, Y_t)}{\sqrt{Var(Y_s)}\sqrt{Var(Y_t)}} = \frac{\sigma(s,t)}{\sigma_s\sigma_t}$$
A continuación se define también una medida de la dependencia lineal, pero adimensional.

5. *La función de autocorrelaciones parciales (PACF) viene dada por la siguiente fórmula*
$$\alpha(s,t) = \frac{Cov(Y_s - \hat{Y}_s^{(s,t)}, Y_t - \hat{Y}_t^{(s,t)})}{\sqrt{Var(Y_s - \hat{Y}_s^{(s,t)}),\ Var(Y_t - \hat{Y}_t^{(s,t)})}},$$
donde $\hat{Y}_j^{(s,t)}$ denota al mejor predictor lineal de Y_j construido a partir de las variables medidas en los instantes comprendidos entre s y t (s y t excluidos). Esta función, también definida en $[-1,1]$, mide la dependencia lineal entre Y_s e Y_t, una vez que se eliminó el efecto lineal que ejercen las variables registradas entre los instantes s y t.

Todas estas funciones dependen del proceso estocástico subyacente, que es desconocido. Por tanto, deben de ser estimadas y, para ello, se asumen ciertas hipótesis sobre el proceso. Una hipótesis habitual será la estacionariedad del proceso, cuya idea es que la ley de probabilidad que rige el comportamiento del proceso no cambia a lo largo del tiempo. Los siguientes conceptos se pueden consultar en Chan and Cryer (2008).

Definición 2.3. *Un proceso estocástico* $\{Y_t\}_t \in \mathbb{Z}$ *se considera estrictamente estacionario si la distribución conjunta de los valores* $\{Y_{t_1}, Y_{t_2}, \ldots, Y_{t_n}\}$ *es idéntica a la distribución conjunta de los valores desplazados* $\{Y_{t_1-k}, Y_{t_2-k}, \ldots, Y_{t_n-k}\}$*, para cualquier conjunto de instantes de tiempo* $t_1, t_2, \ldots, t_n$ *y para cualquier retardo* k*.*

Dado que esta condición resulta difícil de verificar de manera empírica, es común recurrir a una versión más relajada del concepto de estacionariedad.

Definición 2.4. *Un proceso estocástico* $\{Y_t\}_{t\in\mathbb{Z}}$ *se dice que es débilmente estacionario, o estacionario de segundo orden, si cumple las siguientes condiciones:*

- $\mathbb{E}(Y_t) = \mu < \infty$ *para todo* $t \in \mathbb{Z}$*. Es decir, la media del proceso es constante.*
- $\mathbb{E}[(Y_t - \mu)(Y_{t-k} - \mu)] = \gamma_k$ *para todo* $t \in \mathbb{Z}$ *y cualquier retardo* k*, lo que implica que la covarianza entre observaciones con un retardo* k *depende únicamente de dicho retardo y no de los tiempos específicos de las observaciones.*

De esta manera, un proceso es débilmente estacionario si sus dos primeros momentos son constantes e invariables en el tiempo. Específicamente, la media y la varianza no deben cambiar con el tiempo, y la covarianza entre dos observaciones solo depende del retardo temporal entre ellas.

A lo largo de esta memoria, el término *serie estacionaria* o *proceso estacionario* se utilizará para hacer referencia a la estacionariedad de segundo orden.

Es importante introducir el concepto de ruido blanco, un ejemplo fundamental de proceso estacionario.

Definición 2.5. *Un proceso de ruido blanco se define como una secuencia de variables aleatorias* $\{a_t\}$ *incorreladas, con media nula y varianza constante finita* σ_a^2*. Así, se cumplen las siguientes condiciones:*

- $\mathbb{E}[a_t] = 0$*, para todo* t*.*
- $Var(a_t) = \sigma_t^2 = \sigma_a^2$*, para todo* t*.*
- $Cov(a_s, a_t) = \sigma(s,t) = \mathbb{E}(a_s a_t) - \mu_s \mu_t = \begin{cases} \sigma_a^2 & si\ s = t, \\ 0 & si\ s \neq t. \end{cases}$

2.2. Modelos Box-Jenkins

Por definición, las series temporales presentan una estructura de dependencia temporal que es preciso modelizar. La metodología de Box-Jenkins (Box et al. (2015)) busca construir un modelo capaz de capturar la dinámica de la serie, lo que permitirá realizar predicciones sobre los valores futuros de la misma. Los pasos a seguir son los siguientes:

1. Identificación del modelo.
2. Estimación de los parámetros del modelo.
3. Validación del modelo.
4. Predicción de valores futuros basados en el modelo ajustado.

La presente sección se dedica a introducir los modelos $AR(p)$, $MA(q)$ y $ARMA(p,q)$ para describir procesos estacionarios con y sin componente estacional. También se presentan extensiones de estos modelos para procesos no estacionarios.

Más adelante, en la Sección 2.4, se abordan brevemente las etapas que conforman la metodología Box-Jenkins, a saber, técnicas de selección del modelo más adecuado, el problema de estimar los parámetros que determinan estos modelos (por mínimos cuadrados y por máxima verosimilitud), y procedimientos para examinar la validez del modelo seleccionado.

El contenido de ambas secciones se ha basado principalmente en Chan and Cryer (2008) y Shumway et al. (2000).

2.2.1. Modelos para series estacionarias

La primera etapa consiste en identificar un modelo estocástico adecuado. Para ello, es interesante conocer uno de los modelos que ha resultado ser muy práctico en situaciones reales: los modelos $ARMA$, cuya estructura consta de una parte autorregresiva y otra de medias móviles (ver en Chan and Cryer (2008)).

Definición 2.6. *Un modelo autorregresivo de orden p, denotado como $AR(p)$, se define por:*

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + a_t,$$

donde $\{Y_t\}$ es un proceso estocástico, c y ϕ_i, con $i \in \{1, \ldots, p\}$, tal que $\phi_p \neq 0$, son constantes, y $\{a_t\}$ es un proceso de ruido blanco.

El proceso Y_t debe ser estacionario, es decir, el polinomio característico $\phi(z) = 1 - \phi_1 z - \ldots - \phi_p z^p$ no se anula cuando $|z| = 1$.

El modelo es causal si el polinomio característico no tiene raíces de módulo menor o igual a 1, es decir, $\phi(z) \neq 0$ para $|z| \leq 1$.

Usando el operador de retardo B, donde $BY_t = Y_{t-1}$, el modelo $AR(p)$ se puede escribir de forma compacta como:

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

donde $\phi(B) = 1 - \phi_1 B - \ldots - \phi_p B^p$.

Definición 2.7. *Un modelo de medias móviles de orden q, denotado como $MA(q)$, se define por:*

$$Y_t = c + a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \ldots + \theta_q a_{t-q}, \tag{2.2}$$

donde $\{Y_t\}$ es un proceso estocástico, c y θ_i, con $i \in \{1, \ldots, q\}$, tal que $\theta_q \neq 0$, son constantes, y $\{a_t\}$ es un proceso de ruido blanco.

Por un lado, es importante tener en cuenta que el proceso de medias móviles de orden q siempre es estacionario y causal. Por otro lado, el proceso $MA(q)$ será invertible si y solo si se verifica que el polinomio característico $\theta_q(z) = 1 + \theta_1 z + \theta_2 z^2 + \ldots + \theta_q z^q \neq 0$, para todo z tal que $|z| \leq 1$. Como en el caso de los modelos autorregresivos, el modelo dado por la ecuación (2.2) se puede reescribir de manera más compacta empleando el operador retardo B, por lo que se obtiene

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

A continuación, presentamos los modelos $ARMA$, que combinan las estructuras autorregresiva y de medias móviles.

Definición 2.8. *Un modelo ARMA de órdenes p y q, denotado como ARMA(p, q), se define por:*

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \ldots + \theta_q a_{t-q},$$

donde $\{Y_t\}$ es un proceso estocástico, c y ϕ_i, con $i \in \{1, \cdots, p\}$ y θ_j con $j \in \{1, \cdots, q\}$ son constantes, tal que $\phi_p \neq 0$ y $\theta_q \neq 0$, y $\{a_t\}$ es un proceso de ruido blanco.

Para que esta representación corresponda a un proceso estacionario, el polinomio característico asociado a la parte autorregresiva, $\phi_p(z)$, no debe tener ninguna raíz con módulo igual a uno. Además, un proceso $ARMA(p, q)$ será causal si las raíces del polinomio $\phi_p(z)$ tienen módulo estrictamente mayor que uno, y será invertible si el polinomio $\theta_q(z)$ no tiene raíces con módulo mayor que uno.

Finalmente, de forma compacta, el modelo $ARMA(p, q)$ se puede escribir como:

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

Los modelos $ARMA(p, q)$ permiten caracterizar la dependencia regular, es decir, la relación entre observaciones consecutivas o entre el ruido blanco presente en el pasado reciente. Sin embargo, puede haber casos en el que una serie temporal presenta una dependencia estacional, es decir, una relación entre observaciones separadas por múltiplos de un período estacional s. Para abordar este problema, se introducen los procesos $ARMA$ estacionales.

Definición 2.9. *Un modelo ARMA estacional de órdenes P y Q y con período estacional s, denotado como $ARMA(P, Q)_s$, se define por:*

$$Y_t = c + \Phi_1 Y_{t-s} + \Phi_2 Y_{t-s} + \cdots + \Phi_p Y_{t-Ps} + a_t + \Theta_1 a_{t-s} + \Theta_2 a_{t-s} + \cdots + \Theta_q a_{t-Qs}, \tag{2.3}$$

donde $\{Y_t\}$ es un proceso estocástico, c y Φ_i, con $i \in \{1, \cdots, P\}$ y Θ_j con $j \in \{1, \cdots, Q\}$ son constantes, tal que $\Phi_P \neq 0$ y $\Theta_Q \neq 0$, y $\{a_t\}$ es un proceso de ruido blanco.

Los procesos $ARMA$ estacionales son un caso particular de los procesos $ARMA$ no estacionales de órdenes sP y sQ, con muchos coeficientes nulos. Las condiciones de estacionariedad, causalidad e invertibilidad para estos procesos coinciden con las establecidas para los ARMA no estacionales.

Nótese que es posible expresar la ecuación (2.3) esta definición de manera más compacta:

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

donde B^s es el operador de retardo estacional tal que $B^s Y_t = Y_{t-s}$ y

$$\Phi(B^s) = 1 - \Phi_1 B^s - \Phi_s B^{2s} - \cdots - \Phi_P B^{Ps},$$
$$\Theta(B^s) = 1 + \Theta_1 a_{t-s} + \Theta_2 a_{t-2s} + \cdots + \Theta_Q a_{t-Qs}.$$

En orden a tratar con la presencia simultánea de dependencia regular y estacional, pueden considerarse los modelos $ARMA$ estacionales multiplicativos.

Definición 2.10. *Un modelo ARMA estacional multiplicativo de órdenes p, q, P y Q y con período estacional s, denotado como $ARMA(p, q) \times (P, Q)_s$, se define por:*

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

donde $\{Y_t\}$ es un proceso estocástico, c es una constante, $\Phi(B^s)$ y $\phi(B)$ son los polinomios autorregresivos, $\Theta(B^s)$ y $\theta(B)$ son los polinomios de medias móviles y $\{a_t\}$ es un proceso de ruido blanco.

2.2.2. Modelos para series no estacionarias

Hasta aquí se han presentado modelos de series de tiempo estacionarias, pero la hipótesis de estacionariedad es poco común con series de datos reales, de modo que resulta importante extender los procesos $ARMA$ para poder modelizar series no estacionarias.

En particular, se consideran las siguientes posibles causas de no estacionariedad.

1. Tendencia: La media o el nivel de la serie no es constante.
2. Componente estacional: Existe un patrón repetitivo de la serie.
3. Heterocedasticidad: La varianza de la serie no es constante.

Para determinar si una serie presenta una tendencia significativa, se puede aplicar el test de Mann-Kendall, el cual se basa en la correlación de rangos de Kendall. Esta correlación mide la fuerza de la asociación monótona entre dos variables, a través del coeficiente de correlación de Kendall, denotado por τ.

Se evalúa la tendencia mediante el siguiente estadístico:

$$\tau = \frac{S}{D}, \tag{2.4}$$

donde:

- $S = \sum_{i<j} \text{sign}(x_j - x_i)$ es el **score** que resume las concordancias y discordancias entre pares de observaciones.
- $D = \frac{n(n-1)}{2}$ es el valor máximo posible de S.

El coeficiente τ toma valores en el intervalo $[-1, 1]$, donde:

- $\tau > 0$ indica una tendencia creciente.
- $\tau < 0$ indica una tendencia decreciente.
- $\tau = 0$ sugiere ausencia de tendencia monótona.

A partir de este coeficiente, el test de Mann-Kendall plantea el siguiente contraste de hipótesis:

- H_0: No existe tendencia en los datos.
- H_a: Existe una tendencia en los datos, ya sea creciente o decreciente.

La tendencia puede ser resuelta diferenciando la serie de forma regular, por lo que se introduce la definición que sigue.

Definición 2.11. *Un proceso estocástico $\{Y_t\}$ que no es estacionario y presenta una tendencia, se dice integrado de orden $d > 0$, denotado como $\{Y_t\} \sim I(d)$, si la serie transformada $(1-B)^d Y_t$ es estacionaria, mientras que $(1-B)^{d-1} Y_t$ no lo es.*

Sobre la base de esta definición nace un nuevo tipo de modelos, los $ARIMA$, en los cuales se aplican d diferencias regulares a la serie de tiempo tras ajustar un modelo $ARMA$.

Definición 2.12. *Un modelo ARIMA de órdenes p, d y q, denotado como $ARIMA(p,d,q)$, se define por:*

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

donde $\{Y_t\}$ es un proceso estocástico, c es una constante, $\phi(B)$ es el polinomio autorregresivo, $\theta(B)$ es el polinomio de medias móviles y $\{a_t\}$ es un proceso de ruido blanco.

Cabe mencionar que ante una serie temporal no estacionaria con tendencia, para eliminar la tendencia y corregir la estacionariedad basta con aplicar $d \leq 3$ diferencias regulares ($(1-B)^d$), lo cual se puede consultar en Aneiros (2022).

Otra componente que puede presentar una serie temporal es la estacional, para lo que será necesario diferenciar la serie de manera estacional.

Se entiende por estacionalidad a aquellos fenómenos que ocurren en el tiempo y se repiten cada período idéntico de tiempo, lo que se puede identificar a través de las variaciones intertrimestrales o interanuales de las series temporales. Sin embargo, este análisis puede ser engañoso si no se tienen en cuenta ciertos factores. Por ejemplo, las ventas de juguetes pueden variar dependiendo de la estación del año, lo que implica que un aumento de ventas en diciembre respecto a cualquier otro mes no necesariamente refleja un cambio estructural en la demanda. De igual forma, al comparar las variaciones interanuales de variables mensuales, se debe tener en cuenta si las fechas de la Semana Santa varían de un año a otro o si el número de días laborables es distinto entre meses.

Definición 2.13. *Un proceso estocástico $\{Y_t\}_t$ que no es estacionario y presenta dependencia estacional de período s, se considera integrado de orden $D > 0$, denotado como $\{Y_t\} \sim I(D)$, si la serie transformada $(1-B^s)^D Y_t$ es estacionaria, mientras que $(1-B^s)^{D-1} Y_t$ no lo es.*

De nuevo, por este concepto surgen los modelos $ARIMA$ estacionales, en los cuales se aplican D diferencias estacionales a la serie tras ajustar un modelo $ARMA$.

Definición 2.14. *Un modelo ARIMA estacional de órdenes P, D y Q y con período estacional s, denotado como $ARIMA(P,D,Q)_s$, se define por:*

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

donde $\{Y_t\}$ es un proceso estocástico, c es una constante, $\Phi(B^s)$ es el polinomio autorregresivo, $\Theta(B^s)$ es el polinomio de medias móviles y $\{a_t\}$ es un proceso de ruido blanco.

Los modelos $ARIMA$ estacionales multiplicativos permiten tratar con modelos no estacionarios presentando simultáneamente dependencia regular y estacional.

Definición 2.15. *Un modelo ARIMA estacional multiplicativo de órdenes p, d, q, P, D y Q, y con período estacional s, denotado como $ARIMA(p,d,q) \times (P,D,Q)_s$, se define por:*

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

donde $\{Y_t\}$ es un proceso estocástico, c es una constante, $\Phi(B^s)$ y $\phi(B)$ son los polinomios autorregresivos, $\Theta(B^s)$ y $\theta(B)$ son los polinomios de medias móviles y $\{a_t\}$ es un proceso de ruido blanco.

Como último caso, la existencia de heterocedasticidad puede corregirse aplicando la transformación Box-Cox sobre los datos de la serie de tiempo

$$Y_t = \begin{cases} \frac{Y_t^\lambda - 1}{\lambda} & \text{si } \lambda \neq 0, \\ \ln(Y_t) & \text{si } \lambda = 0 \end{cases} \tag{2.5}$$

Esta transformación nos permite estabilizar la varianza de la serie (para más información consultar Box and Cox (1964)).

2.3. Modelos de regresión en el análisis de series temporales

Siguiendo el Capítulo 3 de Aneiros (2022), en el análisis de series temporales, los modelos de regresión dinámica permiten modelar una serie temporal a partir de otra, asumiendo una dependencia entre ellas. Sea $\{Y_t\}_t$ la serie temporal que deseamos predecir, y $\{X_t\}_t$ la serie explicativa que influye sobre Y_t. El modelo de regresión lineal básico que captura esta relación es el siguiente:

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

donde r es el retardo, β_0 y β_1 son los parámetros constantes, y $\{\varepsilon_t\}_t$ es el proceso estocástico de error.

Para que este modelo sea válido, se deben cumplir ciertos requisitos:

1. Ambas series, $\{X_t\}_t$ y $\{Y_t\}_t$, deben ser conjuntamente estacionarias.
2. Es necesario identificar el valor adecuado del retardo r.
3. Es posible que los errores ε_t presenten autocorrelación.

2.3.1. Medición de dependencia entre series temporales

La relación lineal entre dos series temporales se evalúa mediante la **correlación cruzada**, definida como:

$$\rho_{s,t}(X,Y) = \frac{\mathrm{Cov}(X_t, Y_{t+\tau})}{\sigma_X \sigma_Y},$$

donde σ_X y σ_Y son las desviaciones estándar de las series X_t y Y_t, respectivamente.

La **correlación cruzada muestral** es el estimador usado para evaluar el grado de dependencia entre las series:

$$\hat{\rho}_k(X,Y) = \frac{\sum_t (X_t - \bar{X})(Y_{t+k} - \bar{Y})}{\sqrt{\sum_t (X_t - \bar{X})^2 \sum_t (Y_{t+k} - \bar{Y})^2}},$$

donde $\bar{X}$ y $\bar{Y}$ son las medias de las series X_t y Y_t.

Para considerar una correlación significativa, se utiliza el siguiente criterio:

$$|\hat{\rho}_k(X,Y)| \geq \frac{1.96}{\sqrt{T}},$$

donde T es el tamaño de la muestra.

El valor del retardo k que cumpla esta condición se considera como un candidato para el retardo r en el modelo de regresión.

En caso de que las series no sean estacionarias, es necesario aplicar un proceso denominado **preblanqueo**, que consiste en las siguientes etapas:

1. Ajustar un modelo lineal a la serie X_t y transformarla mediante un operador lineal $\pi(B)$, tal que el resultado sea un ruido blanco: $\tilde{X}_t = \pi(B)X_t$.
2. Aplicar la misma transformación a la serie Y_t: $\tilde{Y}_t = \pi(B)Y_t$.
3. Si $\tilde{Y}_t$ es estacionaria, se estudia la correlación cruzada entre $\tilde{X}_t$ y $\tilde{Y}_t$ para determinar el retardo k adecuado.

Nótese que si alguna de las series originales no es estacionaria, será necesario diferenciarlas antes de aplicar el preblanqueo. Como la transformación $\pi(B)$ es lineal, la correlación entre las series transformadas se mantiene en las originales.

2.4. Identificación, estimación y diagnosis del modelo

Tras haber introducido un abanico de modelos para caracterizar series temporales en las Secciones 2.2.1 y 2.2.2, es necesario identificar aquel que mejor describe la serie en estudio. En esta sección se exponen muy brevemente algunas pautas sobre cómo identificar el modelo generador de la serie y, entonces, cómo abordar su estimación y validación.

2.4.1. Identificación del modelo

El análisis preliminar de una serie temporal comienza con una exploración visual y estadística para identificar posibles fuentes de no estacionariedad. Entre las causas más habituales se encuentran la presencia de estacionalidad, tendencia y heterocedasticidad.

En primer lugar, es importante evaluar la posible presencia de estacionalidad, entendida como fluctuaciones que se repiten en intervalos regulares, por ejemplo, de forma trimestral o anual. Esta puede detectarse visualmente a través de los gráficos de la serie temporal, observando patrones que se repiten en los mismos períodos, o mediante el análisis del gráfico de autocorrelaciones simples (ACF), donde se manifiesta mediante picos significativos en retardos estacionales s y sus múltiplos ($2s$, $3s$, etc.). Además, existen pruebas estadísticas específicas que permiten detectar la estacionalidad de manera analítica, como el QS-Test (Maravall (2011)), el test de Friedman (Friedman (1937)) o el test de Kruskal-Wallis (Kruskal and Wallis (1952)), entre otros. Estos procedimientos serán aplicados en la Sección 5.2.1.

Una vez abordada la estacionalidad, se analiza la posible presencia de una tendencia, es decir, una evolución sistemática en el nivel medio de la serie a lo largo del tiempo. Esta puede observarse de forma preliminar en el gráfico de la serie temporal y también a través del comportamiento de los coeficientes de autocorrelación simple, que en presencia de una tendencia suelen tomar valores altos y decrecer lentamente hacia cero. Para eliminar la tendencia se aplican d diferencias regulares. Asimismo, la presencia de una tendencia puede comprobarse mediante el test de Mann-Kendall (McLeod (2022)), el cual evalúa si existe una tendencia monótona, creciente o decreciente, a lo largo del tiempo.

Otro aspecto a considerar es la presencia de heterocedasticidad, es decir, varianza no constante en el tiempo. Este fenómeno puede dificultar el modelado de la serie y debe corregirse antes de aplicar modelos paramétricos. En estos casos, se recurre a transformaciones como la transformación de Box-Cox (2.5), que permite estabilizar la varianza y mejorar las propiedades estadísticas de la serie.

Una vez realizadas las transformaciones necesarias para eliminar la estacionalidad, la tendencia y estabilizar la varianza, es fundamental verificar si la serie resultante es estacionaria. Para ello, se utilizan pruebas de raíces unitarias, como el test de Dickey-Fuller aumentado (Dickey and Fuller (1979)).

El siguiente paso es identificar los órdenes p, q, P y Q basándose en los patrones de las funciones de las autocorrelaciones simples y parciales muestrales, o bien a través de la minimización de algún criterio de información. En el desarrollo de este trabajo se ha adoptado esta última vía.

Como la selección del modelo a partir del estudio de las estimaciones de los coeficientes de las autocorrelaciones nos puede llevar a identificar más de un modelo, se tiene que disponer de métodos que conduzcan a una selección automática, y en este punto es donde se requiere la minimización de

algún criterio de información.

Existen distintas opciones como el criterio de información Bayesiano (BIC), el criterio de información de Akaike (AIC) y el criterio de información de Akaike corregido ($AICc$). En este trabajo se selecciona el modelo $ARMA(p,q)$ que minimiza

$$AIC_c(p,q) = -2\log(L(\hat{\beta})) + 2\frac{kT+k+2}{T-k-2},$$

donde L es la función de verosimilitud, $\hat{\beta}$ es un vector con las estimaciones de máxima verosimilitud de los parámetros del modelo (sin contar la varianza de a_t), k es el número de parámetros a excepción de la varianza del ruído blanco y T es el tamaño de la serie.

2.4.2. Estimación del modelo

Una vez que hemos seleccionado el proceso generador de la serie, el siguiente paso es estimar los parámetros del modelo.

En esta sección se presentan dos métodos para la estimación de los parámetros del modelo $ARMA$, mínimos cuadrados y máxima verosimilitud.

Como los parámetros implicados en un modelo $ARMA(p,q)$ son: $c, \phi_1, \cdots, \phi_p, \theta_1, \cdots, \theta_q$, y la varianza de las innovaciones, σ_a^2, de aquí en adelante se denotará al vector de parámetros estimados por:

$$\hat{\boldsymbol{\beta}} = (\hat{c}, \hat{\phi}_1, \cdots, \hat{\phi}_p, \hat{\theta}_1, \cdots, \hat{\theta}_q, \hat{\sigma}_a^2)^\intercal.$$

Método de mínimos cuadrados

La estimación de los parámetros mediante este método viene dado por la minimización de la suma de los residuos al cuadrado, esto es los estimadores de mínimos cuadrados verifican

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

donde $\hat{a}_t = Y_t - (\hat{c} + \hat{\phi}_1 Y_{t-1} + \cdots + \hat{\phi}_p Y_{t-p} + \hat{\theta}_1 \hat{a}_{t-1} \cdots + \hat{\theta}_q \hat{a}_{t-q}), t \in \{1, \cdots, T\}$, son los residuos.

Pueden surgir dos problemas:

- Si $p > 0$, los valores de los residuos $\hat{a}_1, \cdots, \hat{a}_p$ dependen de los valores de $Y_0, Y_{-1}, \cdots, Y_{1-p}$, que no son observados.

  Una solución es minimizar la suma de los residuos al cuadrado a partir del instante $p+1$, es decir

  $$\hat{\boldsymbol{\beta}} = \min_{\boldsymbol{\beta}} \sum_{t=p+1}^{T} \hat{a}_t^2. \tag{2.6}$$

- Si además $q > 0$, $\hat{a}_{p+1}$ dependerá de los valores $\hat{a}_p, \cdots, \hat{a}_{p+1-q}$, que a su vez son función de los valores no observados de Y_t. Si se conocen $\hat{a}_p, \cdots, \hat{a}_{p+1-q}$, podemos obtener iterativamente los valores $\hat{a}_{p+1}, \cdots, \hat{a}_T$.

  Como solución hemos empleado el método de mínimos cuadrados condicionados, donde se obtienen las estimaciones de los parámetros minimizando la función en β la función objetivo en (2.6) sujeta a que $\hat{a}_p = \cdots = \hat{a}_{p+1-q} = 0$.

Método de máxima verosimilitud

Los estimadores de máxima verosimilitud se obtienen maximizando la función de verosimilitud

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} f_{\boldsymbol{\beta}}(Y_1, \cdots, Y_T),$$

donde $f_{\boldsymbol{\beta}}$ es la función de densidad conjunta asociada al vector aleatorio $(Y_1, \cdots, Y_T)^T$ de un proceso $ARMA(p, q)$ con coeficientes β.

2.4.3. Validación del modelo

El último paso es validar el modelo, lo que consiste en comprobar si se satisfacen las hipótesis de que sus innovaciones a_t son ruido blanco, es decir,

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

A mayores, se comprueba que las innovaciones sean también gaussianas, ya que bajo normalidad, la incorrelación equivale a independencia.

Si alguna de las hipótesis falla, el modelo seleccionado no será valido y habrá que seleccionar otro modelo y repetir este proceso hasta encontrar un modelo que verifique todas las hipótesis. Como las innovaciones son desconocidas, las hipótesis se chequean sobre los residuos del modelo.

Contraste de independencia

El test de Ljung-Box (Ljung and Box (1978)) es una prueba potente bajo normalidad para chequear la hipótesis nula de que las primeras h autocorrelaciones son nulas. El estadístico de contraste es $Q(h) = T(T+2)\sum_{k=1}^{h} \frac{\hat{\rho}_k^2}{T-k}$, donde T es el tamaño de la muestra, $\hat{\rho}_k$ son las estimaciones de las autocorrelaciones simples en el retardo k de los residuos y h es el número de retardos. Bajo la hipótesis nula, este estadístico sigue una distribución χ^2 asintótica con h grados de libertad si las correlaciones se estiman de manera empírica a partir de la serie original. Si las correlaciones se estiman a partir de los residuos de un modelo ajustado (por ejemplo, un modelo ARMA), entonces el número de grados de libertad es $h - p$, donde p es el número de parámetros estimados en el modelo. Para más información sobre este contraste se puede consultar Peña (2005).

Contraste de media nula

Una vez verificado que los residuos son independientes, se realiza un contraste para comprobar si su media es nula, utilizando para ello el test t de Student (Student (1908)).

La hipótesis nula de este contraste establece que los residuos del modelo provienen de variables aleatorias con distribución idéntica, media cero y varianza finita. El estadístico correspondiente a este contraste, bajo el supuesto de que T es grande, es:

$$\frac{\bar{r}}{\hat{s}_r/\sqrt{T}},$$

donde $r = (r_1, r_2, \ldots, r_T)$ representa los residuos del modelo, siendo $\bar{r}$ la media muestral de los mismos y $\hat{s}_r$ la cuasidesviación típica muestral. Bajo la nula, este estadístico sigue una Este estadístico sigue una distribución normal estándar $N(0, 1)$.

Por lo tanto, la hipótesis nula será rechazada con un nivel de significación $\alpha = 0.05$, si se cumple la siguiente condición:

$$|\bar{r}| \geq z_{\alpha/2} \frac{\hat{s}_r}{\sqrt{T}}.$$

Contraste de normalidad

En el caso del contraste de normalidad de una muestra, se pueden utilizar pruebas como la de Jarque-Bera (Jarque and Bera (1987)) y la de Shapiro-Wilk (Shapiro and Wilk (1965)), entre otras. La primera se basa en los coeficientes de curtosis y asimetría, mientras que la segunda se apoya en el gráfico cuantil-cuantil. En este trabajo se empleará el contraste de Shapiro-Wilk, que chequea el contraste:

$$\begin{cases} H_0: & \text{Los residuos del modelo siguen una distrución normal.} \\ H_1: & \text{Los residuos del modelo no siguen una distribución normal.} \end{cases}$$

Usando el siguiente estadístico de contraste:

$$\omega = \sum_{t=1}^{\lfloor T/2 \rfloor} \frac{b_{t,T} \left(y_{(T-t+1)} - y_{(t)} \right)^2}{T s_y^2},$$

donde $y_{(t)}$ es el estadístico ordenado de orden t y las constantes $b_{t,T}$ se obtienen a partir de la inversa de la distribución normal estándar.

Se rechaza la normalidad cuando los valores de ω son pequeños. Los valores de $b_{t,T}$ y la distribución de ω bajo la hipótesis nula fueron tabulados por Shapiro y Wilk.

Capítulo 3

Desagregación temporal

La frecuencia temporal con la que se registran los datos juega un papel muy importante en el análisis de series temporales. Disponer de registros muestreados con alta frecuencia permite detectar ciertos patrones estacionales y comportamientos cíclicos a corto plazo que podrían pasar desapercibidos con una baja frecuencia de muestreo. También, en términos de predicción, una frecuencia más alta puede mejorar la precisión de los pronósticos, especialmente en horizontes temporales cortos.

La simple interpolación lineal o métodos similares no suelen ser adecuados para desagregar series temporales económicas, ya que ignoran las complejas relaciones y patrones estacionales que suelen estar presentes en estos datos. Por lo tanto, se han desarrollado métodos específicos para abordar este problema, como el método de Chow-Lin, Fernández o Litterman.

En este capitulo se proporciona una breve introducción al problema de desagregación temporal de una serie. Mientras que la Sección 3.1 proporciona el marco teórico de este tópico, la Sección 3.2 se centra en los diferentes métodos de desagregación utilizados en este trabajo, siguiendo principalmente la descripción de Melo-Velandia (2010)

3.1. Marco teórico

El objetivo de las técnicas de desagregación temporal es generar una estimación de una serie de tiempo no observada de alta frecuencia (high-frequency), partiendo de una serie de tiempo observada de baja frecuencia (low-frequency). Siguiendo la notación de Ciammola et al. (2005), se denota al vector de baja frecuencia como

$$\mathbf{y}_l = (y_1, y_2, \ldots, y_T)^\intercal,$$

y al correspondiente vector desconocido de observaciones a más alta frecuencia por

$$\mathbf{y}_h = (y_{1,1}, y_{1,2}, \ldots, y_{1,s-1}, y_{1,s}, \ldots, y_{T,1}, y_{T,2}, \ldots, y_{T,s-1}, y_{T,s})^\intercal,$$

donde se ha considerado una serie de tiempo de baja frecuencia con T periodos (trimestres) y s sub-periodos (meses) a desagregar. A lo largo de este trabajo se lleva a cabo únicamente la desagregación mensual, por lo tanto, $s = 3$. Sea $\mathbf{y}_l$ el vector de baja frecuencia de tamaño $(T \times 1)$ e $\mathbf{y}_h$ el vector de alta frecuencia de tamaño $(sT \times 1)$.

La solución básica a este problema consiste en dividir cada valor de $\mathbf{y}_l$ por s, lo cual supone que todos los meses dentro de un trimestre se comportan de la misma manera. Sin embargo, esto no es

cierto para las series de tiempo macroeconómicas, que suelen caracterizarse por estacionalidad, ciclos y efectos calendario, lo cual revela diferencias en frecuencias sub-anuales.

Ante un problema de desagregación de series es muy útil contar con variables de referencia a alta frecuencia, las cuales definirán el tamaño de la serie desagregada. En este caso, se asume que se dispone una matriz $\mathbf{X}_h$ de tamaño $n \times k$, siendo k el número de variables de referencia y $n \geq sT$, siendo n la longitud de cada serie de referencia.

Se supone que existe una relación lineal entre la serie de baja frecuencia y su correspondiente serie de alta frecuencia. Es decir, existe una matriz $\mathbf{C}$ (que se denomina matriz de agregación) tal que $\mathbf{y}_l = \mathbf{C}\mathbf{y}_h$. Asumiendo la existencia de series de referencia que presentan un comportamiento relacionado con la serie de interés, es factible plantear un modelo de regresión a nivel de alta frecuencia.

$$\mathbf{y}_h = \mathbf{X}_h\boldsymbol{\beta} + \mathbf{u}_h \tag{3.1}$$

donde $\boldsymbol{\beta}$ es el vector de coeficientes de regresión de tamaño $k \times 1$ y u_h es el ruido de la serie. Por el momento, se supondrá que

$$\begin{aligned} \mathbb{E}(\mathbf{u}_h|\mathbf{X}_h) &= 0, \\ \mathbb{E}(\mathbf{u}_h\mathbf{u}_h^\intercal|\mathbf{X}_h) &= \mathbf{V}_h \end{aligned} \tag{3.2}$$

sin ninguna forma específica para $\mathbf{V}_h$.

La matriz $\mathbf{C}$ de tamaño $T \times n$ se define como:

$$\mathbf{C} = \mathbf{I}_T \otimes \mathbf{c}^\intercal, \tag{3.3}$$

donde $\mathbf{c}$ es un vector de agregación de dimensión $s \times 1$, cuya forma depende del tipo de variable que representa $\mathbf{y}_l$.

El vector de agregación puede tomar las siguientes formas:

- Si $\mathbf{y}_l$ representa un **flujo**[1], entonces

$$\mathbf{c} = (1, 1, \ldots, 1)^\intercal.$$

- Si $\mathbf{y}_l$ representa un **promedio**, entonces

$$\mathbf{c} = \left(\frac{1}{T}, \frac{1}{T}, \ldots, \frac{1}{T}\right)^\intercal.$$

- Si $\mathbf{y}_l$ es un **stock**[2] **observado al final del período**, entonces

$$\mathbf{c} = (0, 0, \ldots, 1)^\intercal.$$

- Si $\mathbf{y}_l$ es un **stock observado al inicio del período**, entonces

$$\mathbf{c} = (1, 0, \ldots, 0)^\intercal.$$

Multiplicando los dos lados de la ecuación (3.1) por la matriz de agregación $\mathbf{C}$, se obtiene la serie de baja frecuencia:

$$\mathbf{C}\mathbf{y}_h = \mathbf{C}\mathbf{X}_h\boldsymbol{\beta} + \mathbf{C}\mathbf{u}_h.$$

[1] Un *flujo* es una variable que se mide a lo largo de un período de tiempo, como por ejemplo el ingreso mensual.
[2] Un *stock* es una variable que se mide en un punto específico en el tiempo, como el capital al inicio del año o la población al 1 de enero.

$$\mathbf{y}_l = \mathbf{X}_l\boldsymbol{\beta} + \mathbf{u}_l, \tag{3.4}$$

con $\mathbb{E}(\mathbf{u}_l\mathbf{u}_l^\intercal \mid \mathbf{X}_h) = \mathbf{C}\mathbf{V}_h\mathbf{C}^\intercal = \mathbf{V}_l$. De esta forma, se pueden desagregar las series de baja frecuencia en función de las series de alta frecuencia, donde $\mathbf{y}_l = \mathbf{C}\mathbf{y}_h$, $\mathbf{X}_l = \mathbf{C}\mathbf{X}_h$ y $\mathbf{u}_l = \mathbf{C}\mathbf{u}_h$.

Al objeto de hacer más claras las ecuaciones anteriores, se introduce a continuación un sencillo ejemplo siguiendo a Melo-Velandia (2010).

Considérense observaciones trimestrales sobre dos años de una variable $\mathbf{y}_l^\intercal = (y_1, y_2, \ldots, y_8)$, que se quiere desagregar a frecuencia mensual. Además, se dispone de observaciones mensuales de dos variables indicadoras $\mathbf{X}_h = (X_{i,j})$. Entonces, el número de periodos de baja frecuencia es de ocho trimestres (cuatro trimestres por año) ($T = 8$), y el número de subperiodos de alta frecuencia (número de meses en los dos años) es $n = sT = 3 \times 8 = 24$. En este caso, la matriz $\mathbf{X}_h$ tiene veinticuatro filas y dos columnas, de la siguiente manera:

$$\mathbf{X}_h = \begin{pmatrix} x_{1,1} & x_{2,1} & x_{3,1} & \cdots & x_{20,1} & x_{21,1} & x_{22,1} & x_{23,1} & x_{24,1} \\ x_{1,2} & x_{2,2} & x_{3,2} & \cdots & x_{20,2} & x_{21,2} & x_{22,2} & x_{23,2} & x_{24,2} \end{pmatrix}^\intercal$$

Se supone que las observaciones trimestrales de las variables indicadoras se agregan mediante suma. Por lo tanto, los vectores que representan las observaciones trimestrales de estas variables se obtienen sumando las observaciones mensuales correspondientes a cada uno de los meses dentro del trimestre. Es por esta razón que, en este ejemplo, la matriz de agregación toma la siguiente forma:

$$\mathbf{C} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \otimes \begin{pmatrix} 1 & 1 & 1 \end{pmatrix}$$

Al multiplicar $\mathbf{X}_h$ por $\mathbf{C}$, se obtiene el indicador de baja frecuencia $\mathbf{X}_l$:

$$\mathbf{X}_l = \begin{pmatrix} \sum_{n=1}^{3} x_{n,1} & \sum_{n=4}^{6} x_{n,1} & \cdots & \sum_{n=22}^{24} x_{n,1} \\ \sum_{n=1}^{3} x_{n,2} & \sum_{n=4}^{6} x_{n,2} & \cdots & \sum_{n=22}^{24} x_{n,2} \end{pmatrix}^\intercal$$

Este mismo procedimiento se aplica a la variable $\mathbf{y}_h$, obteniendo la variable $\mathbf{y}_l$, donde la suma de los meses se iguala a los valores trimestrales. De esta forma, la variable no observada de alta frecuencia $\mathbf{y}_h$ es:

$$\mathbf{y}_h = (y_{1,1}, y_{1,2}, y_{1,3}, y_{2,1}, y_{2,2}, y_{2,3}, \cdots, y_{8,1}, y_{8,2}, y_{8,3})^\intercal$$

Al multiplicar $\mathbf{y}_h$ por $\mathbf{C}$, se obtiene la variable de baja frecuencia $\mathbf{y}_l$:

$$\begin{aligned}\mathbf{y}_l =& \left(\sum_{n=1}^{3} y_{n,1}, \sum_{n=4}^{6} y_{n,1}, \sum_{n=7}^{9} y_{n,1}, \cdots, \sum_{n=19}^{21} y_{n,1}, \sum_{n=22}^{24} y_{n,1}\right)^{\intercal} \\ =&(y_1, y_2, y_3, y_4, y_5, y_6, y_7, y_8)^{\intercal}\end{aligned}$$

Teniendo en cuenta el marco teórico descrito y la notación expuesta, se introducen en la siguiente sección algunos métodos de desagregación.

3.2. Métodos univariantes de desagregación temporal

Los modelos de Chow and Lin (1971), Litterman (1983) y Fernandez (1981) son técnicas utilizadas para la desagregación temporal, basadas en regresión lineal tal y como se expresa en la ecuación (3.4).

Estas metodologías se emplean para desagregar una serie de baja frecuencia en una serie de alta frecuencia. Suponen que existe una relación lineal entre la serie de baja frecuencia y la serie correspondiente de alta frecuencia. La regresión lineal se utiliza para estimar los coeficientes de dicha relación, los cuales luego se aplican a la serie de alta frecuencia para obtener la serie desagregada. El problema de la desagregación temporal se resuelve identificando el mejor estimador lineal insesgado de la variable $\boldsymbol{\beta}$ en la ecuación de la regresión (3.4). A continuación, se explican los tres métodos de desagregación mencionados.

3.2.1. Método de Chow-Lin

En este método la solución óptima se obtiene a través de la suma de residuos al cuadrado y derivando con respecto a $\boldsymbol{\beta}$, empleando la teoría de la estimación lineal insesgada óptima (Hoff (2022)), por lo que los parámetros estimados son

$$\hat{\boldsymbol{\beta}} = \left[\mathbf{X}_l^{\intercal}\mathbf{V}_l^{-1}\mathbf{X}_l\right]^{-1}\mathbf{X}_l^{\intercal}\mathbf{V}_l^{-1}\mathbf{y}_l$$

$$\hat{\boldsymbol{\beta}} = \left[(\mathbf{C}\mathbf{X}_h)^{\intercal}(\mathbf{C}\mathbf{V}_h\mathbf{C}^{\intercal})^{-1}(\mathbf{C}\mathbf{X}_h)\right]^{-1}(\mathbf{C}\mathbf{X}_h)^{\intercal}(\mathbf{C}\mathbf{V}_h\mathbf{C}^{\intercal})^{-1}\mathbf{y}_l \tag{3.5}$$

donde $\hat{\boldsymbol{\beta}}$ el estimador de mínimos cuadrados de $\boldsymbol{\beta}$ en la regresión (3.4), y $\mathbf{V}_h$ la matriz de varianzas y covarianzas correspondiente a los datos en alta frecuencia. Si se quiere expresar esta matriz en términos de baja frecuencia, se puede hacer una transformación mediante la matriz de agregación temporal $\mathbf{C}$ que consiste en multiplicar $\mathbf{V}_h$ por $\mathbf{C}$ y su transpuesta, es decir, $\mathbf{C}\mathbf{V}_h\mathbf{C}^{\intercal}$. De este modo, es posible obtener una predicción para la serie no observada a partir de los datos agregados.

$$\hat{\mathbf{y}}_h = \mathbf{X}_h\hat{\boldsymbol{\beta}} + \mathbf{V}_h\mathbf{C}^{\intercal}(\mathbf{C}\mathbf{V}_h\mathbf{C}^{\intercal})^{-1}(\mathbf{y}_l - (\mathbf{C}\mathbf{X}_h)\hat{\boldsymbol{\beta}}) \tag{3.6}$$

El estimador de $\boldsymbol{\beta}$ y, en consecuencia, la serie estimada $\hat{\mathbf{y}}_h$ está condicionada a la forma de $\mathbf{V}_h$ para la cual se proponen diferentes supuestos partiendo del mismo modelo. La solución presentada por Chow-Lin supone el simple proceso de Markov para $\mathbf{u}_h$

$$\mathbf{u}_t = \rho\mathbf{u}_{t-1} + \boldsymbol{\epsilon}_t, \tag{3.7}$$

con $\mathbb{E}(\boldsymbol{\epsilon}_t^2) = \sigma_\epsilon^2$, $t = 1, ..., n$ y $u_0 = 0$. Bajo esta suposición, se tiene que la matriz de covarianza $\mathbf{V}_h$

tiene una forma de Toeplitz

$$\mathbf{V}_h = \frac{\sigma_\epsilon^2}{1-\rho^2} \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & 1 \end{pmatrix}, \tag{3.8}$$

Nótese que $\mathbf{V}_h$ depende de los valores de σ_ϵ^2 y ρ, de modo que la estimación de la matriz de covarianzas requiere la estimación de estos parámetros. La propuesta de Chow y Lin para estimar ρ y σ_ϵ es un método iterativo. Sin embargo, estas iteraciones presentan una mayor complejidad cuando se trata de una desagregación temporal más detallada. Autores como Bournay and Laroque (1979) sugieren que, si se asume que los errores de alta frecuencia siguen una distribución normal, es posible estimar $\boldsymbol{\beta}$, ρ y σ_ϵ utilizando el método de máxima verosimilitud.

3.2.2. Método de Fernández

El método de Fernandez (1981), adaptando el enfoque de Chow y Lin, propone una estructura de residuos basada en un paseo aleatorio:

$$\mathbf{u}_t = \mathbf{u}_{t-1} + \boldsymbol{\epsilon}_t, \;\; \epsilon_t \sim \mathcal{N}(0, \sigma_\epsilon^2) \tag{3.9}$$

donde $\boldsymbol{\epsilon}_t$ es ruido blanco con media cero y varianza σ_ϵ^2. Aplicando primeras diferencias al modelo (3.1) mediante la matriz $\mathbf{D}$ de tamaño $sT \times sT$:

$$\mathbf{D}\mathbf{y}_h = \mathbf{D}\mathbf{X}_h\boldsymbol{\beta} + \mathbf{D}\mathbf{u}_h, \tag{3.10}$$

donde

$$\mathbf{D} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{pmatrix},$$

Fernández introduce una transformación que vincula las primeras diferencias de alta y baja frecuencia. De esta manera, $\mathbf{y}_h$ se convierte en un estimador lineal insesgado óptimo, resultando en las siguientes predicciones:

$$\hat{\mathbf{y}}_h = \mathbf{X}_h\hat{\boldsymbol{\beta}} + (\mathbf{D}^\intercal\mathbf{D})^{-1}\mathbf{C}^\intercal \left(\mathbf{C}(\mathbf{D}^\intercal\mathbf{D})^{-1}\mathbf{C}^\intercal\right)^{-1}(\mathbf{y}_l - \mathbf{C}\mathbf{X}_h\hat{\boldsymbol{\beta}}),$$

con la estimación de $\boldsymbol{\beta}$ dada por

$$\hat{\boldsymbol{\beta}} = \left[\mathbf{X}_h^\intercal\mathbf{C}^\intercal \left(\mathbf{C}(\mathbf{D}^\intercal\mathbf{D})^{-1}\mathbf{C}^\intercal\right)^{-1}\mathbf{C}\mathbf{X}_h\right]^{-1} \mathbf{X}_h^\intercal\mathbf{C}^\intercal \left(\mathbf{C}(\mathbf{D}^\intercal\mathbf{D})^{-1}\mathbf{C}^\intercal\right)^{-1}\mathbf{y}_l.$$

El supuesto del paseo aleatorio para los errores (3.9) permite a Fernández evitar la estimación de parámetros adicionales, como el ρ del modelo original de Chow y Lin.

3.2.3. Método de Litterman

El método de Litterman (1983), similar a Fernández, generaliza el enfoque para la desagregación temporal, asumiendo una relación lineal entre la serie a estimar y las series de referencia, como en el modelo (3.1). La dinámica de los residuos se modeliza como:

$$\begin{aligned}\mathbf{u}_t &= \mathbf{u}_{t-1} + \boldsymbol{\epsilon}_t \\ \boldsymbol{\epsilon}_t &= \rho\boldsymbol{\epsilon}_{t-1} + \boldsymbol{\epsilon}_t\end{aligned}$$

donde $\boldsymbol{\epsilon}_t$ es ruido blanco con varianza σ_e^2 y ρ representa la autocorrelación de ϵ_t. En las ecuaciones anteriores, se observa que $\boldsymbol{\epsilon}_t$ depende de $\boldsymbol{\epsilon}_{t-1}$ y $\boldsymbol{\epsilon}_t$, y sustituyendo la segunda ecuación en la primera, se tiene:

$$\mathbf{u}_t - \mathbf{u}_{t-1} = \rho(\mathbf{u}_{t-1} - \mathbf{u}_{t-2}) + \boldsymbol{\epsilon}_t$$

Reordenando los términos

$$\boldsymbol{\epsilon}_t = \mathbf{u}_t - (1+\rho)\mathbf{u}_{t-1} + \rho\mathbf{u}_{t-2} \tag{3.11}$$

Matricialmente, la ecuación (3.11) se expresa como:

$$\mathbf{e} = \mathbf{H}\mathbf{D}\mathbf{u}_h$$

donde $\mathbf{e} = [e_1, \ldots, e_n]^\intercal$ y $\mathbf{H}$ es una matriz de tamaño $n \times n$ definida como:

$$\mathbf{H} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ -\rho & 1 & 0 & \cdots & 0 & 0 \\ 0 & -\rho & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -\rho & 1 \end{pmatrix}.$$

Despejando u_h:

$$\mathbf{u}_h = (\mathbf{H}\mathbf{D})^{-1}\mathbf{e} = \mathbf{D}^{-1}\mathbf{H}^{-1}\mathbf{e}$$

Por lo tanto, la matriz de covarianza de los residuos $\mathbf{V}_h$ se calcula como:

$$\mathbf{V}_h = \mathbb{E}[\mathbf{u}_h\mathbf{u}_h^\intercal] = (\mathbf{H}\mathbf{D})^{-1}\mathbb{E}[\mathbf{e}\mathbf{e}^\intercal]((\mathbf{H}\mathbf{D})^{-1})^\intercal = (\mathbf{D}^\intercal\mathbf{H}^\intercal\mathbf{H}\mathbf{D})^{-1}\sigma_e^2$$

Sustituyendo $\mathbf{V}_h$ en las soluciones de Chow y Lin (3.5) y (3.6), se obtienen los estimadores para $\mathbf{y}_h$ y $\boldsymbol{\beta}$:

$$\hat{\mathbf{y}}_h = \mathbf{X}_h\hat{\boldsymbol{\beta}} + (\mathbf{D}^\intercal\mathbf{H}^\intercal\mathbf{H}\mathbf{D})^{-1}\mathbf{C}\left(\mathbf{C}(\mathbf{D}^\intercal\mathbf{H}^\intercal\mathbf{H}\mathbf{D})^{-1}\mathbf{C}^\intercal\right)^{-1}\left(\mathbf{y}_l - \mathbf{X}_l\hat{\boldsymbol{\beta}}\right),$$

y

$$\hat{\boldsymbol{\beta}} = \left[\mathbf{X}_l^{-1}\left(\mathbf{C}(\mathbf{D}^\intercal\mathbf{H}^\intercal\mathbf{H}\mathbf{D})^{-1}\mathbf{C}^\intercal\right)^{-1}\mathbf{X}_l\right]^{-1}\mathbf{X}_l^{-1}\left(\mathbf{C}(\mathbf{D}^\intercal\mathbf{H}^\intercal\mathbf{H}\mathbf{D})^{-1}\mathbf{C}^\intercal\right)^{-1}\mathbf{y}_l.$$

Nótese que el método de Fernández es un caso particular del de Litterman, con $\rho = 0$ y también un caso límite del de Chow-Lin cuando $\rho \to 1$.

En definitiva, se han presentado muy brevemente en este capítulo los fundamentos sobre los que pivotan tres importantes métodos de desagregación de series. Todos ellos se emplearán más adelante con las series de ocupados y parados, utilizando como indicadores de alta frecuencia las afiliaciones y el paro registrado, respectivamente, con el fin de obtener series mensuales. Previamente, en el siguiente capítulo se aborda el problema de cómo proceder para corregir las series ante la presencia de patrones estacionales, efectos calendario y datos atípicos.

Capítulo 4

Corrección de series temporales

Este capítulo aborda todo lo relacionado con la corrección de series temporales y su estructura. El contenido se organiza del siguiente modo. En la Sección 4.1 se justifica la necesidad de corregir los efectos de calendario de una serie temporal y se introducen los conceptos clave relacionados con este modelo. La Sección 4.2 se ocupa del problema de los de datos atípicos en el marco de series temporales. En la Sección 4.3 se exploran los detalles técnicos y metodológicos asociados con el proceso de corrección. Finalmente, en la Sección 4.4 se explican diferentes test que se utilizan para evaluar si hay presencia o no de estacionalidad residual.

Para la elaboración de este capítulo se han considerado diferentes trabajos, destacando entre ellos INE (2002a), INE (2002b), IGE (2017) y Maravall and Caporello (2004).

4.1. Motivación y preliminares

Una serie temporal puede presentar diferentes patrones subyacentes, que permiten descomponerla en diversas componentes. Las principales son:

- **Tendencia** (T_t): Representa los movimientos a largo plazo de la serie (superiores a 8 años) y refleja la evolución estructural, como el crecimiento económico, cambios en la población, inflación, entre otros.
- **Ciclo** (C_t): Oscilaciones económicas que duran entre 2 y 8 años, generadas por factores como cambios en la industria o la moda.
- **Componente estacional** (S_t): Movimiento periódico de menos de un año, como los picos y valles regulares asociados con estaciones o fechas específicas.
- **Efectos de calendario** (CAL_t): Movimientos relacionados con el calendario, como las fiestas móviles o el número de días laborables en cada mes.
- **Componente irregular** (I_t): Movimientos erráticos y no predecibles, como huelgas, guerras, elecciones, etc.

En el análisis de series temporales socioeconómicas es habitual trabajar con datos corregidos de efectos estacionales y de calendario. Esta práctica permite centrarse en los cambios reales de las variables, eliminando las fluctuaciones que se repiten cada cierto tiempo, como las provocadas por las

estaciones del año o por diferencias en el número de días laborables. De este modo, se obtiene una visión más clara de cómo evolucionan realmente las series.

Con el objetivo de coordinar y armonizar las estadísticas oficiales entre los países de la Unión Europea, Eurostat ha establecido una serie de directrices metodológicas relativas al tratamiento de los efectos estacionales y de calendario en las series temporales. Estas recomendaciones, recogidas en el manual Eurostat (2024), tienen como finalidad garantizar la comparabilidad internacional y la coherencia en la interpretación de los indicadores económicos.

En línea con estas directrices, se promueve el uso de la metodología TRAMO-SEATS para la corrección de los efectos estacionales y de calendario. A fin de facilitar su aplicación, Eurostat impulsa el uso de la herramienta JDemetra+, que incorpora esta metodología en un software accesible y compatible con los requisitos técnicos establecidos.

El modelo de descomposición puede adoptar una forma aditiva o multiplicativa, según cómo se combinen las componentes de la serie temporal. En el modelo aditivo, se asume que las distintas componentes afectan a la serie de forma independiente y se suman entre sí:

$$y_t = T_t + C_t + S_t + CAL_t + I_t.$$

En cambio, el modelo multiplicativo considera que las componentes están interrelacionadas y se representa como:

$$y_t = T_t \cdot C_t \cdot S_t \cdot CAL_t \cdot I_t.$$

Este modelo es apropiado cuando la amplitud de las fluctuaciones aumenta con el nivel de la serie: es decir, cuando los valores estacionales o las irregularidades son más intensos cuando la serie está en niveles altos, y más suaves cuando los valores son bajos.

En la práctica, es complicado diferenciar las componentes de ciclo y tendencia. Por este motivo, se engloban en una única componente denominada **ciclotendencia** (P_t), lo que da lugar a la siguiente descomposición en el caso aditivo:

$$y_t = P_t + S_t + I_t + CAL_t.$$

La desestacionalización de una serie temporal consiste extraer las componentes estacionales y los efectos de calendario. Este proceso permite obtener una serie ajustada, que ofrece una estimación más precisa de la tendencia y el ciclo subyacente, eliminando las fluctuaciones estacionales y los efectos específicos del calendario.

Es importante tener en cuenta que las series desestacionalizadas dependen del método utilizado y de las hipótesis del modelo elegido para dicha corrección. Un ajuste estacional incorrecto puede generar señales falsas y resultados erróneos. Existen dos enfoques principales para el ajuste estacional de series temporales, ambos reconocidos como válidos por el Sistema Estadístico Europeo (ESS). Por un lado, están los métodos basados en modelos $ARIMA$, que ajustan la serie utilizando modelos estadísticos adaptados a su estructura particular. Por otro lado, existen los enfoques basados en filtros fijos, que aplican filtros lineales predefinidos e independientes de la serie para extraer las componentes estacionales.

El INE recomienda el método TRAMO-SEATS, que es el empleado en este trabajo. Los detalles del mismo se explican en la Sección 4.3.

4.2. Valores atípicos

Un dato atípico o outlier es una observación que se desvía de forma significativa del resto de los datos en un conjunto de observaciones. Estos valores pueden surgir por diversos motivos, como errores humanos o eventos inesperados, tales como la crisis económica de 2008 o la pandemia del COVID-19.

La presencia de outliers en un conjunto de datos puede afectar negativamente el ajuste de un modelo estadístico, por lo que es fundamental identificarlos y tratarlos adecuadamente. Este tratamiento cobra especial importancia cuando los valores atípicos se encuentran al final de la serie temporal, ya que una interpretación incorrecta de los mismos podría alterar significativamente el comportamiento del modelo ajustado.

En esta sección se introduce el concepto de dato atípico y se explican los diferentes tipos de atípicos que existen ya que cuando se ajusten los modelos mas adelante estos estarán presentes. Para elaborar esta sección se ha consultado IGE (2017).

El impacto de un dato atípico sobre las series observadas se puede modelizar de la siguiente manera:

$$y_t = \sum_{j=1}^{k} \xi_j(B) w_j I_t^{(\tau_j)} + x_t,$$

donde y_t denota la serie de tiempo observada, x_t es una serie que sigue el modelo $ARIMA$, w_j es un impacto inicial del valor atípico en el tiempo $t = \tau_j$, e $I_t^{(\tau_j)}$ es una variable indicadora tal que vale 1 para $t = \tau_j$ y 0 en caso contrario. La función $\xi_j(B)$ determina la dinámica del valor atípico que ocurre en el tiempo $t = \tau_j$, y B es el operador retardo (es decir, $B^k x_t = x_{t-k}$).

A continuación, se describen los cuatro tipos de valores atípicos más comunes en el análisis de series temporales:

- Atípico aditivo (AO): Se presenta como un punto atípico que ocurre en un momento específico t_0. En este caso, la función dinámica es $\xi_j(B) = 1$, dando lugar a la variable indicadora:

$$AO_t^{t_0} = \begin{cases} 1 & \text{si } t = t_0, \\ 0 & \text{si } t \neq t_0. \end{cases}$$

- Cambio de nivel (LS): Representa un cambio estructural permanente en el nivel de la serie a partir de cierto instante t_0. La dinámica de este cambio de nivel está dada por la función $\xi_j(B) = \frac{1}{1-B}$, lo que da lugar a la siguiente variable de regresión:

$$LS_t^{t_0} = \begin{cases} 0 & \text{si } t < t_0, \\ 1 & \text{si } t \geq t_0. \end{cases}$$

- Cambio temporal (TC): Corresponde a un efecto temporal en el instante t_0 que decae exponencialmente en los siguientes períodos. La dinámica de este cambio temporal está dada por $\xi_j(B) = \frac{1}{1-\delta B}$, lo que resulta en la variable de regresión:

$$TC_t^{t_0} = \begin{cases} 0 & \text{si } t < t_0, \\ \delta^{t-t_0} & \text{si } t \geq t_0. \end{cases}$$

  Aquí, δ es la tasa de decrecimiento hacia el nivel anterior, donde $0 < \delta < 1$. Es importante notar que cuando δ tiende a 0, el TC se convierte en un AO, y cuando δ tiende a 1, el TC se convierte en un LS.

- Atípico estacional (SO): Se representa un cambio brusco en el patrón estacional en el tiempo t_0, y mantiene el nivel de la serie con un cambio significativo en los períodos siguientes. La variable de regresión que modela este atípico es la siguiente:

$$SO_t^{t_0} = \begin{cases} 0 & \text{si } t < t_0, \\ 1 & \text{si } t \geq t_0 \text{ y } t \text{ está en el mismo mes o trimestre que } t_0, \\ -\frac{1}{s-1} & \text{si } t \neq t_0 \text{ y } t \text{ está en un período diferente,} \end{cases}$$

  donde s representa el número de períodos en el ciclo estacional.

4.3. Metodología TRAMO-SEATS

La metodología TRAMO-SEATS es una herramienta desarrollada por Víctor Gómez y Agustín Maravall en el Banco de España para el análisis de series temporales. Combina dos programas: TRAMO (Time Series Regression with $ARIMA$ Noise, Missing Observations and Outliers) y SEATS (Signal Extraction in $ARIMA$ Time Series). Aunque SEATS depende de TRAMO para su funcionamiento, TRAMO puede utilizarse de forma independiente.

A continuación se detallan ambos programas y para una descripción más completa, puede consultarse Gómez and Maravall Herrero (1998).

TRAMO

TRAMO actúa como un filtro que limpia la serie temporal, corrigiendo anomalías y preparando el terreno para el análisis posterior. Identifica y corrige automáticamente diversos tipos de valores atípicos y utiliza variables de regresión para modelar eventos específicos del calendario, como la influencia de la Pascua.

El modelo propuesto para el proceso de desestacionalización tiene la siguiente forma:

$$Y_t = h_t\beta + x_t,$$

donde:

- Y_t es la serie original.
- $\beta = (\beta_1, \ldots, \beta_n)$ es un vector de coeficientes de regresión.
- $h_t = (h_{1t}, \ldots, h_{nt})$ son n variables de regresión, que corresponden a las distintas variables de efectos de calendario y valores atípicos.
- x_t es una componente de error que sigue el proceso general $ARIMA$.

A continuación, se explica cómo se construyen las variables regresoras de los efectos de calendario, concretamente los efectos de días laborables, año bisiesto y Semana Santa. La metodología empleada para esta construcción está basada en INE (2019) y Gómez and Gómez (1998).

- **Ciclo Semanal**: En el análisis de series temporales que dependen del día de la semana, se emplean variables regresoras diseñadas para capturar las diferencias entre días laborables (lunes a viernes) y fines de semana (sábado y domingo).

La variable regresora de días laborables se construye de la siguiente manera:

$$D_t = \sum_{i=1}^{5} X_{i,t} - \frac{5}{2} \sum_{i=6}^{7} X_{i,t},$$

donde $X_{i,t}$ indica la presencia del día i en el período t.

El efecto del Ciclo Semanal sobre la serie se modeliza como:

$$CS_t = \alpha D_t.$$

- **Año bisiesto**: Uno de los efectos más conocidos es el causado por los años bisiestos, que agregan un día más en el mes de febrero en comparación con los años no bisiestos.

  Para analizar este efecto de manera aislada, se considera el periodo de los años 1901 a 2099, en el cual un año de cada cuatro es bisiesto y el promedio de días de febrero se calcula como:

$$\frac{28 + 28 + 28 + 29}{4} = 28.25$$

  Teniendo en cuenta este promedio, el efecto del año bisiesto sobre los meses de febrero en años no bisiestos se puede modelar como la diferencia entre 28 días y el valor promedio de 28.25, es decir:

$$28 - 28.25 = -0.25$$

  Para los meses de febrero de años bisiestos, el efecto se expresa como la diferencia entre 29 días y 28.25 días:

$$29 - 28.25 = 0.75$$

  Es importante señalar que el efecto para los meses que no sean febrero es nulo. Este enfoque se justifica porque se asume que el impacto de los años bisiestos es proporcional a la diferencia entre el número medio de días en febrero y el número de días de cada mes de febrero, de forma que, en promedio, el efecto es cero. Si no se hace este ajuste, se estaría alterando la media de la serie en los meses de febrero, lo cual no sería apropiado.

  En resumen, el regresor que captura el efecto de los años bisiestos será una serie temporal determinista $b_{am,t}$ de la misma longitud que la serie de interés y_t, con los siguientes valores:

$$b_{am,t} = \begin{cases} 0 & \text{si } m \neq 2, \\ -0.25 & \text{si } m = 2 \text{ y el año no es bisiesto}, \\ 0.75 & \text{si } m = 2 \text{ y el año es bisiesto}. \end{cases}$$

  De este modo, el efecto del año bisiesto se expresa como:

$$LY_t = \beta b_{am,t}$$

- **Semana Santa**: La Semana Santa es una festividad que se celebra entre el 22 de marzo y el 25 de abril, dependiendo de la fecha en que se ubique el domingo de Pascua, que corresponde al primer domingo después de la primera luna llena tras el equinoccio de primavera.

  La variable que trata de modelizar el efecto de esta fiesta modeliza un cambio constante en el nivel de actividad diaria durante los d días previos a la Semana Santa. El valor de d suele ser proporcionado por el usuario.

  El valor asignado a marzo es igual a $p_M - m_M$, donde p_M es la proporción de los d días que caen en ese mes y m_M es el valor medio de las proporciones de los d días que caen en marzo durante

un largo período de tiempo. El valor asignado a abril es $p_A - m_A$, donde p_A y m_A se definen de manera análoga. Para el resto de los meses, la variable toma el valor cero.

Tomando que $m_M = m_A = \frac{1}{2}$, se define la variable regresora de la Semana Santa como:

$$P_t = \begin{cases} p_t - \frac{1}{2} & \text{si } t = \text{marzo ó abril,} \\ 0 & \text{en otro caso.} \end{cases}$$

En consecuencia, el efecto de la Semana Santa es:

$$E_t = \gamma P_t$$

.

Combinando todos los efectos de calendario anteriores, se llega a que dichos efectos se modelizan como:

$$CAL_t = CS_t + LY_t + E_t = \alpha D_t + \beta b_{am,t} + \gamma P_t.$$

y se cuantifica mediante el siguiente modelo:

$$Y_t = \alpha D_t + \beta b_{am,t} + \gamma P_t + \frac{\theta_q(B)\Theta_Q(B^s)}{\phi_p(B)\Phi_P(B^s)(1-B)^d(1-B^s)^D} a_t. \tag{4.1}$$

La identificación de los efectos de calendario se realiza mediante contrastes de significación estadística con las hipótesis nulas: $\alpha = 0$ (sin efecto de Ciclo Semanal), $\beta = 0$ (sin efecto de año bisiesto) y $\gamma = 0$ (sin efecto de Semana Santa) en el modelo (4.1). La serie ajustada por los efectos de calendario, denotada como x_t, se obtiene restando de la serie original los efectos significativos. Si los tres efectos resultan ser estadísticamente significativos, la fórmula para x_t es:

$$x_t = Y_t - \hat{\gamma} P_t - \hat{\alpha} D_t - \hat{\beta} b_{am,t}$$

El modelo que se ha identificado y estimado permite descomponer la serie corregida de los efectos de calendario en sus componentes subyacentes: tendencia, estacionalidad e irregularidad. Este modelo supone que cada uno de sus componentes sigue un proceso $ARIMA$, y los modelos correspondientes deben ser consistentes con el que describe la serie desagregada x_t.

SEATS

SEATS descompone la serie temporal linealizada en las componentes: tendencia, estacionalidad, componente irregular, acompañadas de su error (ruido blanco). No solo estima estas componentes, sino que también ofrece predicciones para cada una de ellas, incluyendo sus respectivos errores estándar. Finalmente, SEATS incorpora los efectos deterministas en cada componente para obtener las componentes finales. Este proceso de estimación se basa en la aplicación de filtros de Wiener-Kolmogorov, utilizando la información preprocesada por TRAMO. Para la elaboración de todo lo que sigue se ha empleado Maravall and Caporello (2004) y JDemetra+ Development Team (2024).

El programa SEATS parte de la suposición de que la serie temporal linealizada x_t sigue un modelo $ARIMA$, que escrito de manera compacta se tiene que

$$\varphi(B) x_t = \theta(B) a_t, \tag{4.2}$$

donde $\varphi(B)$ viene dado por:

$$\begin{aligned} \varphi(B) &= \phi(B)\Phi(B_s)(1-B)^d(1-B_s)^D \\ &= (1 + \phi_1 B + \cdots + \phi_p B^p)(1 + \Phi_1 B + \cdots + \Phi_P B^P)(1-B)^d(1-B^s)^D. \end{aligned}$$

Se supone, por simplicidad, que la serie temporal x_t se descompone en k componentes estocásticas, que son ortogonales entre sí. La agregación de estas componentes genera la serie original x_t, y se expresa como sigue:

$$x_t = \sum_{i=1}^{k} x_{i,t}, \tag{4.3}$$

donde $t = 1, \ldots, T$, siendo T el total de observaciones de la serie. El subíndice i hace referencia a las componentes ortogonales: tendencia, estacionalidad, irregularidad y error, con k denotando el número total de componentes. Exceptuando la componente de error, que se modeliza como ruido blanco, se asume que cada componente sigue un proceso $ARIMA$, el cual puede ser representado de forma análoga a la ecuación (4.2) como:

$$\varphi_i(B)x_{i,t} = \theta_i(B)a_{i,t},$$

donde $\varphi_i(B) = \phi_i(B)\delta_i(B)$, siendo $x_{i,t}$ la i-ésima componente no observada, y los polinomios $\varphi_i(B)$ y $\theta_i(B)$ tienen órdenes p_i y q_i, respectivamente. Además, $a_{i,t}$ representa la perturbación asociada a dicha componente. Es importante destacar que las perturbaciones $a_{i,t}$ y $a_{j,t}$ son independientes entre sí para $i \neq j$ y cualquier valor de t.

Dado que la agregación de modelos $ARIMA$ produce nuevamente un modelo de este tipo, la serie x_t debe ajustarse también a dicha clase de modelos. La idea fundamental es que cada componente puede representarse mediante un modelo $ARIMA$, y todos ellos son coherentes con la serie resultante de su combinación. Así, el modelo identificado y estimado para la serie observada x_t puede descomponerse en modelos equivalentes para cada una de sus componentes.

Formalmente, esto puede expresarse considerando que la serie observada x_t y cada una de sus k componentes $x_{i,t}$ siguen modelos $ARIMA$, de la forma $x_t = \frac{\theta(B)}{\varphi(B)}a_t$ y $x_{i,t} = \frac{\theta_i(B)}{\varphi_i(B)}a_{i,t}$, respectivamente. Al sustituir estas expresiones en la ecuación (4.3), se obtiene:

$$\sum_{i=1}^{k} \frac{\theta_i(B)}{\varphi_i(B)}a_{i,t} = \frac{\theta(B)}{\varphi(B)}a_t.$$

De esta ecuación, se pueden deducir las siguientes relaciones:

$$\varphi(B) = \prod_{i=1}^{k} \varphi_i(B),$$

y

$$\theta(B)a_t = \sum_{i=1}^{k} \varphi_{(i)}(B)\theta_i(B)a_{i,t}, \quad \text{donde} \quad \varphi_{(i)}(B) = \prod_{\substack{j=1 \\ j\neq i}}^{k} \varphi_j(B).$$

Los polinomios autoregresivos AR para cada componente, $\varphi_i(B)$, se obtienen de manera sencilla a partir de la factorización del polinomio AR total $\varphi(B)$, como se muestra en la siguiente expresión:

$$\varphi(B) = \prod_{i=1}^{k} \varphi_i(B). \tag{4.4}$$

En cuanto a los polinomios de medias móviles MA para las componentes, así como las varianzas de las innovaciones $V(a_i)$, no pueden ser derivados directamente de la relación

$$\theta(B)a_t = \sum_{i=1}^{k} \varphi_{n,i}(B)\theta_i(B)a_{i,t},$$

donde $\varphi_{n,i}(B)$ es el producto de todos los $\varphi_j(B)$, para $j = 1, \ldots, k$, excepto $\varphi_i(B)$. En este contexto, es necesario hacer suposiciones adicionales para resolver el problema de subidentificación, dado que existe un número infinito de descomposiciones que pueden ser válidas. Por lo tanto, aunque los polinomios MA y las varianzas de innovaciones no pueden identificarse directamente a partir del modelo de x_t, se puede proceder de la siguiente manera:

Para abordar este problema y lograr una descomposición única, se asume que el orden del polinomio MA de cada componente no es mayor que el orden del polinomio AR correspondiente, es decir, $p_i \leq q_i$. Esta suposición lleva a una solución canónica, en la cual todo el ruido blanco aditivo se asigna exclusivamente a la componente de error.

Para entender cómo SEATS realiza la factorización de los polinomios AR en la ecuación (4.4), es importante centrarse en el cálculo de sus raíces. Se considera nuevamente la ecuación original (4.2) y multiplicando ambos lados por x_{t-k}, con $k > q$, se tiene que:

$$\mathbb{E}(\varphi(B)x_t x_{t-k}) = \mathbb{E}(\theta(B)a_t x_{t-k}). \tag{4.5}$$

Por un lado, el término en el lado derecho desaparece de la siguiente forma:

$$\mathbb{E}[\theta(B)a_t x_{t-k}] = \theta(B)\mathbb{E}[a_t x_{t-k}] = \theta(B)\mathbb{E}(a_t)\mathbb{E}(x_{t-k}) = 0,$$

ya que a_t es ruido blanco (con media cero) y es independiente de x_{t-k}. Por otro lado, el término en el lado izquierdo, al tomar las esperanzas, se convierte en:

$$\mathbb{E}[\varphi(B)x_t x_{t-k}] = \varphi(B)\mathbb{E}[x_t x_{t-k}] = \varphi(B)\gamma_k,$$

donde γ_k es la autocovarianza entre x_t y x_{t-k}, y el operador de B actúa sobre el índice k. Sustituyendo en la ecuación (4.5), obtenemos:

$$\gamma_k + \varphi_1\gamma_{k-1} + \cdots + \varphi_p\gamma_{k-p} = 0. \tag{4.6}$$

La función γ_k es la solución de la ecuación diferencial homogénea (4.6), la cual está asociada a la siguiente ecuación característica:

$$z^p + \varphi_1 z^{p-1} + \cdots + \varphi_{p-1}z + \varphi_p = 0. \tag{4.7}$$

Las raíces de esta ecuación característica pueden ser tanto reales como complejas. Si se denotan las raíces de la ecuación (4.7) como $z_1, z_2, \ldots, z_p$, estas corresponden a las inversas de las raíces del polinomio $\varphi(B) = 0$, es decir, $z_i = B_i^{-1}$.

Como se puede ver en JDemetra+ Development Team (2024), el polinomio completo AR, $\varphi(B)$, asigna las raíces a las distintas componentes en función de su módulo y frecuencia, de la siguiente manera:

- La asignación de raíces de $(1-B)^d$ se realiza a la componente de tendencia.
- Para las raíces de $(1-B^s)^D = \big((1-B)(1+B+\cdots+B^{s-1})\big)^D$, se efectúa la siguiente asignación:
  1. Las raíces de $(1-B)^D$ se asignan a la componente de tendencia.
  2. Las raíces del polinomio $(1+B+\cdots+B^{s-1})^D$ corresponden a la componente estacional.
- Las raíces del polinomio $\phi_p(B)$ se asignan según el comportamiento de las raíces del polinomio característico:

$$\phi_p(z) = z^p + \phi_1 z^{p-1} + \cdots + \phi_p \quad \text{donde} \quad z = B^{-1}.$$

Las raíces se distribuyen de la siguiente manera:

1. Para raíces reales positivas:
   a) Si el módulo es mayor o igual a k_1, se asigna a la componente de tendencia.
   b) Si el módulo es menor que k_1, se asigna a la componente irregular.
2. Para raíces reales negativas:
   a) Si $s \neq 1$ y el módulo es mayor o igual a k_s, se asigna a la componente estacional.
   b) Si $s \neq 1$ y el módulo es menor que k_2, se asigna a la componente irregular.
   c) Si $s = 1$, se asigna a la componente irregular.
3. Para raíces complejas:
   a) Si la frecuencia ω está cerca de la frecuencia estacional, se asigna a la componente estacional.
   b) En caso contrario, se asigna a la componente irregular.

- Las raíces de $\Phi_P(B)$ se asignan de acuerdo con las propiedades de las raíces del polinomio $\Phi_P(z) = z^s + \phi_s$, donde $z = B^{-1}$.

  1. Si $\phi_s > 0$, las raíces se asignan a la componente irregular.
  2. Si $\phi_s < 0$, se define ϕ como la raíz real positiva de $(-\phi_s)^{\frac{1}{s}}$, de modo que el polinomio $\Phi_P(z)$ se puede expresar como:

  $$\Phi_P(z) = (z - \phi)(z^{s-1} + \phi z^{s-2} + \phi^2 z^{s-3} + \cdots + \phi^{s-1}).$$

  Las raíces se asignan de la siguiente manera:
  a) Si $D = 1$, la raíz AR correspondiente a $(1 - \alpha B)$ se asigna a la tendencia, mientras que las otras $s - 1$ raíces se asignan a la componente estacional.
  b) Si $D = 0$, la raíz se asigna a la componente estacional si $\phi_s < -0.2$ o si los tests de estacionalidad indican su presencia. En caso contrario, se asigna a la componente irregular.

En consecuencia, al realizar la factorización inicial del polinomio AR, se logra identificar los polinomios AR correspondientes a las componentes principales de tendencia, estacionalidad e irregularidad. Estas componentes contienen las raíces AR asociadas a cada una de ellas, lo que nos lleva a la ecuación (4.4).

Una vez definido un modelo teórico para las componentes, el siguiente paso es estimarlas, es decir, obtener las series temporales de cada $x_{i,t}$ a partir de los datos observados de x_t. Este proceso de estimación se lleva a cabo mediante Wiener-Kolmogorov. El procedimiento consiste en aplicar un filtrado de la serie x_t de la siguiente manera:

$$\hat{x}_{i,t} = v_i(B, F)x_t,$$

donde $v_i(B, F)$, con $F = B^{-1}$, son los filtros de SEATS. Estos filtros tienen como objetivo minimizar el error cuadrático medio entre el valor estimado y la componente teórica. Formalmente, podemos representar este proceso como la solución de un problema de optimización restringida:

$$\begin{aligned} \min_{x_{i,t}} \quad & \mathbb{E}\left(x_{i,t} - \hat{x}_{i,t}\right)^2 \\ \text{sujeto a} \quad & x_{i,t} = \frac{\theta_i(B)}{\phi_i(B)} a_{i,t}, \quad i = 1, \ldots, k. \end{aligned}$$

La solución a este problema nos permite obtener las estimaciones para cada $x_{i,t}$ de la forma siguiente:

$$\hat{x}_{i,t} = k_i \frac{\psi_i(B)\psi_i(F)}{\psi(B)\psi(F)} x_t,$$

donde $\psi(B) = \frac{\theta(B)}{\phi(B)}$ y $k_i = \frac{V(a_i)}{V(a)}$, siendo $V(a_i)$ y $V(a)$ las varianzas de $a_{i,t}$ y a_t, respectivamente.

El filtro de Wiener-Kolmogorov está diseñado para adaptarse tanto a la componente a estimar como al modelo de la serie. Esto implica que tanto el estimador de la componente como el propio filtro reflejan las características del conjunto de datos, lo que permite que el filtro se ajuste a cada serie individual. En la práctica, el filtro se aplica a una extensión de x_t que incluye predicciones y retrocesos derivados del modelo $ARIMA$. Este modelo es crucial para el procedimiento SEATS, ya que una especificación incorrecta puede dar lugar a una descomposición errónea.

4.4. Diagnosis del modelo

Una vez eliminada la estacionalidad y los efectos de calendario de la serie, se debe evaluar el modelo mediante el análisis de los residuos, tal como se explica en la Sección 2.4.3, es decir, verificando que los residuos se comporten como ruido blanco. Además, en esta sección se describen dos tests que permiten comprobar la presencia de estacionalidad residual: el QS-test y el F-test, que se detallan a continuación. Las hipótesis para estos contrastes son las siguientes:

$$\begin{cases} H_0: & \text{No existe estacionalidad residual,} \\ H_1: & \text{Existe estacionalidad residual.} \end{cases}$$

QS-test

El QS-test, propuesto por Maravall (2012), es una variante del contraste de Ljung-Box, pero calculado específicamente en los retardos estacionales, y sólo considerando las autocorrelaciones positivas. De manera más precisa, el estadístico QS se define como:

$$QS = T(T+2) \sum_{k=1}^{k} \left[\text{máx}\left(0, \hat{\gamma}_k \cdot l\right)\right]^2 \frac{1}{n - k \cdot l},$$

donde $k = 2$, lo que significa que solo se tienen en cuenta los primeros y segundos retardos estacionales. Así, la prueba verifica la correlación entre la observación real y las observaciones con retardo de uno y dos años. Es importante notar que cuando se trabaja con datos mensuales, $l = 12$, por lo que solo se consideran las autocovarianzas $\hat{\gamma}_{12}$ y $\hat{\gamma}_{24}$. En el caso de los datos trimestrales, $l = 4$.

Bajo la hipótesis nula H_0, que postula que los datos siguen una distribución independiente, el estadístico de la prueba sigue una distribución χ^2_{df}, siendo df los grados de libertad.

F-test

El F-test, basado en el modelo de análisis de varianza (ANOVA), tiene como hipótesis nula que las medias de todos los grupos que se comparan son idénticas. Este test nos permite analizar si existen diferencias entre los grupos, que en nuestro caso corresponden a los distintos períodos, como meses o trimestres.

El estadístico utilizado en este test se define como:

$$F = \frac{RSS_1/(k-1)}{RSS/(N-k)},$$

donde:

$$RSS_1 = \sum_{j=1}^{k} n_j \left(\text{res}_j - \text{res}_{..}\right)^2,$$
$$RSS = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(\text{res}_{ij} - \text{res}_j\right)^2.$$

Aquí, res_{ij} representa el residuo correspondiente al grupo j, con $j = 1, \ldots, k$, y $i = 1, \ldots, n_j$ indica la posición de cada elemento dentro del grupo. Además, $\text{res}_{..}$ es la media global de los residuos y res_j es la media muestral de los residuos para el grupo j.

Este estadístico sigue una distribución de Fisher-Snedecor con $k-1$ y $N-k$ grados de libertad, donde k es el número de grupos (12 para datos mensuales y 4 para datos trimestrales), y $N = \sum_{j=1}^{k} n_j$ es el tamaño total de la muestra.

El F-test sobre las variables ficticias estacionales verifica la presencia de estacionalidad residual. El modelo utilizado en el trabajo emplea variables ficticias estacionales (efecto medio y 11 variables ficticias estacionales para datos mensuales, efecto medio y 3 para datos trimestrales) para describir el comportamiento de la serie temporal. La prueba se emplea para verificar si las variables ficticias estacionales no son estadísticamente significativas en conjunto. Cuando se rechaza esta hipótesis, se asume que la estacionalidad residual está presente.

Esta prueba está basada en las pruebas χ^2 y F-test basadas en el modelo para efectos estacionales fijos propuestas por Lytras et al. (2007), que se basa en las estimaciones de las variables ficticias de regresión y las correspondientes t-estadísticas del modelo $RegARIMA$, en el cual la parte $ARIMA$ del modelo tiene la forma $(0,1,1)(0,0,0)$. Las consecuencias de una especificación incorrecta del modelo se discuten en Lytras et al. (2007).

Para una serie temporal mensual, la estructura del modelo RegARIMA es la siguiente:

$$(1-B)(y_t - \beta_1 M_{1,t} - \cdots - \beta_{11} M_{11,t} - \gamma X_t) = \mu + (1-B)a_t,$$

donde:

- $M_{j,t} = \begin{cases} 1 & \text{si el mes } j = 1, \ldots, 11, \\ -1 & \text{en diciembre,} \\ 0 & \text{de lo contrario.} \end{cases}$ son las variables ficticias.
- y_t es la serie temporal original.
- B es el operador retardo.
- X_t son otras variables de regresión utilizadas en el modelo (por ejemplo, valores atípicos, efectos del calendario, variables de regresión definidas por el usuario, variables de intervención).
- μ es el efecto medio.
- a_t es una variable de ruido blanco con media cero y varianza constante.

En el caso de una serie trimestral, el modelo estimado tiene la forma:

$$(1-B)(y_t - \beta_1 M_{1,t} - \cdots - \beta_3 M_{3,t} - \gamma X_t) = \mu + (1-B)a_t,$$

donde:

- $M_{j,t} = \begin{cases} 1 & \text{si el trimestre } j = 1, \ldots, 3, \\ -1 & \text{en el cuarto trimestre,} \\ 0 & \text{de lo contrario.} \end{cases}$ son las variables ficticias.

El estadístico del test chi-cuadrado es:

$$\hat{\chi}^2 = \hat{\beta}^T [\text{Var}(\hat{\beta})]^{-1} \hat{\beta},$$

Dado que la $\text{Var}(\hat{\beta})$, calculada utilizando la varianza estimada de a_t, puede ser muy diferente de la varianza real en muestras pequeñas, esta prueba se corrige utilizando el estadístico F que viene dado por:

$$F = \frac{\hat{\chi}^2_{s-1}}{n-d-k} \times \frac{n-d}{k},$$

donde:

- n es el tamaño de la muestra.
- d es el grado de diferenciación.
- s es la frecuencia de la serie temporal.
- k es el número total de regresores en el modelo RegARIMA (incluyendo las variables ficticias estacionales $M_{j,t}$ y el intercepto).

Este capítulo ha sentado las bases para trabajar con las series temporales, eliminando la estacionalidad y los efectos del calendario. Ahora que las series están ajustadas y libres de esos factores, podemos usarlas para ajustar modelos y analizar las relaciones entre las variables.

Parte II

Resultados

Capítulo 5

Análisis exploratorio de los datos

En este capítulo se realiza un estudio preliminar de los datos empleados en este trabajo. Las variables de interés, su origen y el procedimiento de obtención de las mismas se detallan en la Sección 5.1. En la Sección 5.2 se muestran los resultados de un análisis exploratorio de estas variables para comprender su comportamiento. En particular, se analiza qué variables requieren corrección de estacionalidad y las que presentan una tendencia significativa. Finalmente, en la Sección 5.3, se estudian las correlaciones existentes entre los ocupados y las afiliaciones, así como entre los parados y el paro registrado.

5.1. Variables consideradas para la modelización

Las variables empleadas son proporcionadas por entidades gubernamentales mediante formatos abiertos y de acceso público. En la Tabla 5.1, se presentan dichas entidades, junto con sus correspondientes abreviaturas, que se utilizarán de ahora en adelante, y los enlaces web donde pueden ser consultadas.

Organismos públicos proveedores de datos	**Abreviatura**	**Referencia**
Instituto Nacional de Estadística	INE	https://www.ine.es/
Ministerio de Economía, Comercio y Empresa	MINECO	https://portal.mineco.gob.es/

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

En la Tabla 5.2 se recogen las variables que se van a emplear en los próximos capítulos acompañadas por la fuente de la que provienen, su frecuencia (T trimestral y M mensual) y el período que comprenden los datos cuando se descargan:

Variable	Fuente	Frecuencia	Período
Ocupados	INE	T	2002-2024
Parados	INE	T	2002-2024
Tasa de paro	INE	T	2002-2024
Paro	MINECO	M	2002-Marzo 2025
Afiliaciones a la Seguridad Social	MINECO	M	2002-Marzo 2025

Tabla 5.2: Información principal de las variables empleadas.

5.1.1. Automatización del proceso de la descarga de datos

Para simplificar el acceso a estas variables y evitar la necesidad de ingresar a las páginas web cada vez que se deseen actualizar los datos, se ha implementado un proceso automatizado de descarga de la información utilizando el lenguaje de programación R. A continuación, se describe este procedimiento desarrollado para la descarga de los datos en función de la fuente de origen.

INE

Los datos relacionados con la EPA provienen del INE. Este organismo ofrece un servicio de acceso a sus bases de datos mediante su API en formato JSON, lo que permite realizar consultas directas a través de solicitudes URL.

El formato estándar para realizar estas peticiones está estructurado de la siguiente manera:

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

donde:

- **id_tabla** corresponde al código identificador único de cada tabla dentro de la base de datos del INE.
- **nult** especifica la cantidad de períodos de datos que se desean obtener.
- **date** define el rango de fechas para la consulta de los datos, con el formato AAAAMMDD:AAAAMMDD.

Para la descarga específica de los datos de la EPA, la URL solicitada debe ajustarse a los parámetros correspondientes según la tabla y el rango de fechas deseado. Por ejemplo, en el caso de la descarga de los ocupados, la URL correspondiente es

https://servicios.ine.es/wstempus/js/ES/DATOS_TABLA/65109?nult=1000

El formato en que se presentan los datos recuperados a través de las URLs es de tipo JSON (JavaScript Object Notation). Para automatizar la descarga de las series de índices solicitadas, simplemente se puede utilizar la función `fromJSON` de la librería `jsonlite` (Ooms (2014)) en R. Esta función permite leer los datos obtenidos desde la URL y acceder a las series almacenadas dentro del conjunto de datos descargado.

Veamos un ejemplo de este proceso, con la descarga de los ocupados:

```
datos2<-jsonlite::fromJSON("https://servicios.ine.es/wstempus/js/ES/DATOS_TABLA/
65109?nult=1000")

ocupados_ts<-ts(rev(datos2$Data[[1]]$Valor), start = c(rev(datos2$Data[[1]]$Anyo)
[1],datos2$Data[[1]]$FK_Periodo[1]%%6), freq = length(unique(datos2$Data[[1]]$FK_
Periodo)))
```

Este procedimiento de acceso a la base de datos del INE desde R mediante solicitudes URL es aplicable a cualquier serie estadística disponible en su portal, lo que permite a los analistas de ABANCA automatizar la descarga de todos los datos necesarios para el seguimiento macroeconómico.

MINECO

En relación con los datos publicados por el Ministerio de Economía, Comercio y Empresa, existen dos métodos principales para acceder a ellos. El primero consiste en navegar por su base de datos disponible en el siguiente enlace:

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

que permite seleccionar manualmente las series deseadas y proceder con su descarga. El segundo método implica la descarga de un archivo comprimido (**.zip**) que contiene múltiples archivos en formato **.csv**, los cuales incluyen todas las series disponibles.

Para automatizar este proceso, se opta por la segunda opción. Utilizando la función `download.file` de la librería `utils` (R Core Team (2023)) es posible descargar el archivo comprimido. Posteriormente, se emplea la función `unzip` de la misma librería para descomprimir el archivo y extraer las series necesarias.

En este caso, las series de interés incluyen el paro registrado y las afiliaciones a la Seguridad Social. A continuación, se presenta el código utilizado para la descarga de todas ellas y lectura de la primera.

```
# Descarga del archivo zip
download.file(url = "https://portal.mineco.gob.es/economiayempresa/
EconomiaInformesMacro/Documents/bdsicecsv.zip",
destfile = paste0("C:\\Users\\u033810\\Desktop", "\\mineco.zip"), mode = "wb")

# Descomprimimos solo las series que necesitamos
archivos <- c("170000n.csv", "190000m.csv")
unzip(zipfile = paste0("C:\\Users\\u033810\\Desktop", "\\mineco.zip"),
files = archivos, overwrite = TRUE, exdir = "dir")

# Paro registrado
paro <- fread(paste0("C:\\Users\\u033810\\Desktop\\datos_mineco\\", archivos[1]),
header = TRUE, sep = ";", dec = ",", fill = TRUE, encoding = "UTF-8")[, 1:3]
paro_ts <- ts(paro$Observaciones, start = c(min(paro$Año), min(paro$Mes)), freq =
12)
```

5.2. Análisis de las variables

Series de la EPA

Tras la descarga de los datos, se procede a realizar un análisis exploratorio de los mismos. En primer lugar se construyen gráficos de líneas de las series de los ocupados, parados y tasa de paro, las cuales se muestran en la Figura 5.1 de izquierda a derecha, respectivamente. Esta representación gráfica permite estudiar la evolución del empleo a lo largo del intervalo de tiempo que comprende desde el primer trimestre de 2002 hasta el cuarto trimestre de 2024.

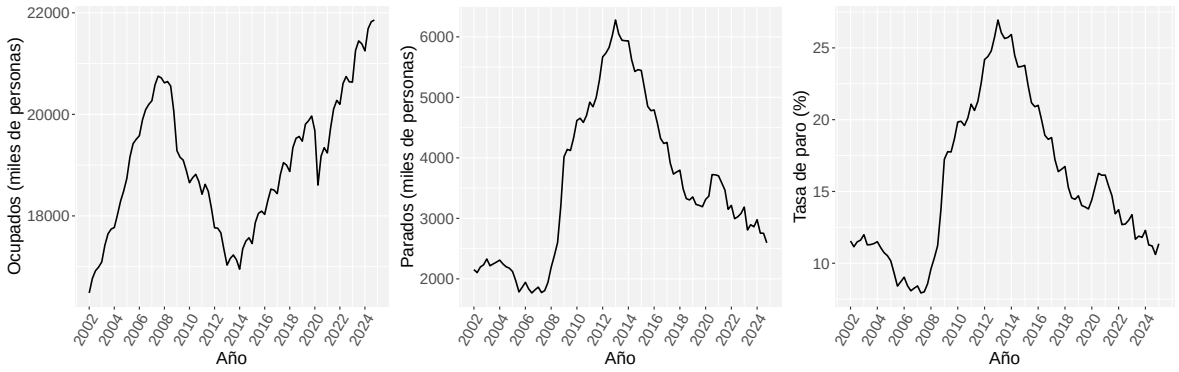

Figura 5.1: Series de ocupados, parados y tasa de paro de la EPA.

A la vista de la Figura 5.1, parece claro que las tres series presentan tendencias cambiantes en intervalos específicos de tiempo. No obstante, este aspecto se analizará en detalle en la Sección 5.2.2. En estas series también se observa una componente estacional, lo que quizá se hace más claro en la Figura 5.2, donde se muestra la evolución por trimestres de los ocupados desde 2002 hasta 2024.

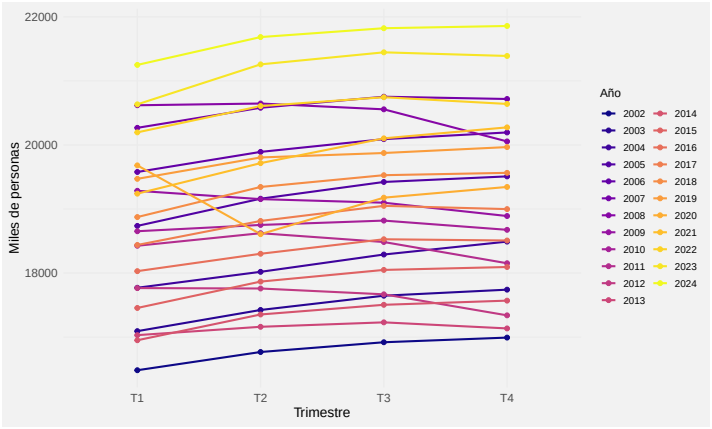

Figura 5.2: Evolución trimestral de los ocupados.

Observando la Figura 5.2 se puede ver que en el tercer trimestre de todos los años, el número de ocupados tiende a aumentar en comparación con el resto de trimestres lo que se refleja en un incremento correspondiente de los parados. Hay que destacar también que en el año 2020 se refleja una bajada en

el segundo trimestre, coincidiendo con el el confinamiento por la pandemia del COVID-19.

Para ampliar el análisis de estas variables, en la Figura 5.3 se representan la variación interanual de los ocupados (izquierda) y de los parados (derecha). Este indicador muestra cómo cambia el valor de una serie en un mes determinado respecto al mismo mes del año anterior, lo que permite identificar ciertos cambios relevantes en el trimestre.

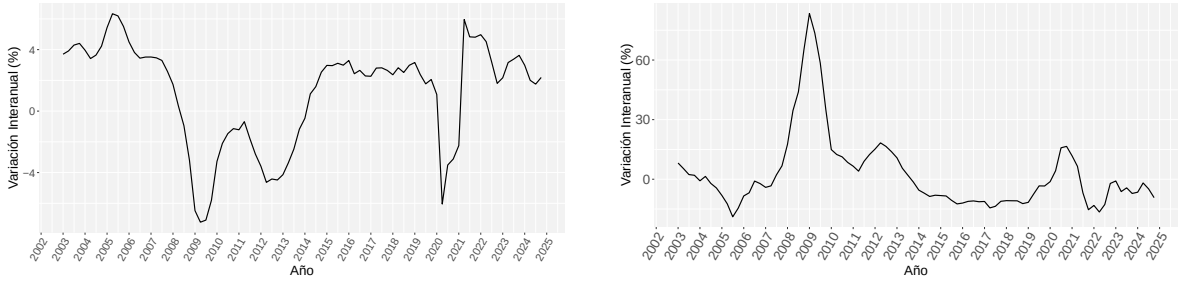

Figura 5.3: Variación interanual de los ocupados (izquierda) y de los parados (derecha).

En la serie de los ocupados, se observa un notable descenso en la variación interanual durante la crisis de 2008-2009, seguido de fuerte incremento. Posteriormente, se registra una recuperación entre 2013 y 2019, antes de experimentar otra gran disminución en 2020, coincidiendo con la pandemia de COVID-19. A partir de entonces, se aprecia un repunte significativo, seguido por cierta estabilización con fluctuaciones. Este comportamiento indica que la serie de los ocupados no es estacionaria, ya que parece que presenta tanto tendencia como heterocedasticidad.

En cuanto a la serie de los parados, también muestra un comportamiento no estacionario, con variaciones abruptas a lo largo del tiempo. Se destacan un fuerte aumento en la variación interanual alrededor de 2009-2010 y varios repuntes posteriores, como en 2021, junto con periodos de relativa estabilidad. Estas fluctuaciones, junto con la ausencia de una varianza constante y la posible presencia de tendencias, sugieren que esta serie tampoco es estacionaria.

No obstante, se chequeará más adelante mediante pruebas estadísticas precisas si ambas series presentan estacionalidad y tendencia.

Series paro registrado y afiliaciones a la Seguridad Social

Se continúa el análisis exploratorio con las series descargadas del Ministerio de Economía, Comercio y Empresa (MINECO). La Figura 5.4 proporciona una visualización inicial de la evolución de estas series, que comprenden los periodos de enero de 2002 hasta marzo de 2025 tanto para el paro registrado como para las afiliaciones a la Seguridad Social.

A partir de la Figura 5.4 se identifica la presencia de características propias de series no estacionarias, como tendencia, heterocedasticidad y estacionalidad.

La serie de paro registrado muestra un nivel bajo hasta 2008, momento en el que inicia una tendencia creciente, coincidiendo con la crisis económica. A partir de 2014, se observa un descenso en el paro con la recuperación de la crisis, interrumpido por un nuevo incremento en 2020 debido a la pandemia de COVID-19. Por su parte, la serie de afiliaciones a la Seguridad Social experimenta un rápido crecimiento en los primeros años de la serie, seguido de una caída significativa durante la crisis económica. Posteriormente, se recupera mostrando una tendencia al alza hasta 2020, cuando se produce un descenso asociado a la pandemia. Los valores observados durante la pandemia podrían considerarse

atípicos.

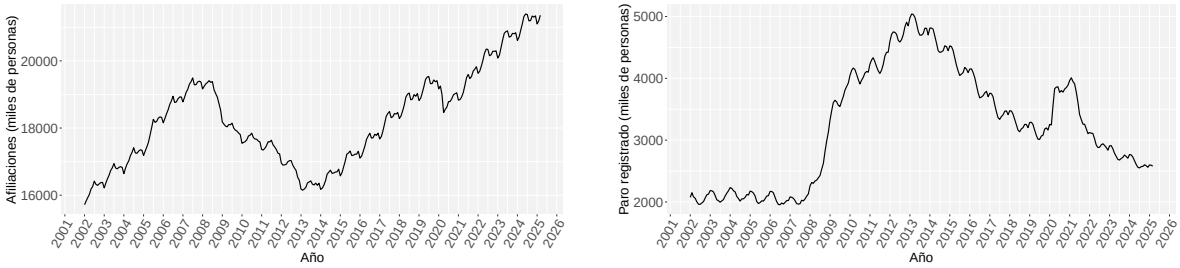

Figura 5.4: Evolución de las series de afiliaciones (izquierda) y paro registrado (derecha).

5.2.1. Estacionalidad

La estacionalidad pronunciada puede generar distorsiones al intentar ajustar un modelo estadístico. Por esta razón, se procede en primer lugar a chequear si las series presentan o no estacionalidad. Para ello, se utiliza la librería `seastest` (Ollech (2021)), la cual incluye las pruebas estadísticas que se emplearán: el QS-Test (Maravall (2011)), el test de Friedman (Friedman (1937)), el de Kruskal-Wallis (Kruskal and Wallis (1952)) y el de Welch (Welch (1951)).

Todas las pruebas mencionadas arrojaron p-valores iguales a cero para las series de ocupados, parados, tasa de paro, afiliaciones y paro registrado. Por lo tanto, estas series presentan estacionalidad.

5.2.2. Tendencia

Otra cuestión relevante en el análisis de series temporales es la posible existencia de una tendencia. Para chequearla, se puede aplicar el test de Mann-Kendall, disponible en R a través de la librería `Kendall` (McLeod (2022)), mediante la función `MannKendall`.

En la Tabla 5.3 se presentan los p-valores del test de Mann-Kendall, el valor del estadístico correspondiente (denotado por τ) y la conclusión sobre la existencia o no de tendencia en las series analizadas. En caso afirmativo, se indica también el tipo de tendencia detectada.

Series	**P-valor**	$\boldsymbol{\tau}$	**Tendencia**
Ocupados	0.000	0.366	Sí, tendencia creciente
Parados	0.154	0.101	No
Tasa de Paro	0.443	0.054	No
Afiliaciones	0.000	0.469	Sí, tendencia creciente
Paro Registrado	0.000	0.147	Sí, tendencia creciente

Tabla 5.3: Información sobre el test de Mann-Kendall.

Conviene subrayar que el test de Mann-Kendall chequea la existencia de una tendencia creciente o

decreciente sobre todo el rango temporal examinado. Así, la significación encontrada en el sentido de tendencia creciente para las series de ocupados y afiliaciones es del todo compatible con los perfiles de los gráficos de líneas de ambas series.

Análogamente, es razonable que la prueba no resulte significativa para las series de parados y tasa de paro porque no se aprecia en los perfiles de estas series una tendencia monótona a lo largo de todo el rango temporal. Sin embargo, para estas series sí existen tendencias de sentido contrario bien definidas en subintervalos del rango de tiempo. En definitiva, la prueba de Mann-Kendall no es apropiada para detectar "cambios de tendencia".

5.3. Correlación entre las variables

En esta última sección del capítulo se analiza la relación existente entre dos pares de series temporales: por un lado, la serie de ocupados de la EPA y las afiliaciones a la Seguridad Social; y por otro, la serie de parados de la EPA y el paro registrado.

El objetivo de este análisis es evaluar en qué medida las afiliaciones y el paro registrado pueden ser buenos indicadores para desagregar las series de ocupados y parados, respectivamente. A priori, se ha optado por estos dos indicadores debido a su coherencia con la lógica económica y a la recomendación de los analistas del entorno macroeconómico.

Gráficamente, las series de cada uno de los pares mencionados, las series de ocupados y afiliaciones, por un lado, y las de parados y paro registrado, por otro, muestran una evolución similar, como puede observarse en la Figura 5.5, donde se representan dichas variables respectivamente. Además, en ambas gráficas se aprecia que, durante el periodo correspondiente a la pandemia del COVID-19, el número de afiliaciones y el paro registrado superan al de ocupados y parados, respectivamente. Este hecho supone un problema que será analizado en detalle más adelante.

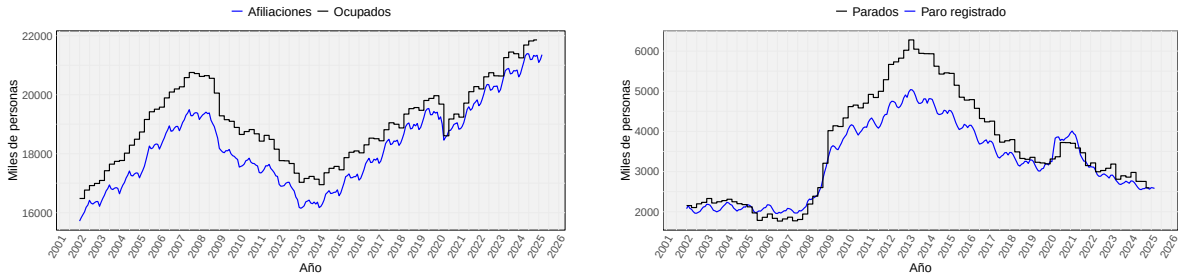

Figura 5.5: Comparativa entre el número de afiliaciones a la Seguridad Social y el de ocupados (izquierda) y el número de paro registrado y de parados (derecha).

A modo de observación, en las Figura 5.5, la series de ocupados y parados se presentan asumiendo que todos los meses tienen el mismo valor que el trimestre al que hacen referencia, mientras que las series de las afiliaciones y del paro registrado tienen una frecuencia mensual.

Para complementar la explicación anterior, se representan también en la Figura 5.6 las variaciones interanuales de estas series. En el lado izquierdo las relativas a los números de ocupados y de afiliaciones, y en el derecho las de parados y del paro registrado. En ambas se observa que las variaciones de las variables de afiliaciones y ocupados siguen un patrón similar, al igual que las de parados y paro registrado.

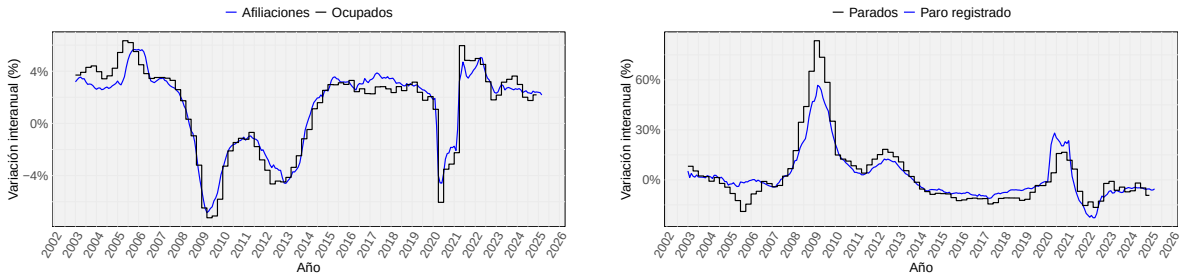

Figura 5.6: Variación interanual de los ocupados y afiliaciones (izquierda) y de los parados y paro registrado (derecha).

A continuación, se va a comprobar estadísticamente la relación entre ambas series mediante las correlaciones cruzadas.

Dado que las series de la EPA son de frecuencia trimestral, mientras que las de afiliaciones y paro registrado son mensuales, es necesario homogeneizar sus frecuencias. Para ello, se han transformado las series mensuales en series trimestrales utilizando la media de los valores mensuales de cada trimestre.

No obstante, estas series presentan tendencia y estacionalidad, lo que implica que no son estacionarias. Trabajar con series no estacionarias puede dar lugar a correlaciones espurias, por lo que es imprescindible transformarlas previamente en series estacionarias. Este proceso se lleva a cabo aplicando, en primer lugar, una diferenciación regular y, a continuación, una diferenciación estacional a cada una de las series (ocupados, parados, afiliaciones y paro registrado, todas ya transformadas a frecuencia trimestral).

Una vez diferenciadas, se verifica la estacionariedad mediante el test de Dickey-Fuller aumentado (Dickey and Fuller (1979)). En todos los casos, se obtiene un p-valor de 0.01, por lo que se rechaza la hipótesis nula de no estacionariedad y se concluye que las series ya pueden considerarse estacionarias.

A continuación, se aplica el método de preblanqueo, descrito en la Sección 2.3.1, con el fin de eliminar la autocorrelación de las series y analizar correctamente la relación entre ellas mediante correlaciones cruzadas. Este análisis permite identificar el retardo temporal en el que se produce la mayor correlación entre las series.

En la Figura 5.7 se muestra el gráfico de correlaciones cruzadas entre ocupados y afiliaciones. Se observa que la mayor correlación se produce con un retardo de cero trimestres, lo que indica una fuerte correlación contemporánea positiva: cuando aumentan las afiliaciones, lo hacen también los ocupados en el mismo trimestre.

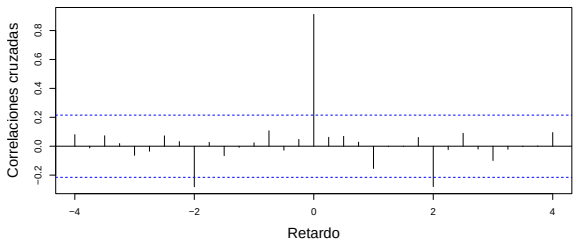

Figura 5.7: Correlaciones cruzadas entre ocupados y afiliaciones después del preblanqueo.

Por otro lado, la Figura 5.8 representa la correlación cruzada entre parados y paro registrado. En este caso, la mayor correlación se observa con un retardo de un trimestre, lo que sugiere que el paro registrado anticipa parcialmente los movimientos de la serie de parados.

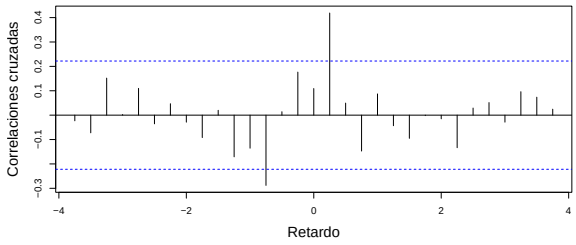

Figura 5.8: Correlaciones cruzadas entre parados y paro registrado después del preblanqueo.

En conclusión, los resultados muestran una relación significativa entre las series de afiliaciones y paro registrado y las series de la EPA. Estas observaciones son relevantes y se tendrán en cuenta en el Capítulo 6, donde se abordará la desagregación temporal de las series de la EPA a partir de las series mensuales de ocupados y paro registrado.

Capítulo 6

Desagregación

El objetivo de este capítulo es desagregar las series temporales de ocupados y parados, que tienen una frecuencia trimestral, para transformarlas en series mensuales a través de variables de más alta frecuencia. Además, se discutirán las dificultades encontradas al utilizar los distintos métodos de desagregación y cómo se han solucionado. Posteriormente, se procederá a corregirlas de estacionalidad y efecto calendario, proceso que se explicará en el Capítulo 7.

El capítulo se divide en dos secciones en las que se abordan la desagregación de la serie de ocupados (Sección 6.1) y la de la serie parados (Sección 6.2). Los detalles de la función utilizada para la desagregación temporal se encuentran disponibles en el Apéndice A.

6.1. Desagregación de la serie de ocupados

En primer lugar, se lleva a cabo la desagregación temporal de la serie de ocupados, utilizando como indicador el número de afiliaciones a la Seguridad Social. Este indicador ha demostrado ser adecuado para la desagregación de series temporales, tal como se ha justificado previamente en la Sección 5.3. Para realizar dicha desagregación, se aplicarán los tres procedimientos presentados en la Sección 3.2, a saber los métodos de Chow-Lin, Fernández y Litterman, los cuales permiten desagregar series trimestrales, como la de los ocupados, a partir de series mensuales, como la de las afiliaciones a la Seguridad Social.

Los tres procedimientos de desagregación se basan en un ajuste lineal entre la variable a desagregar y la serie indicadora, con errores dependientes siguiendo estructuras paramétricas específicas según el método. Por consiguiente, es fundamental la estimación de los parámetros que determinan la regresión y la estructura de dependencia, así como chequear que son significativamente no nulos. A tal efecto, se obtendrá el correspondiente estadístico de prueba que, bajo la nula, es aproximadamente normal, de modo que se examina si en valor absoluto supera 1.96, el umbral crítico para un nivel de significación del 5 %. Si así fuese, se rechaza la hipótesis nula de que el parámetro es igual a cero, y por tanto se considera estadísticamente significativo.

Tras proceder de esta forma y obtenidas las series desagregadas, es necesario chequear que se cumplen las hipótesis asociadas a los residuos del ajuste lineal, en función del método de desagregación utilizado. En particular, para que el método aplicado sea válido, los residuos del modelo desagregado deben seguir determinadas estructuras estocásticas: en el caso del método de Chow-Lin, deben ajustarse a un proceso $AR(1)$; para el método de Fernández, a un proceso $ARIMA(0,1,0)$; y, en el caso del

método de Litterman, a un proceso $ARIMA(1,1,0)$. Realizar esta comprobación supone corroborar la significación de los coeficientes del ARIMA correspondiente y chequear que los residuos del ajuste de estas estructuras son ruido blanco.

Para comprobar la media nula se emplea el test de la t de Student (Student (1908)), para la independencia se utiliza el test de Ljung-Box (Ljung and Box (1978)), para la homocedasticidad se aplica el test de Ljung-Box (Ljung and Box (1978)) sobre los residuos y para comprobar la normalidad se emplea el test de Doornik-Hansen (Doornik and Hansen (2008)). En la Tabla 6.1 se muestran los p-valores de los contrastes aplicados a los residuos de los modelos estocásticos ajustados sobre los residuos de la serie desagregada de ocupados.

Contraste	**Chow-Lin**	**Fernández**	**Litterman**
Media nula	0.635	0.940	0.940
Independencia	0.000	0.000	0.000
Homocedasticidad	0.637	0.017	0.017
Normalidad	0.003	0.000	0.000

Tabla 6.1: P-valores de los contrastes aplicados a los residuos de los modelos estocásticos ajustados ($AR(1)$, $ARIMA(0,1,0)$ y $ARIMA(1,1,0)$) sobre los residuos de la serie desagregada de ocupados.

De la Tabla 6.1 se concluye que en ningún caso se cumple la hipótesis de independencia de los residuos de los modelos ajustados. Por lo tanto, ninguno de los tres métodos de desagregación es válido porque no se satisfacen las hipótesis estructurales requeridas, como se explica en la Sección 2.4.3.

A raiz de los análisis de las series y la lógica económica, se puede pensar que el hecho de que no haya un buen ajuste con ninguno de los métodos podría verse influenciado por el atípico causado por la pandemia de COVID-19. Para analizar este fenómeno de manera más detallada, se decide dividir las series de ocupados y afiliaciones en dos períodos: uno antes del COVID y otro posterior a su aparición. Ambas series se recortan hasta el primer trimestre de 2020 y en marzo de 2020, respectivamente.

Entonces, se establece un segundo período, post-COVID, que comienza en el tercer trimestre de 2020 para la serie de ocupados y en julio de 2020 para la serie de afiliaciones. A continuación, se aplican nuevamente los métodos de desagregación temporal, utilizando el método de Chow-Lin como ejemplo para ambos períodos y siguiendo el mismo camino que en los análisis previos. En la Tabla 6.2 se recogen los p-valores de los contrastes aplicados a los residuos del modelo estocástico ajustado $AR(1)$ sobre los residuos de la serie desagregada de ocupados antes y después del COVID.

Al observar la Tabla 6.2, tanto antes como después del COVID, se puede concluir que los residuos en ambos casos pasan sin problema los tests de diagnosis. Los residuos del ajuste cumplen las tres principales hipótesis, además de la normalidad: media nula, independencia y aleatoriedad. En consecuencia, los residuos se comportan como ruido blanco, lo que respalda la adecuación del modelo y del método de desagregación. Realizando el mismo análisis con los otros dos métodos restantes (Fernández y Litterman), también se cumplen las hipótesis.

Contraste	**Pre-COVID**	**Post-COVID**
Media nula	0.508	0.9934
Independencia	0.118	0.249
Homocedasticidad	0.374	0.166
Normalidad	0.350	0.587

Tabla 6.2: P-valores de los contrastes aplicados a los residuos de los modelos estocásticos ajustados $AR(1)$ sobre los residuos de la serie desagregada de ocupados antes y después del COVID.

En vista de estos resultados, se puede inferir que la relación entre las afiliaciones y los ocupados durante el período del COVID presenta ciertas irregularidades. Para abordar este problema, surge la idea de corregir este efecto empleando los Expedientes de Regulación Temporal de Empleo (ERTE).

Antes de proceder con el análisis, es importante explicar en qué consiste esta medida y cómo afectó a los trabajadores ocupados en ese periodo, lo que se puede consultar en Luís Jiménez Arellano (2020) y Diario de Navarra (2018).

En respuesta a los efectos de la pandemia de COVID-19, una de las medidas adoptadas en España fue el uso de los ERTEs, regulado por el artículo 47 del Estatuto de los Trabajadores. Este mecanismo tiene como objetivo principal permitir que las empresas en dificultades temporales puedan suspender los contratos de trabajo o reducir las jornadas laborales de sus empleados sin llegar a un despido definitivo.

A continuación, se exponen las principales características de los ERTEs en España:

- **No conlleva despidos permanentes:** La principal ventaja de los ERTEs es que no implican la pérdida definitiva del puesto de trabajo. Al finalizar el período de suspensión o reducción, los empleados recuperan sus puestos en la empresa.
- **Flexibilidad en su aplicación:** A diferencia del ERE (Expediente de Regulación de Empleo), que se aplica generalmente a empresas más grandes o en situaciones específicas, el ERTE puede implementarse en cualquier tipo de empresa, independientemente de su tamaño o del número de empleados afectados.
- **Uso de la prestación por desempleo:** Durante la vigencia del ERTE, los empleados pueden hacer uso de su derecho a la prestación por desempleo, lo que afecta al total de días que pueden cobrar esta ayuda en el futuro, si en algún momento pierden definitivamente su empleo.
- **Ayuda económica parcial:** Durante los primeros 180 días de aplicación del ERTE, los empleados pueden percibir hasta el 70 % de su salario habitual, aunque, en muchos casos, las empresas negocian un complemento para reducir la pérdida salarial, si bien no cubre completamente el sueldo.

El ERTE es una solución temporal diseñada para proteger tanto a las empresas como a los empleados, permitiendo que, tras la crisis, las empresas puedan recuperar su actividad sin que los trabajadores pierdan sus empleos de manera definitiva.

Durante la pandemia, muchos trabajadores en ERTE fueron clasificados como ocupados en la EPA debido a que mantenían un vínculo formal con su empleo. Sin embargo, en los casos de ERTE de

suspensión, estos trabajadores no realizaban ninguna hora de trabajo, lo que generó una distorsión en las estadísticas. Por ello, para obtener una imagen más fiel del empleo efectivo, los analistas consideraban necesario restar a estos trabajadores de la cifra de ocupados, ya que no estaban contribuyendo activamente a la producción ni al mercado laboral en ese periodo.

Por esta razón, para la desagregación de los ocupados, se debe restar de los datos de los ocupados inicialmente registrados la cantidad correspondiente a los ERTE en los años 2020 y 2021. De este modo, se consigue lo que serían los ocupados efectivos. Estos datos se han obtenido de la página del Ministerio de Inclusión, Seguridad Social y Migraciones.

Con esta corrección y aplicando los tres métodos de desagregación empleados anteriormente, se vuelven a analizar los resultados y además se comprueba que todos los coeficientes del modelo sean estadísticamente significativos. En la Tabla 6.3 se presentan los p-valores de los contrastes realizados sobre los residuos de los modelos estocásticos ajustados $AR(1)$, $ARIMA(0,1,0)$ y $ARIMA(1,1,0)$ sobre los residuos de la serie desagregada de los ocupados efectivos.

Contraste	**Chow-Lin**	**Fernández**	**Litterman**
Normalidad	0.000	0.000	0.000
Media nula	0.965	0.672	0.672
Independencia	0.513	0.000	0.000
Homocedasticidad	0.873	0.001	0.001

Tabla 6.3: P-valores de los contrastes aplicados a los residuos de los modelos estocásticos ajustados ($AR(1)$, $ARIMA(0,1,0)$ y $ARIMA(1,1,0)$) sobre los residuos de la serie desagregada de ocupados efectivos.

Según los p-valores de la Tabla 6.3, se observa que el único método que cumple la diagnosis de los residuos, a excepción de la normalidad, es el método de Chow-Lin. Además, comparando estos resultados junto con los anteriores, se constata que los diagnósticos de los residuos para los métodos de Fernández y Litterman coinciden. De hecho, se obtiene que el valor estimado de ρ es 0, lo que confirma que ambos métodos coinciden, tal como se indicó en la Sección 3.2.3.

De toda la argumentación previa cabe concluir que el método adecuado para realizar la desagregación temporal de la serie de ocupados es el método de Chow-Lin. Específicamente:

- La serie de baja frecuencia, correspondiente a los ocupados anuales, se denota por $\mathbf{y}_l \in \mathbb{R}^{T\times 1}$, donde $T = 92$ trimestres correspondientes al periodo 2002-2024.
- La matriz $\mathbf{X}_h \in \mathbb{R}^{n\times k}$ representando las series indicadoras de alta frecuencia solamente incluye en este caso a la serie mensual de afiliaciones a la Seguridad Social (así, $k = 1$) y, como el número de subperíodos es $s = 3$, se tiene $n = T \times s + 3 = 92 \cdot 3 + 3 = 279$.

Para la estimación, se ha asumido un modelo autorregresivo de primer orden ($AR(1)$) para los residuos del modelo, según la ecuación (3.7), donde los parámetros estimados resultan:

$$\hat{\rho} = 0.760, \quad \hat{\sigma}_\epsilon^2 = 1.077.$$

Con estos valores se construye la matriz de covarianzas de alta frecuencia $\mathbf{V}_h$, cuya forma está dada por la ecuación (3.8).

Posteriormente, se estima el vector de coeficientes $\hat{\boldsymbol{\beta}}$ a través de la expresión:

$$\hat{\boldsymbol{\beta}} = \left[(\mathbf{C}\mathbf{X}_h)^\intercal(\mathbf{C}\mathbf{V}_h\mathbf{C}^\intercal)^{-1}(\mathbf{C}\mathbf{X}_h)\right]^{-1}(\mathbf{C}\mathbf{X}_h)^\intercal(\mathbf{C}\mathbf{V}_h\mathbf{C}^\intercal)^{-1}\mathbf{y}_l,$$

donde:

- $\mathbf{C}$ es la matriz de agregación temporal definida en la ecuación (3.3),
- $\mathbf{X}_h$ es la matriz de la variable indicadora (afiliaciones),
- $\mathbf{y}_l$ es el vector de ocupados en baja frecuencia,
- $\mathbf{V}_h$ es la matriz de covarianzas definida anteriormente.

Cabe mencionar que la estimación de $\hat{\boldsymbol{\beta}}$, aunque se obtiene, no se muestra debido a sus dimensiones.

Finalmente, con todos estos elementos, se obtiene la serie desagregada de ocupados en alta frecuencia, $\hat{\mathbf{y}}_h$, mediante la siguiente expresión:

$$\hat{\mathbf{y}}_h = \mathbf{X}_h\hat{\boldsymbol{\beta}} + \mathbf{V}_h\mathbf{C}^\intercal\left(\mathbf{C}\mathbf{V}_h\mathbf{C}^\intercal\right)^{-1}\left(\mathbf{y}_l - \mathbf{C}\mathbf{X}_h\hat{\boldsymbol{\beta}}\right).$$

Todos los elementos que intervienen en esta última fórmula han sido definidos previamente, y su combinación permite obtener la desagregación deseada de la serie de ocupados.

Finalmente, se representa gráficamente en la Figura 6.1 la serie desagregada de los ocupados por el método de Chow-Lin junto con la serie original de ocupados. Se ve que la desagregación es suave y el pico que se observa durante los años 2020 y 2021 se debe al restar los ertes a los ocupados durante esta época.

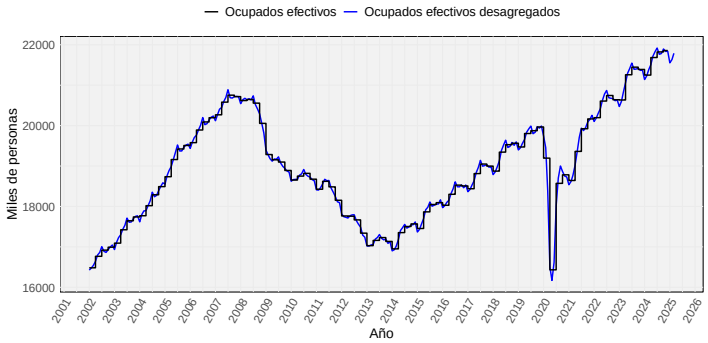

Figura 6.1: Representación gráfica de la serie de ocupados efectivos teniendo en cuenta los ERTEs y su serie desagregada.

Todo lo que se ha realizado para los ocupados, explicándolo con detalle, se ha replicado para los parados, aunque se presentará de manera más breve, ya que el procedimiento es análogo.

6.2. Desagregación de la serie de parados

En esta sección se explica la desagregación temporal de la serie de parados utilizando el indicador de paro registrado, que, como se ha mostrado en la Sección 5.3, se considera un buen indicador para llevar a cabo la desagregación mediante métodos como Chow-Lin, Fernández o Litterman.

A continuación, se han aplicado los tres métodos mencionados de desagregación y generado las series desagregadas con cada uno de ellos. Como en la sección previa se procedió a chequear las hipótesis estructurales sobre los residuos resultantes de cada uno de los tres ajustes, resultando los p-valores que se proporcionan en la Tabla 6.4.

Contraste	**Chow-Lin**	**Fernández**	**Litterman**
Media nula	0.330	0.987	0.987
Independencia	0.000	0.000	0.000
Homocedasticidad	0.005	0.012	0.012
Normalidad	0.000	0.000	0.000

Tabla 6.4: P-valores de los contrastes aplicados a los residuos de los modelos estocásticos ajustados ($AR(1)$, $ARIMA(0,1,0)$ y $ARIMA(1,1,0)$) sobre los residuos de la serie desagregada de parados.

Al igual que en el caso de los ocupados, ninguno de los tres ajustes conduce a residuos independientes, lo que indica que ninguno de los tres procedimientos es aceptable en este caso.

Como antes, se realiza un análisis separado para el periodo del COVID, utilizando las mismas fechas pre y post COVID, tanto para la serie de parados como el paro registrado. Con esta nueva división, se lleva a cabo la desagregación temporal, utilizando el método de Chow-Lin como ejemplo para ambas series, siguiendo el mismo enfoque utilizado en los análisis previos. Los p-valores de los contrastes sobre los residuos de los ajustes $AR(1)$ del error de la serie desagregada de parados antes y después del COVID se presentan en la Tabla 6.5.

Contraste	**Pre-COVID**	**Post-COVID**
Media nula	0.879	1.000
Independencia	0.476	0.059
Homocedasticidad	0.833	0.306
Normalidad	0.180	0.697

Tabla 6.5: P-valores de los contrastes aplicados a los residuos de los modelos estocásticos ajustados $AR(1)$ sobre los residuos de la serie desagregada de parados antes y después del COVID.

Al observar los resultados de la Tabla 6.5, tanto antes como después del COVID, se puede concluir que los residuos superan los tests. Por lo tanto, volvemos a considerar la serie de los ERTEs que se ha utilizado en el caso de los ocupados. En este caso, los trabajadores en ERTE total aunque técnicamente no estaban desempleados, compartían muchas características con los parados: no trabajaban, dependían de una prestación pública y su reincorporación era incierta. Por esta razón, sumarlos al número de parados ofrecía una mejor aproximación al impacto real de la crisis en el empleo.

Tras añadir los ERTEs en los años 2020 y 2021 a la serie de parados, se procede a desagregarla con los tres métodos y chequear las hipótesis estructurales de los residuos resultantes. Los p-valores obtenidos se proporcionan en la Tabla 6.6.

Contraste	**Chow-Lin**	**Fernández**	**Litterman**
Media nula	0.969	0.887	0.887
Independencia	0.966	0.040	0.040
Homocedasticidad	0.153	0.046	0.046
Normalidad	0.000	0.000	0.000

Tabla 6.6: P-valores de los contrastes aplicados a los residuos de los modelos estocásticos ajustados ($AR(1)$, $ARIMA(0,1,0)$ y $ARIMA(1,1,0)$) sobre los residuos de la serie desagregada de parados efectivos.

De la Tabla 6.6 se concluye que los residuos de los modelos estocásticos asociados a los métodos de Fernández y Litterman no superan la validación del modelo, al no cumplir con los contrastes de independencia y homocedasticidad. En cambio, los residuos del modelo estocástico asociado al método de Chow-Lin sí superan dichas pruebas, lo que lo convierte en la opción más adecuada para llevar a cabo la desagregación de la serie de parados, utilizando como variable indicadora el paro registrado.

De acuerdo con el análisis previo, se adopta la solución obtenida con Chow-Lin para desagregar la serie de parados tomando como indicadora el paro registrado. Denotando a la primera como $\mathbf{y}_l \in \mathbb{R}^{T\times 1}$, donde $T = 92$ corresponde al número de trimestres comprendidos entre los años 2002 y 2024.

Como la variable indicadora de alta frecuencia es el paro registrado, que se organiza en la matriz $\mathbf{X}_h \in \mathbb{R}^{n\times k}$. En este caso, el número total de observaciones es $n = (92 \cdot 3) + 3 = 279$, siendo $s = 3$ el número de subperiodos mensuales por trimestre y $k = 1$ el número de variables indicadoras.

Para modelar la dinámica de los residuos del modelo, se adopta un proceso autorregresivo de primer orden $AR(1)$, definido por la ecuación (3.7), donde los parámetros estimados son:

$$\hat{\rho} = 0.705, \quad \hat{\sigma}_\epsilon^2 = 1.408.$$

Se representa gráficamente en la Figura 6.2 la serie desagregada de los parados utilizando el método de Chow-Lin, junto con la serie de parados teniendo en cuenta los ERTEs. En ella, se puede observar que la desagregación es suave, y el pico más elevado en comparación con la serie original se debe a la suma de los ERTEs, como se comentó anteriormente.

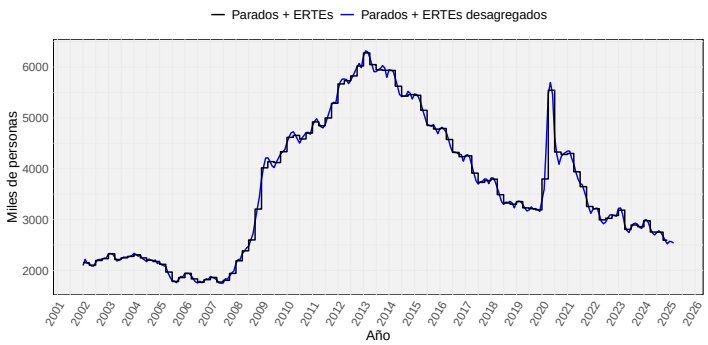

Figura 6.2: Representación gráfica de la serie de parados teniendo en cuenta los ERTEs y su serie desagregada.

Una vez desagregadas las series de ocupados y parados, se deben corregir de efectos estacionales y de calendario. Este paso es importante para garantizar que las series reflejen correctamente las variaciones estructurales y no aquellas causadas por factores estacionales o los días laborales. En el siguiente capítulo se explica el procedimiento práctico utilizado para llevar a cabo estas correcciones.

Capítulo 7

Corrección de estacionalidad y calendario

En este capítulo se aplica todo el proceso de corrección y desestacionalización de series económicas explicado en el Capítulo 4. El objetivo es obtener series ajustadas que reflejen las fluctuaciones económicas reales, eliminando las distorsiones introducidas por los efectos de calendario y la estacionalidad. De esta manera, se busca facilitar la comparación entre diferentes períodos, permitiendo un análisis más preciso.

El capítulo se divide en tres secciones. En la Sección 7.1 se detallan los modelos empleados para corregir de estacionalidad y efectos de calendario las series de ocupados, parados y sus desagregaciones obtenidas en el Capítulo 6. Para ello, antes de nada se analiza cuáles de dichos efectos son significativos en cada caso. Posteriormente, en la Sección 7.2 se explica la validación de los modelos $ARIMA$ ajustados en la Sección 7.1.

Por último, en la Sección 7.3 se pone en práctica la metodología de corrección sobre una batería de indicadores particularmente relevantes para ABANCA en términos de análisis del mercado laboral. En concreto, se abordan las afiliaciones a la Seguridad Social organizadas por sectores de actividad económica, siguiendo la clasificación establecida por la CNAE-2009, lo que permite un análisis más detallado del comportamiento del empleo en distintos ámbitos.

Todo el proceso realizado en el presente capítulo se lleva a cabo utilizando el programa R, y se encuentra descrito en detalle en el Apéndice A.

7.1. Corrección de las series

Antes de corregir las series, y ya analizado si tienen estacionalidad o no (Sección 5.2.1) cabe detectar qué efectos de calendario y atípicos están presentes en cada caso, con el fin de decidir qué elementos incluir en el modelo tal y como se explicó en la Sección 4.3.

A continuación, en la Tabla 7.1 se presentan los p-valores correspondientes a los contrastes de significación de estos efectos en las distintas series analizadas, y en la Tabla 7.2 los valores atípicos detectados en cada una de las series.

Efecto	**Ocupados**		**Parados**	
	Original	**Desagregado**	**Original**	**Desagregado**
Días laborables	0.334	0.929	0.483	0.000
Año bisiesto	0.371	0.935	0.039	0.020
Pascua	0.379	0.817	0.558	0.020

Tabla 7.1: P-valores de los efectos de calendario en las series de ocupados, parados y sus desagregaciones.

Observando la Tabla 7.1, en la serie de ocupados, los efectos de calendario, como los días laborables, el año bisiesto y la Pascua, no son significativos, por lo que no se incluirán en el modelo. Por otro lado, en la serie de parados solo el efecto del año bisiesto es significativo. Lo mismo ocurre con las series desagregadas del Capítulo 6: los ocupados desagregados no muestran efectos de calendario significativos, mientras que en los parados desagregados, tanto los días laborables, como el año bisiesto y la pascua son significativos.

	Ocupados	**Parados**
Original	LS (2009-01-01), LS (2020-04-01)	LS (2008-10-01), TC (2009-01-01), LS (2020-07-01), TC (2021-10-01)
Desagregado	LS (2020-02-01), LS (2020-03-01), LS (2020-04-01), TC (2020-05-01), LS (2020-07-01), LS (2020-08-01), AO (2020-10-01)	AO (2002-02-01), LS (2020-03-01), TC (2020-04-01), TC (2020-05-01), LS (2020-07-01), LS (2020-08-01), LS (2020-10-01)

Tabla 7.2: Valores atípicos detectados en las series de ocupados, parados y sus desagregaciones.

Los valores atípicos identificados en la Tabla 7.2 se corresponden principalmente con dos periodos de especial impacto económico en España: la crisis financiera de 2008-2009 y la crisis provocada por la pandemia del COVID-19 a partir de 2020. En las series originales, se observa un cambio de nivel (LS) en los ocupados en el primer trimestre de 2009 y en el segundo trimestre de 2020, lo que refleja el fuerte ajuste del mercado laboral tras el estallido de la burbuja inmobiliaria y, más tarde, la interrupción repentina de la actividad económica por las restricciones sanitarias. Por su parte, los parados presentan tanto cambios de nivel como cambios de tendencia (TC) en momentos clave de ambas crisis: principios de 2009, mediados de 2020 y finales de 2021, estos últimos probablemente relacionados con la finalización progresiva de los ERTEs.

En las series desagregadas, se identifican múltiples valores atípicos concentrados en 2020, especialmente entre febrero y octubre, lo cual es coherente con los efectos inmediatos y volátiles de la pandemia sobre el empleo. Además, en el caso de los parados, aparece un valor atípico aditivo (AO) en febrero de 2002, posiblemente asociado a dinámicas específicas del mercado laboral anterior a la crisis financiera. Si se quiere consultar la definición de los tipos de atípicos con más detalle se puede consultar la Sección 4.2.

A continuación, en la Sección 7.2, se realiza el análisis de residuos de los modelos $ARIMA$ ajustados a las series desagregadas corregidas, con el fin de validar estos ajustes.

7.2. Validación de los modelos empleados para la corrección

En la Tabla 7.3 se resume si se ha aplicado o no la transformación logarítmica a los datos, así como los órdenes de los modelos seleccionados y del Criterio de Información Bayesiana (BIC), como medida de bondad del ajuste.

Serie	**Transformación logarítmica**	**p**	**d**	**q**	**P**	**D**	**Q**	**BIC**
Ocupados	Sí	2	1	0	0	1	1	1077.132
Parados	No	1	1	0	1	0	1	1120.513
Ocupados desagregados	No	1	1	0	1	1	1	2804.479
Parados desagregados	Sí	1	1	2	0	1	1	2880.868

Tabla 7.3: Detalles de los modelos ajustados: transformación logarítmica, órdenes del modelo y BIC.

Para llevar a cabo la validación de los residuos de los modelos ajustados, se aplican diversas pruebas. En primer lugar, se comprueba si los residuos tienen media nula empleando el test de la t de Student (Student (1908)). Además, se emplea la prueba de Dickey-Fuller (Dickey and Fuller (1979)) para examinar la estacionariedad de los mismos, la prueba Shapiro-Wilks (Shapiro and Wilk (1965)) para comprobar su normalidad, y la prueba de Ljung-Box (Shapiro and Wilk (1965)) para comprobar la independencia. Los p-valores de las pruebas mencionadas para las distintas series se proporcionan en la Tabla 7.4.

Serie	**Media Nula**	**Estacionariedad**	**Normalidad**	**Independencia**
Ocupados	0.870	0.010	0.528	0.891
Parados	0.875	0.010	0.926	0.901
Ocupados desagregados	0.946	0.010	0.636	0.719
Parados desagregados	0.563	0.010	0.039	0.801

Tabla 7.4: P-valores obtenidos tras la validación de los residuos de los modelos.

Con un 5 % de significación, todas las pruebas realizadas concluyen que no hay evidencias estadísticas para dejar de asumir las cuatro hipótesis estructurales chequeadas. Cabe destacar que, en el caso del modelo de los parados desagregados, se rechaza la normalidad de estos residuos a un nivel de significación del 5 % pero esta hipótesis no es estrictamente necesaria para que el modelo sea válido.

Las funciones de autocorrelación simple y parcial de los residuos de los modelos se muestran en las Figuras 7.1 y 7.2, respectivamente. En general, las autocorrelaciones se mantienen dentro de las bandas de confianza, con algunas excepciones puntuales que podrían atribuirse a simple ruido.

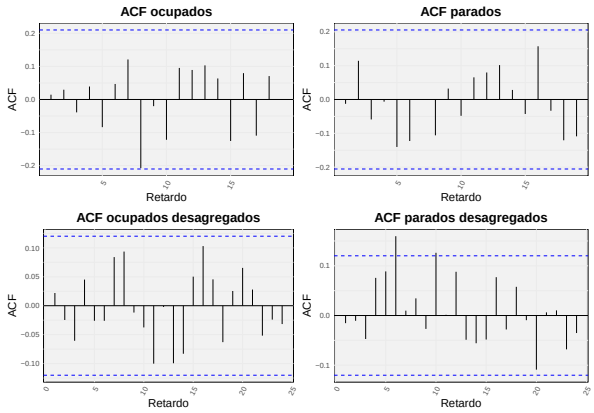

Figura 7.1: Funciones de las autocorrelaciones simples de los residuos de los modelos.

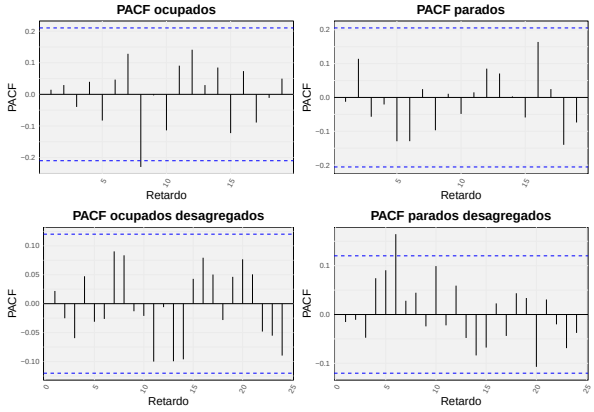

Figura 7.2: Funciones de las autocorrelaciones parciales de los residuos de los modelos.

En la Sección 4.4, se ha explicado la metodología empleada para verificar la presencia de estacionalidad residual y el efecto del ciclo semanal, tanto en la serie ajustada estacionalmente como en su componente irregular. Los resultados de estos tests se resumen en la Tabla 7.5 en la cual podemos ver que las series ya no tienen estacionalidad residual.

Test aplicado	**Ocupados**	**Parados**	**Ocupados desagregados**	**Parados desagregados**
F-test para estacionalidad en la componente irregular	0.914	0.819	1.000	1.000
F-test para estacionalidad en la serie ajustada	0.876	0.906	1.000	1.000
Q-test de Ljung-Box en la componente irregular	1.000	0.952	0.129	1.000
Q-test de Ljung-Box en la serie ajustada	1.000	1.000	1.000	1.000
F-test para efectos de calendario en la componente irregular	–	0.721	–	0.963
F-test para efectos de calendario en la serie ajustada	–	0.866	–	0.644

Tabla 7.5: P-valores de los tests aplicados sobre estacionalidad y efectos de calendario.

Los p-valores asociados a estos test son mayores que los niveles de significación habituales (1 %,

5 % y 10 %) por lo que no hay evidencia suficiente para rechazar la hipótesis nula, es decir, tanto en la serie ajustada estacionalmente como en la componente irregular no existe la estacionalidad residual.

Una vez hecha la validación de los modelos, en la Figura 7.3 se muestran las series corregidas de estacionalidad para los ocupados y los ocupados efectivos desagregados, junto con las series originales sin corregir. En cuanto a los parados y su desagregación, la figura presenta las series corregidas tanto de estacionalidad como de efectos de calendario, junto con las series sin corregir.

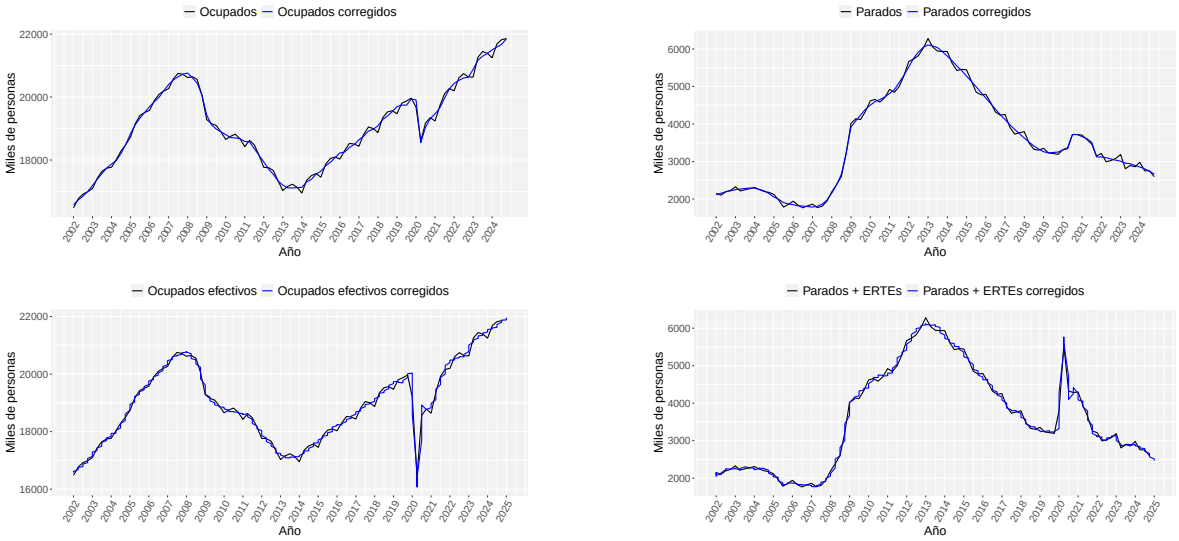

Figura 7.3: Series corregidas y no corregidas de ocupados y parados, junto con sus respectivas desagregaciones.

7.3. Corrección de la serie de afiliaciones por CNAE

En este trabajo también se ha aplicado la metodología de corrección estacional y de calendario a las series de afiliaciones por sectores según la clasificación CNAE-2009. Este ejercicio no solo sirve como análisis en sí mismo, sino que ilustra cómo puede extenderse la metodología a otras series para las que no existen versiones corregidas, ofreciendo una herramienta práctica para analistas de ABANCA.

Para simplificar la presentación, las actividades económicas se han agrupado en cuatro grandes bloques, conforme a la CNAE-2009, cuya clasificación detallada puede consultarse en INE (2022). Los datos utilizados abarcan el período de enero de 2009 a diciembre de 2024.

- **Agricultura, Ganadería, Silvicultura y Pesca (A):**
  - Actividades relacionadas con el aprovechamiento de recursos naturales, vegetales o animales.
  - Incluye producción agrícola, cría de animales, silvicultura, pesca y recolección de productos naturales.
- **Industrias Extractivas, Manufactureras, Suministro de Energía y Agua (B-E):**
  - Incluye industrias extractivas y manufactureras, así como el suministro de energía eléctrica, gas, vapor, agua y gestión de residuos.

- **Construcción (F):**
  - Actividades de edificación, reformas y obras civiles, tanto de estructuras nuevas como prefabricadas.
- **Servicios (G-T):**
  - Incluye comercio, transporte, hostelería, sanidad, educación, actividades recreativas, financieras y de servicios personales y sociales.

En la Figura 7.4 se muestran las series de afiliaciones por CNAE-2009 agrupadas por sector, junto con sus respectivas versiones corregidas de estacionalidad y efectos de calendario.

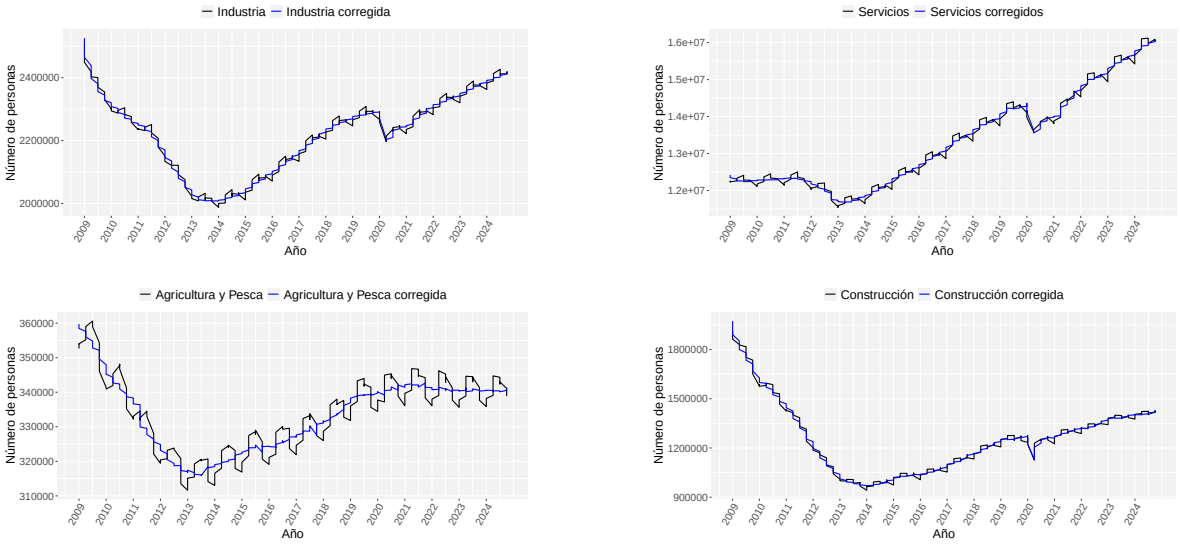

Figura 7.4: Series corregidas y sin corregir de las afiliaciones por CNAE-2009.

Esto ilustra como la metodología de corrección estacional y de calendario puede aplicarse de forma eficaz a otras variables relevantes para el análisis económico que no se publican corregidas de estos efectos.

Capítulo 8

Construcción de los indicadores adelantados de empleo

Con todo lo desarrollado en los capítulos anteriores, se ha alcanzado la fase del trabajo en la que es posible implementar el cálculo del indicador adelantado de la tasa de paro para España.

El capítulo se estructura como sigue: en la Sección 8.1 se explica cómo se construye el indicador, mientras que en la Sección 8.2 se presenta una simulación que ilustra el procedimiento para obtener los datos de dicho indicador.

8.1. Construcción del indicador adelantado de la tasa de paro

El indicador tiene como objetivo proporcionar una estimación temprana y de mayor frecuencia sobre la evolución del desempleo. El proceso de construcción del indicador se basa en la desagregación de las cifras trimestrales de ocupados y parados que publica la EPA a partir de los datos mensuales de afiliaciones a la Seguridad Social y el paro registrado. Con esta información mensual, que se publican con menor desfase temporal que la EPA, es posible estimar mensualmente el número de ocupados y parados. Gracias a esta estimación mensual, se puede calcular de forma anticipada la tasa de paro, entendida como el cociente entre el número de parados y la suma de ocupados más parados, es decir, la población activa.

Además, se contempla tanto una versión corregida de estacionalidad y efectos de calendario como una versión sin corregir, con el fin de garantizar una comparación adecuada con la EPA, que se publica sin corregir de efectos de calendario y estacionalidad. Esta doble aplicación del indicador facilita un análisis más detallado de las dinámicas subyacentes del mercado laboral y, al mismo tiempo, permite contrastar directamente los resultados estimados con los datos oficiales.

Por un lado, en la Figura 8.1 se presentan las series de la tasa de paro original publicada por la EPA y la tasa de paro corregida, calculada utilizando las series de ocupados y parados corregidas, ambas trimestrales. Cabe destacar que el pico que se observa en la gráfica se debe a que la tasa de paro corregida de estacionalidad está calculada teniendo en cuenta el efecto de los ERTEs, a diferencia de la tasa de paro que publica la EPA.

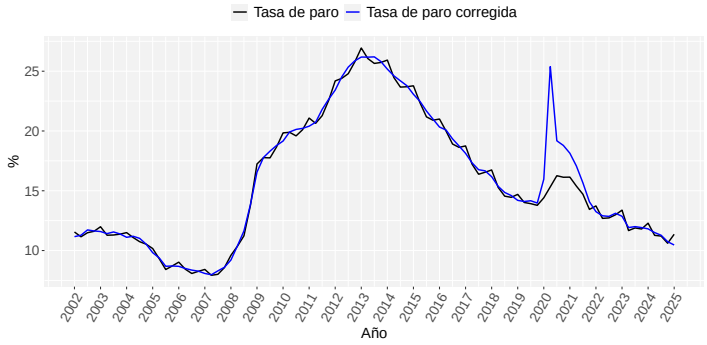

Figura 8.1: Tasa de paro de la EPA y la calculada con el indicador corregida de efectos de estacionalidad y de calendario.

Por otro lado, en la Figura 8.2 se muestra la serie de la tasa de paro mensual estimada a partir del indicador, junto con su versión corregida de estacionalidad y efectos de calendario.

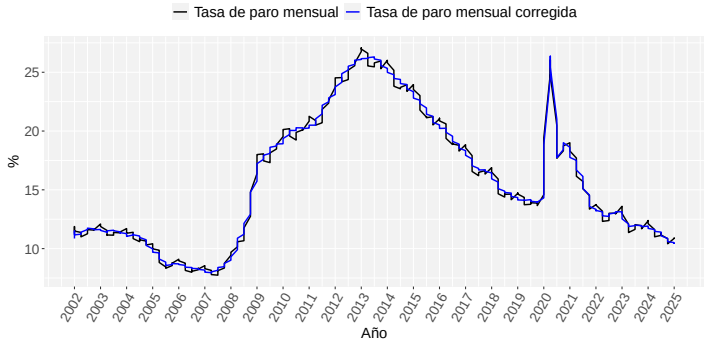

Figura 8.2: Tasa de paro mensual estimada con el indicador, junto con su serie corregida de estacionalidad y efectos de calendario.

8.2. Comparativa del indicador con los datos oficiales

Gracias a las técnicas aplicadas en este trabajo, que incluyen la desagregación de las series temporales, y aprovechando la publicación anticipada de los datos mensuales de paro registrado y afiliaciones a la Seguridad Social, es posible predecir los primeros meses del trimestre en curso para las series de ocupados y parados. Esto permite obtener una tasa de paro no solo desagregada históricamente, sino también adelantada en el tiempo, facilitando un análisis más detallado del mercado laboral.

En esta sección se presenta la estimación mensual de la tasa de paro correspondiente a los meses del primer trimestre de 2025, así como la media trimestral resultante. Estas estimaciones se compararán posteriormente con el dato oficial publicado por la EPA, lo que permitirá evaluar la precisión del indicador adelantado desarrollado.

Por lo tanto, como recordatorio, el punto de partida del trabajo son las siguientes estadísticas de mercado laboral:

- Los datos de la EPA (ocupados, parados, tasa de paro), que son trimestrales y se publican casi

un mes después de que se cierre el trimestre en cuestión.

- Los datos de afiliaciones y paro registrado, que son mensuales y se publican dos días tras el cierre del mes.

Debido a la fecha de realizacion de este trabajo, la información disponible era:

- EPA: La EPA del cuarto trimestre (4T) está disponible. La del primer trimestre (1T) no se publica hasta finales de abril de 2025.
- Datos Mensuales: Se dispone de los datos mensuales de parados y afiliaciones correspondientes a los meses transcurridos del primer trimestre.

Y, a modo ilustrativo, un ejemplo del proceso a lo largo del trimestre en curso sería el siguiente:

- 11 de febrero: Se tienen los datos de parados y afiliaciones de Enero de 2025. Con estos datos, se calcula la tasa de paro mensual de enero de 2025.
- 11 de marzo: Se tienen los datos de parados y afiliaciones de enero y febrero de 2025. Con estos datos, se calculan las tasas de paro mensuales de enero y febrero de 2025.
- 11 de abril: Se tienen los datos de parados y afiliaciones de enero, febrero y marzo de 2025. Con estos datos, se calculan las tasas de paro mensuales de enero, febrero y marzo de 2025, permitiendo una anticipación de la tasa de paro del primer trimestre de 2025.

Con el fin de facilitar la comprensión de las fechas previamente mencionadas, en la Figura 8.3 se muestra un esquema explicativo.

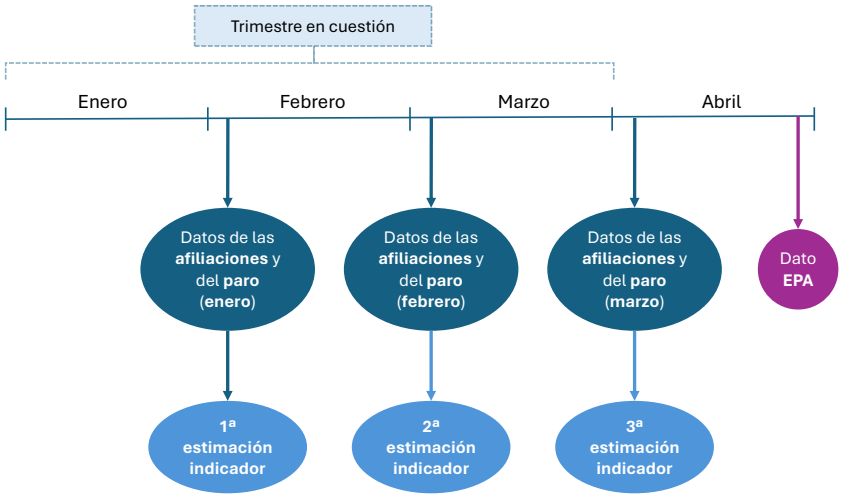

Figura 8.3: Línea temporal ilustrativa del proceso de construcción del indicador adelantado de la tasa de paro.

Entonces, siguiendo el proceso expuesto y teniendo en cuenta que en este momento ya se conocen los datos oficiales del primer trimestre, se pueden analizar los datos predichos con el método empleado

y compararlos con los datos oficiales, con el fin de evaluar la eficacia de la solución propuesta en el presente trabajo.

La Tabla 8.1 recoge los datos oficiales de ocupados, parados y tasa de paro en el primer trimestre de 2025, según la EPA, conjuntamente con las estimaciones arrojadas por el indicador. Además, se presentan las estimaciones mensuales para esas series en ese trimestre.

Indicador	**Enero**	**Febrero**	**Marzo**	**1T 2025**
Ocupados (predichos)	21552.220	21635.570	21789.430	21659.080
Parados (predichos)	2628.297	2644.661	2643.396	2638.785
Tasa de paro (% predicha)	10.869	10.892	10.819	10.860
Ocupados (real)	–	–	–	21765.400
Parados (real)	–	–	–	2789.200
Tasa de paro (% real)	–	–	–	11.360

Tabla 8.1: Comparación de la tasa de paro, ocupados y parados estimados con los datos reales para el primer trimestre de 2025.

En cuanto a los ocupados, la estimación para el primer trimestre de 2025 fue de 21659.080 miles de personas, mientras que el dato real registrado por la EPA fue de 21765.400 miles, resultando un error relativo inferior al 0.500 % entre el dato real y el predicho. Por otro lado, los parados fueron estimados en 2638.790 miles, frente a los 2789.200 miles registrados oficialmente, lo que representa una ligera desviación, con un error relativo del 5,390 %. Aunque se podría decir que las estimaciones no presentan diferencias muy significativas respecto a los valores oficiales, los resultados son útiles para comprender las tendencias generales del mercado laboral, proporcionando datos anticipados que permiten prever la evolución del mercado con un significativo adelanto respecto al dato oficial, que se publica con considerable retraso.

En cuanto a la tasa de paro, la estimación para el primer trimestre de 2025 fue del 10.860 %, mientras que el dato oficial publicado por la EPA alcanzó el 11.360 %. La diferencia entre ambos valores es de 0.500 puntos porcentuales, lo que representa un error relativo del 4.400 % respecto al dato oficial. Esta discrepancia, aunque moderada, es razonable teniendo en cuenta que la estimación se basa en datos mensuales sin los ajustes metodológicos completos de la EPA. Así, el indicador adelantado proporciona una aproximación útil y oportuna para anticipar la evolución del mercado laboral.

En resumen, se ha conseguido construir una serie temporal de tasas de paro mensuales, basada en datos de alta frecuencia (mensuales), para obtener una estimación temprana de la tasa de paro trimestral, que se publica con retraso a través de la EPA.

Capítulo 9

Conclusiones y líneas futuras

Para cerrar esta memoria, este capítulo expone las principales conclusiones derivadas del análisis desarrollado, basadas en los resultados presentados en el Capítulo 8.

En la Sección 9.1 se recogen y comentan los principales resultados obtenidos a lo largo del proyecto. Finalmente, en la Sección 9.2 se presentan distintas propuestas y líneas de trabajo futuras que se derivan del desarrollo realizado.

9.1. Conclusiones

El objetivo de este trabajo es proponer una metodología que permita a ABANCA estimar de forma adelantada la tasa de paro, aprovechando la disponibilidad mensual de los datos de afiliación a la Seguridad Social y del paro registrado. Para ello, se plantean una serie de pasos que culminan en la construcción del indicador deseado.

En una primera fase, se descargaron los datos de ocupados y parados procedentes de la EPA, junto con los datos mensuales de afiliaciones y paro registrado. Estos se transformaron en series temporales, lo que permitió su análisis desde una perspectiva dinámica. A continuación, se realizó un análisis de correlación para estudiar la relación entre las afiliaciones y los ocupados, así como entre el paro registrado y el número de parados. Los resultados mostraron una relación significativa en ambos casos, lo cual justifica el uso de las series mensuales (afiliaciones y paro registrado) para la desagregación temporal de los datos trimestrales de ocupados y parados, respectivamente.

Un aspecto especialmente relevante en este proceso fue el tratamiento de los ERTEs durante la pandemia del COVID-19. La presencia de los ERTEs alteró significativamente las estadísticas laborales, afectando tanto a los ocupados como a los parados, y generando problemas en los modelos de desagregación temporal. Estos modelos, al aplicar los distintos métodos, no lograban superar la validación.

Para resolver esta limitación, fue necesario realizar un ajuste específico: se incorporaron los datos de ERTEs restándolos y sumándolos a las series de ocupados y parados, respectivamente. Esta corrección fue crucial, ya que permitió adaptar los modelos a la realidad del mercado laboral en ese periodo. Gracias a este proceso, los modelos desagregados superaron la validación del modelo, lo que pone de manifiesto la utilidad clave de considerar los ERTEs como una variable de ajuste indispensable en el análisis.

Una vez obtenidas las series mensualizadas de ocupados y parados, se construye el indicador de tasa de paro dividiendo el número de parados entre la suma de ocupados y parados. Esto permite estimar una tasa de paro adelantada que puede ser comparada con el dato oficial, una vez publicado por la EPA, para evaluar su precisión y utilidad como herramienta de previsión.

Como parte de este proceso, ABANCA dispone también de una metodología propia para la corrección de la estacionalidad y de los efectos de calendario, aplicable a cualquier serie temporal. Esto es especialmente relevante, ya que los datos no suelen publicarse ajustados, y dicha corrección permite eliminar fluctuaciones estacionales, facilitando un análisis más preciso.

9.2. Líneas futuras

A raíz del desarrollo de este Trabajo de Fin de Máster, se identifican diversas líneas de trabajo que podrían explorarse en el futuro.

En primer lugar, dado que el objetivo principal ha sido la estimación anticipada de la tasa de paro en España y el análisis del comportamiento del empleo, una posible extensión natural sería aplicar esta metodología a otros contextos geográficos, como Galicia y Portugal. Estas regiones tienen especial relevancia para ABANCA, por lo que adaptar el modelo a dichos territorios podría aportar valor añadido desde el punto de vista estratégico y operativo.

Asimismo, la técnica de desagregación temporal empleada en este estudio podría trasladarse a otras variables macroeconómicas de gran interés, como el Producto Interior Bruto (PIB), cuyo seguimiento mensual resulta clave para el diagnóstico y la planificación económica.

Apéndice A

Funciones empleadas

La función que se ha utilizado para la desagregación temporal es `temporaldisaggregation` del paquete `rjd3toolkit` (Palate et al. (2025a)), la cual se define de la siguiente manera:

```
temporaldisaggregation(
series,
constant = TRUE,
trend = FALSE,
indicators = NULL,
model = c("Ar1", "Rw", "RwAr1"),
freq = 4,
conversion = c("Sum", "Average", "Last", "First", "UserDefined"),
conversion.obsposition = 1,
rho = 0,
rho.fixed = FALSE,
rho.truncated = 0,
zeroinitialization = FALSE,
diffuse.algorithm = c("SqrtDiffuse", "Diffuse", "Augmented"),
diffuse.regressors = FALSE
)
```

- `series`: Serie temporal que se va a desagregar. Debe ser un objeto de tipo `ts`.
- `constant`: Indica si se incluye un término constante en el modelo. Este parámetro se utiliza solo con el modelo `Ar1` cuando la inicialización a cero es `FALSE`.
- `trend`: Especifica si se incorpora una tendencia lineal.
- `indicators`: Lista de variables indicadoras de alta frecuencia que se usarán en la desagregación temporal. Debe ser de tipo `ts`.
- `model`: Modelo para el término de error a nivel desagregado. Las opciones son: `Ar1` (método de Chow-Lin), `Rw` (método de Fernández), y `RwAr1` (método de Litterman).
- `freq`: Frecuencia de la variable que se desagrega. Este parámetro se utiliza solo si no se proporcionan indicadores.
- `conversion`: Modo de agregación a utilizar.
- `conversion.obsposition`: Este parámetro se usa solo cuando el modo de agregación es `UserDefined`.

- `rho`: Se aplica únicamente a los modelos `Ar1` o `RwAr1`. Representa el valor inicial del parámetro ρ.
- `rho.fixed`: Si es `TRUE` no se estima ρ y se toma el valor específico de ρ.
- `rho.truncated`: Si la estimación de ρ es menor se fija al valor aquí especificado. Por defecto es 0 porque valores negativos de ρ puede dar problemas.
- `zeroinitialization`: Si es `TRUE` se toman valores iniciales 0 para la estimación. Se recomienda dejarlo en `FALSE`.
- `diffuse.algorithm`: Tipo de algoritmo usado para la estimación por filtro de Kalman.
- `diffuse.regressors`: Si es `TRUE` se usa inicialización difusa para los regresores en la estimación por filtro de Kalman.

Para más información del resto se puede consultar el manual de R del paquete `rjd3bench` (Palate (2025)).

Para llevar a cabo el proceso de corrección de estacionalidad y efectos de calendario es necesario instalar y cargar dos librerías en R, las cuales son:

- `Rjd3tramoseats` (Palate et al. (2025b)).
- `rjd3toolkit` (Palate et al. (2025a)).

La función empleada es `tramoseats_fast` se define de la siguiente manera:

```
tramoseats_fast(
ts,
spec = c("rsafull", "rsa0", "rsa1", "rsa2", "rsa3", "rsa4", "rsa5"),
context = NULL,
userdefined = NULL
)
```

donde

- `ts`: Una serie temporal univariante.
- `spec`: La especificación del modelo. Puede ser el nombre de una especificación predefinida o una especificación definida por el usuario. Las predefinidas se explican en la Tabla A.1.
- `context`: El diccionario de variables.
- `userdefined`: Un vector que contiene las variables adicionales.

Las especificaciones pueden ser las siguientes:

Identificador	Log	Valores atípicos	Efectos de calendario	Arima
RSA0	NA	NA	NA	Modelo Airline(+media)
RSA1	automático	AO/LS/TC	NA	Modelo Airline(+media)
RSA2	automático	AO/LS/TC	2 vars td + Pascua	Modelo Airline(+media)
RSA3	automático	AO/LS/TC	NA	automático
RSA4	automático	AO/LS/TC	2 vars td + Pascua	automático
RSA5	automático	AO/LS/TC	7 vars td + Pascua	automático
RSAfull	automático	AO/LS/TC	automático	automático

Tabla A.1: Especificaciones del modelo ARIMA en función de varios parámetros.

Inicialmente, para corregir las series de ocupados, parados o cualquier otra serie utilizada a lo largo del trabajo, se va a emplear la especificación **RSA3**. Como se puede ver en la Tabla A.1, en esta especificación no se consideran efectos del calendario. Estos efectos se van a definir utilizando varias funciones de la librería `rjd3toolkit`, específicamente las funciones `set_tradingdays` y `set_easter`, las cuales se definen de la siguiente manera:

```
set_tradingdays(
x,
option = c(NA, "TradingDays", "WorkingDays", "TD3", "TD3c", "TD4", "None",
"UserDefined"),
calendar.name = NA,
uservariable = NA,
stocktd = NA,
test = c(NA, "None", "Remove", "Add", "Separate_T", "Joint_F"),
coef = NA,
coef.type = c(NA, "Fixed", "Estimated"),
automatic = c(NA, "Unused", "FTest", "WaldTest", "Aic", "Bic"),
pftd = NA,
autoadjust = NA,
leapyear = c(NA, "LeapYear", "LengthOfPeriod", "None"),
leapyear.coef = NA,
leapyear.coef.type = c(NA, "Fixed", "Estimated")
)
```

```
set_easter(
x,
enabled = NA,
julian = NA,
duration = NA,
test = c(NA, "Add", "Remove", "None"),
coef = NA,
coef.type = c(NA, "Estimated", "Fixed"),
type = c(NA, "Unused", "Standard", "IncludeEaster", "IncludeEasterMonday")
)
```

Para obtener más información sobre los detalles de estas dos funciones se puede consultar el manual del paquete `rjd3toolkit` (Palate et al. (2025a)).

Bibliografía

Aneiros, G. (2022). *Series de tiempo*. Apuntes de la asignatura, Universidade da Coruña.

Bournay, J. and Laroque, G. (1979). Réflexions sur la méthode d'elaboration des comptes trimestriels. In *Annales de l'INSEE*, pages 3–30. JSTOR.

Box, G. E. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 26(2):211–243.

Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.

Chan, K.-S. and Cryer, J. D. (2008). *Time Series Analysis With Applications in R*. Springer.

Chow, G. C. and Lin, A.-l. (1971). Best linear unbiased interpolation, distribution, and extrapolation of time series by related series. *The review of Economics and Statistics*, pages 372–375.

Ciammola, A., Di Palma, F., and Marini, M. (2005). Temporal disaggregation techniques of time series by related series: A comparison by a monte carlo experiment. Technical report, Eurostat, Working Paper.

Diario de Navarra (2018). Qué es un ERTE y sus consecuencias. https://www.diariodenavarra.es/noticias/negocios/dn-management/2018/08/21/que-erte-sus-consecuencias-606635-2541.html. Accedido 15 mayo 2025.

Dickey, D. A. and Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American statistical association*, 74(366a):427–431.

Doornik, J. A. and Hansen, H. (2008). An omnibus test for univariate and multivariate normality. *Oxford bulletin of economics and statistics*, 70:927–939.

EPA (2008). Metodología 2005. *Descripción de la encuesta, definiciones e instrucciones para la cumplimentación del cuestionario. Madrid*.

Eurostat (2024). *ESS Guidelines on Seasonal Adjustment*. https://ec.europa.eu/eurostat/documents/3859598/19355229/KS-GQ-24-012-EN-N.pdf/be954db6-64f5-c1a2-a0b8-0b4de2a5c707?version=1.0&t=1718182927069.

Eurostat (2024). Metadata annex: Unemployment rate by sex and age. https://ec.europa.eu/eurostat/cache/metadata/Annexes/une_rt_m_esms_an_1.pdf. Accedido el 19 de mayo de 2025.

Fernandez, R. B. (1981). A methodological note on the estimation of time series. *The Review of Economics and Statistics*, 63(3):471–476.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the american statistical association*, 32(200):675–701.

Gómez, V. and Gómez, V. (1998). *Butterworth filters: a new perspective*. Dirección General de Análisis y Programación Presupuestaria. https://www.sepg.pap.hacienda.gob.es/sitios/sepg/es-ES/Presupuestos/DocumentacionEstadisticas/Documentacion/Documents/DOCUMENTOS%20DE%20TRABAJO/D98008.pdf.

Gómez, V. and Maravall Herrero, A. (1998). *Guide for using the programs TRAMO and SEATS: beta version: December 1997*. Banco de España. Servicio de Estudios.

Hoff, P. (2022). Best Lineal Unbiased Estimation. https://www2.stat.duke.edu/~pdh10/Teaching/721/Materials/ch2blue.pdf.

IGE (2017). Banco de series de conxuntura. Ciclotendencia e series corrixidas de estacionalidade e calendario. Notas explicativas. Technical report, Instituto Galego de Estatística. Galicia, España.

INE (2002a). Ajuste estacional y extracción de señales en la Contabilidad Trimestral Nacional Trimestral. *Boletín Trimestral de Coyuntura*, 84. España.

INE (2002b). Extracción de señales y ajuste estacional en la CNTR. *Boletín Trimestral de Coyuntura*, 85. España.

INE (2019). Manual de usuario de JDemetra+. pages 123–130. https://www.ine.es/clasifi/manual_jdemetra.pdf.

INE (2022). *CNAE-2009*. https://www.ine.es/daco/daco42/clasificaciones/cnae09/notasex_cnae_09.pdf.

INE (2023). Actualización de modelos de ajuste estacional para las series de empleo por sexo. https://www.ine.gob.cl/docs/default-source/ocupacion-y-desocupacion/publicaciones-y-anuarios/separatas/tematicas/actualizacion-ajuste-estacional-2020.pdf. Chile.

Jarque, C. M. and Bera, A. K. (1987). A test for normality of observations and regression residuals. *International Statistical Review/Revue Internationale de Statistique*, pages 163–172.

JDemetra+ Development Team (2024). Theory behind seasonal adjustment. https://jdemetradocumentation.github.io/JDemetra-documentation/pages/theory/. Accedido el 19 de mayo de 2025.

Kruskal, W. H. and Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621.

Litterman, R. B. (1983). A random walk, markov model for the distribution of time series. *Journal of Business & Economic Statistics*, 1(2):169–173.

Ljung, G. M. and Box, G. E. (1978). On a measure of lack of fit in time series models. *Biometrika*, 65(2):297–303.

Luís Jiménez Arellano (2020). Qué es un ERTE y otras seis claves sobre los expedientes de regulación temporal. https://www.elespanol.com/invertia/economia/empleo/20200314/erte-claves-expedientes-regulacion-temporal/474454644_0.html. Accedido 15 mayo 2025.

Lytras, D. P., Feldpausch, R. M., and Bell, W. R. (2007). Determining seasonality: a comparison of diagnostics from x-12-arima. *US Census Bureau*.

Maravall, A. (2011). Seasonality tests and automatic model identification in tramo-seats. *Bank of Spain: Madrid, Spain*.

Maravall, A. (2012). Update of seasonality tests and automatic model identification in tramo-seats. *Banco de Espana*.

Maravall, A. and Caporello, G. (2004). Program TSW. Revised manual. Version may 2004. *Documentos ocasionales-Banco de España*, (8):5–58.

McLeod, A. (2022). *Kendall: Kendall Rank Correlation and Mann-Kendall Trend Test*. R package version 2.2.1.

Melo-Velandia, L. F. (2010). Una metodología multivariada de desagregación temporal. *Borradores de Economía; No. 586*.

Ministerio de Inclusión, S. S. y. M. (2025). Metodología de estimación de factores estacionales para series de afiliación mensuales y diarias. https://www.seg-social.es/wps/wcm/connect/wss/6e40c419-9051-462b-a338-477d91beb4e1/Metodologia+y+Fuentes+de+Ajuste+Estacional+2024.pdf. Accedido el 19 de mayo de 2025.

Ollech, D. (2021). *seastests: Seasonality Tests*. R package version 0.15.4. https://CRAN.R-project.org/package=seastests.

Ooms, J. (2014). The jsonlite package: A practical and consistent mapping between json data and r objects. *arXiv:1403.2805 [stat.CO]*.

Palate, J. (2025). *rjd3bench: Interface to 'JDemetra+ 3.x' time series analysis software*. R package version 2.1.0.

Palate, J., la Tente, A. Q., Barthelemy, T., and Smyk, A. (2025a). *rjd3toolkit: Utility Functions around 'JDemetra+ 3.0'*. R package version 3.3.0.

Palate, J., la Tente, A. Q., Barthelemy, T., and Smyk, A. (2025b). *rjd3tramoseats: Seasonal Adjustment with TRAMO-SEATS in 'JDemetra+ 3.x'*. R package version 3.3.0.

Peña, D. (2005). *Análisis de series temporales*. Alianza.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

SEPE (2022). Conceptos. https://www.sepe.es/HomeSepe/es/que-es-el-sepe/estadisticas/datos-avance/conceptos.html.

Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4):591–611.

Shumway, R. H., Stoffer, D. S., and Stoffer, D. S. (2000). *Time series analysis and its applications*, volume 3. Springer.

Social, S. (2025). Afiliación de trabajadores. https://www.seg-social.es/wps/portal/wss/internet/Trabajadores/Afiliacion/7332.

Student (1908). The probable error of a mean. *Biometrika*, pages 1–25.

Vázquez, J. (2018). D'economía blog. http://www.deconomiablog.com/2018/03/el-mercado-de-trabajo-poblacion-activa.html.

Welch, B. L. (1951). On the comparison of several mean values: an alternative approach. *Biometrika*, 38(3/4):330–336.