Trabajo Fin de Máster

Proyecciones del PIB de Galicia a partir de modelos estadísticos

Teresa Veiga Rodríguez

Máster en Técnicas Estadísticas

Curso 2017-2018

Propuesta de Trabajo Fin de Máster

Título en galego: Proxeccións do PIB de Galicia a partir de modelos estatísticos

Título en español: Proyecciones del PIB de Galicia a partir de modelos estadísticos

English title: Forecasting models for Galician GDP

Modalidad: Modalidad B

Autor/a: Teresa Veiga Rodríguez, Universidade de Vigo

Director: Germán Aneiros Pérez, Universidade da Coruña

Tutora: Belén María Fernández de Castro, ABANCA

Breve resumen del trabajo:

Desde el Departamento de Planificación Estratégica y PMO de ABANCA se lleva a cabo con frecuencia semanal un seguimiento de la economía gallega. Surge en este ámbito de estudio la necesidad de un modelo para obtener proyecciones del principal agregado macroeconómico, el Producto Interior Bruto, que tiene un desfase de publicación de unos 50 días. En esta memoria se recogen las metodologías basadas en el análisis de series temporales a las que se ha recurrido para abordar y resolver el problema, así como los conocimientos que ha sido necesario adquirir.

Don Germán Aneiros Pérez, Profesor Titular de la Universidade da Coruña y doña Belén María Fernández de Castro, Especialista de ABANCA, informan que el Trabajo Fin de Máster titulado

Proyecciones del PIB de Galicia a partir de modelos estadísticos

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

En A Coruña, a 26 de enero de 2018.

El director:

Don Germán Aneiros Pérez

La tutora:

Doña Belén María Fernández de Castro

La autora:

Doña Teresa Veiga Rodríguez

Agradecimientos

Agradecer en primer lugar a la empresa ABANCA, por brindarme la oportunidad de tener una primera toma de contacto con el mundo laboral, que ha sido muy enriquecedora. A todos las compañeras y compañeros que he tenido la oportunidad de conocer en el transcurso de las prácticas y, en particular, a mi tutora Belén Fernández de Castro.

En el ámbito académico, agradecer a todo el profesorado del Máster en Técnicas Estadísticas los conocimientos que me han aportado y haber inculcado en mi un gran interés por esta área de estudio. Mención especial a mi director académico, Germán Aneiros Pérez.

Índice general

Resumen

Resumen en español

El Producto Interior Bruto es uno de los principales agregados trimestrales macroeconómicos utilizados en los procesos de seguimiento macroeconómico que se integran en los procesos de presupuestación y planificación de entidades financieras como ABANCA Corporación Bancaria S.A. El departamento de Planificación Estratégica y PMO de ABANCA realiza este seguimiento de forma semanal centrándose en la región de Galicia, ya que es la principal área de desarrollo de su actividad. Dado que la publicación del PIB tiene un desfase de unos 50 días tras el final del trimestre, resulta de suma importancia un modelo que proporcione proyecciones del PIB.

A lo largo de este trabajo se estudian distintas metodologías para abordar el problema en cuestión, en concreto los modelos Box–Jenkins y las regresiones dinámicas. Para facilitar el acceso y manejo de los modelos ajustados con estas metodologías, así como para simplificar el seguimiento macroeconómico, se ha desarrollado una aplicación que permite al usuario interactuar sin necesidad de trabajar con código. Por último, a raíz de las necesidades que han surgido en el desarrollo de este trabajo, ha sido necesario hacer una revisión de la teoría de corrección de estacionalidad y calendario de las series temporales.

English abstract

The Gross Domestic Product is one of the main macroeconomic quarterly aggregated figures used in macroeconomic monitoring processes that are integrated in the planning and budgeting process of financial organisations such as ABANCA Corporación Bancaria S.A. The Department of Strategic Planning and PMO of ABANCA carries out this monitoring on a weekly basis focusing in the region of Galicia, since this is its main area of activity. In view of the fact that the publication of the GDP lags up to 50 days after a quarter end, a model providing projections of the GDP is of utmost importance. The aforementioned

Throughout this paper, different methodologies are considered to address the aforesaid problem in question, specifically Box-Jenkins models and dynamic regressions. In order to facilitate the access and management of the models adjusted with these methodologies, as well as to simplify the macroeconomic monitoring, an application has been developed. Code-illiterate users are able to interact with the application since knowledge of the code is not necessary. Finally, as a result of the needs that have arisen in the development of this paper, it has been necessary to review the seasonality and calendar correction theory in time series.

Prefacio

En el marco del seguimiento macroeconómico que se lleva a cabo desde el Departamento de Planificación Estratégica y PMO de ABANCA Corporación Bancaria S.A. (de ahora en adelante, ABANCA) para la economía gallega, surge la necesidad de contar con un modelo que proporcione proyecciones del Producto Interior Bruto de Galicia, por ser el principal medidor de la economía. Esta modelización esta motivada por el hecho de que este agregado macroeconómico tiene un desfase de publicación que está en torno a los 50 días.

Para abordar el problema se recurre al análisis de series temporales, rama de la Estadística cuyo propósito es doble. Por un lado, entender y modelizar el proceso estocástico que ha generado la serie y, por otro lado, poder construir predicciones. Es en este ámbito teórico en el que se desarrollan dos metodologías diferentes pero muy vinculadas, la metodología Box-Jenkins y los modelos de regresión dinámica. Los modelos ajustados se han volcado en una aplicación, que permite al usuario no conocedor del lenguaje de programación empleado acceder a la herramienta desarrollada.

El trabajo se ha dividido en dos partes muy diferenciadas. Una primera parte en la que se trata toda la metodología necesaria para desarrollar y dar solución al problema en cuestión y una segunda parte donde se aborda la cuestión desde un punto de vista práctico, se describen las variables y se muestran los resultados obtenidos. A continuación se detallan y describen los capítulos de cada una de estas partes.

Metodología

En el primer capítulo se motiva la necesidad de un modelo para el Producto Interior Bruto y se introducen los conceptos necesarios, tanto para adquirir unas nociones básicas de economía y comprender qué es y qué representa la serie objeto de estudio, como para poder desarrollar la teoría estadística empleada en este trabajo.

Los Capítulos 2, 3 y 4 se desarrollan en el marco del estudio de las series temporales. En el Capítulo 2 se revisa la metodología Box-Jenkins, que incluye un abanico muy amplio de procesos bajo los cuales gran parte de las series temporales pueden ser modelizadas. A continuación, en el Capítulo 3 se describen los modelos de regresión dinámica, que son una generalización de los modelos de regresión lineal en el contexto de las series temporales. Cuando se trabaja y analiza el comportamiento de variables económicas que están indexadas en el tiempo, es muy habitual trabajar con las series corregidas de calendario y estacionalidad. Uno de los motivos por los que se ha abordado esta metodología en el Capítulo 4 es que la serie que se pretende modelizar está sometida a este tratamiento, y será necesario que aquellas series candidatas a ser variables explicativas también lo estén.

Se ha desarrollado una aplicación en la que se han volcado los modelos desarrollados y que permite al usuario no conocedor del lenguaje de programación empleado, R (R Core Team (2015)), acceder a la metodología y los modelos implementados. Asimismo, se ha automatizado la descarga de los indicadores de los que se hace seguimiento a través de la API que ofrece el Instituto Galego de Estatística (IGE). Una breve introducción a los conocimientos necesarios para desarrollar lo anterior se detallan en el Capítulo 5.

Parte Práctica

En el Capítulo 6 se describen las variables empleadas para el estudio y se muestra la pestaña de la aplicación que se ha diseñado para poder llevar a cabo una parte del seguimiento macroeconómico de manera más cómoda y eficiente a como se estaba haciendo en la Entidad.

En los capítulos 7 y 8 se ponen en práctica las distintas metodologías estudiadas para el tratamiento de series temporales. Se han programado diversas rutinas que permiten elegir de manera automática los mejores modelos, tanto para el caso de la metodología ARIMA como para elegir la mejor combinación de variables explicativas. Toda esta metodología se ha volcado dentro de la aplicación y, en el cierre de cada capítulo, se muestra la interfaz y el funcionamiento.

En el Capítulo 9 se comparan los modelos ajustados por las distintas metodologías estudiadas y se motiva el uso de los modelos dinámicos; cuya principal ventaja es la capacidad de anticiparse a los shocks económicos, al contrario que los modelos Box-Jenkins.

Se concluye esta memoria con el Capítulo 10, donde se hace un breve análisis de lo que ha sido el trabajo y se muestran las líneas de trabajo que han quedado abiertas o que la realización de este trabajo ha permitido abrir.

Parte I

Metodología

Capítulo 1

Conceptos previos y motivación del problema

Este es un capítulo introductorio en el que se tratan los conceptos necesarios para comprender esta memoria y se motiva el objetivo; que es la búsqueda de un modelo estadístico para predecir el Producto Interior Bruto.

En la Sección 1.1, siguiendo Mochón (2006), se describe qué es el Producto Interior Bruto, comúnmente denominado PIB, además de las dos maneras que existen de cuantificarlo y los distintos enfoques para su construcción. En la Sección 1.2 se describe el seguimiento macroeconómico que se lleva a cabo en ABANCA, seguimiento totalmente integrado en el proceso de planificación y presupuestación de la Entidad. En este marco, se motiva el problema a tratar.

Por otro lado, ya desde un enfoque más técnico, en la Sección 1.3 se definen una serie de conceptos fundamentales para el estudio de series temporales, como el de proceso estocástico y las diferentes medidas estadísticas que se emplean para describirlo. Se verá que para poder hacer inferencia es necesario imponer ciertas condiciones de estabilidad al proceso y se detallan tres maneras de describir los procesos estocásticos.

1.1. El PIB

La macroeconomía se ocupa de estudiar el funcionamiento de la economía en su conjunto, siendo uno de los objetivos agregar los distintos bienes y servicios hasta reducirlos a un solo bien genérico. El más importante de los agregados es el Producto Interior Bruto, comúnmente denominado PIB, que se define como el valor monetario total de los bienes y servicios producidos para el mercado durante un año dentro de las fronteras de un país.

Para poder entender esta definición, siguiendo Mochón (2006, Cap.1), se analizan cada una de las palabras que la integran:

El valor monetario total...

Una economía produce una gran cantidad de bienes y servicios muy dispares, medidos cada uno de ellos en distintas unidades, ¿cómo se podría contabilizar de manera conjunta las ventas de manzanas con el gasto en educación?. Con el objetivo de englobarlos todos en una única cifra se suma el valor monetario de cada uno de ellos (euros).

...de los bienes y servicios finales...

Los bienes son aquellos objetos y mercancías tangibles fabricados por una economía: vehículos, ropa, viviendas, etcétera. Los servicios son actividades intangibles que buscan satisfacer las necesidades de los individuos: la consulta del dentista, una entrada de cine y la educación, por ejemplo.

Los bienes intermedios son los que se emplean en el proceso de producción de otro bien. Por ejemplo, para producir una barra de pan el panadero necesitará comprar los distintos ingredientes a proveedores, que a su vez quizás también se abastecen de otros proveedores. En este caso, la barra de pan es un bien final, mientras que la harina y el resto de ingredientes son bienes intermedios.

Los servicios como un corte de pelo se usan en el momento en que se producen, por lo que siempre son servicios finales. Sin embargo, los servicios que una empresa presta a otra se consideran servicios intermedios. Por ejemplo, el servicio de limpieza que una empresa de limpieza presta a un hotel es un servicio intermedio del servicio final, que es el alojamiento que el hotel presta a sus clientes.

En el cálculo del PIB sólo se contabilizan los bienes y servicios finales ya que, si se contabilizaran los intermedios se produciría una contabilización múltiple (en el caso del panadero, al incluir la harina y la barra de pan, se estaría contando dos veces la harina). El PIB también incluye las llamadas existencias finales que son aquellos bienes y servicios que, estando destinados a ser un producto final, no se han integrado en el proceso productivo al final del período de cálculo.

...producidos...

Sólo se incluyen los bienes y servicios producidos en el año en cuestión. La compra de activos financieros, la tierra y los recursos naturales que hay en ella no se consideran bienes ni servicios producidos. Tampoco se incluye las compras de bienes de segunda mano, por ser bienes que ya fueron producidos (en el momento de la primera compra).

...para el mercado...

Los bienes y servicios que se incluyen son sólo aquellos que se producen con la intención de ser vendidos en el mercado. Se incluirán las barras de pan que el panadero hornea para vender, pero no aquellas que hornea para su propio consumo.

...durante un año dado...

El PIB es una variable de flujo, representa la cantidad producida en un período.

...dentro de las fronteras del país

Solamente se incluyen los bienes y servicios producidos dentro del país, sea por trabajadores y empresas nacionales o extranjeras. Por ejemplo, si un inglés trabaja en España se contabiliza en el PIB español; mientras que un español que trabaja en Londres se contabilizará en el PIB inglés.

1.1.1. Enfoques en el cálculo del PIB

El cálculo del PIB se puede abordar desde tres ángulos, que son la producción, el gasto y las rentas. A continuación se estudia cada uno de ellos:

El enfoque del gasto: es el resultado de sumar el valor de todas las compras de bienes y servicios finales por los distintos grupos de la economía:

Consumo de bienes y servicios: es el gasto de bienes (perecederos y no perecederos) y servicios realizado por las familias.

Inversión privada de bienes y servicios: es la suma de:

- Planta y equipos comprados por las empresas: almacenes, fábricas, maquinarias etcétera.
- Construcción residencial: aunque la mayoría de la adquisición de nuevas viviendas la hacen las familias, se incluye en este bloque.
- La variación de existencias, son los bienes que se han producido pero aún no se han vendido (los bienes que están en las estanterías de las tiendas, los que están en producción y las materias primas que se van a emplear, entre otras).

Gasto público: es el gasto realizado por el sector público:

- Las compras de los distintos niveles de la Administración pública, central, autonómica y local. Incluye los salarios de funcionarios y las materias primas, entre otras.

- Servicios: como los que prestan los legisladores y los policías.

Exportaciones netas: son la diferencia entre las exportaciones e importaciones de bienes y servicios.

El enfoque de la producción: se calcula como el valor añadido que se genera a medida que se transforma el bien o el servicio en las diferentes ramas de la actividad económica, que son:

- **Agricultura, ganadería, silvicultura y pesca.**
- **Industria.**
- **Construcción.**
- **Servicios.**

El valor añadido es la diferencia entre el precio de venta de un bien, sin tener en cuenta los impuestos indirectos, y el coste de los bienes intermedios adquiridos para su producción. En el caso de una empresa, se define como el total de ventas menos el valor de los bienes intermedios utilizados en la producción.

En el caso del panadero que produce barras de pan que son vendidas por otra empresa, esto se resume en el Cuadro 1.1 (Mochón (2006, p.10)):

Empresa	Coste productos intermedios	Precio venta	Valor añadido
Agrícola	0 €	5 €	5-0=0€
Harinera	5€	15€	15-5=10€
Panadería	15€	25€	25-15=10€
Distribuidora	25€	36€	36-25=11€
			Total= 36€

Cuadro 1.1: Ejemplo del cálculo desde el enfoque de la producción.

El enfoque de los ingresos, renta o costes: es el resultado de sumar:

- Salarios y Cotizaciones de la Seguridad Social.
- Intereses, rentas o alquileres.
- Impuestos indirectos y subvenciones.
- Depreciación o amortización.
- Beneficios.

1.1.2. PIB nominal y real

El PIB mide el gasto total de bienes y servicios en todos los mercados de la economía. Si de un año para otro hay un crecimiento, puede deberse a dos circunstancias:

- La economía está produciendo más bienes y servicios.
- Esos bienes y servicios se están vendiendo a precios más altos.

Una manera de averiguar si este aumento se debe a que la economía está produciendo más bienes y servicios es calcular el PIB con los precios fijados a un año elegido, resultando el PIB a precios constantes. Con esta magnitud se podrá estudiar la variación en cuanto a volumen de producción, convirtiéndose en un buen medidor de la situación económica. Se define por tanto:

- El PIB a precios corrientes o nominal se mide con los precios existentes cuando se realiza la producción.
- El PIB a precios constantes o real se mide con los precios existentes en un año base específico. Permite conocer la evolución del PIB descontadas las variaciones en los precios.

Por poner un ejemplo, imagínese que todos los bienes y servicios producidos por una economía son las barras de pan producidas por un panadero y los datos de la evolución de la producción y de los precios en los tres últimos años se muestran en el Cuadro 1.2.

Año	Producción	Precio unidad
2015	200	1€
2016	250	1.05€
2017	225	1.15€

Cuadro 1.2: Producción y precios de la barra de pan para el ejemplo del PIB real y del PIB nominal.

El PIB real y el PIB nominal del año 2017 tomando como base el año 2015 son:

$$\text{PIB}_{2017}^{nominal} = 225 \cdot 1.15 = 258.75€ \qquad \text{PIB}_{2017}^{real} = 225 \cdot 1 = 225€$$

1.1.3. La serie del PIB

La serie trimestral del PIB con la que se suele trabajar y con la que se lleva a cabo este estudio es el PIB a precios constantes o PIB real, cuya frecuencia es trimestral. Para poder analizar la evolución a lo largo del tiempo lo que se hace es transformar esta cantidad monetaria en un índice con base el año 2010[1].

Cuando en prensa se lee "...El PIB crece en Galicia un 3.2 respecto al año pasado..." o se hacen estudios y predicciones, no se emplea el índice bruto, sino que se trabaja con el índice corregido de calendario y estacionalidad y con un tratamiento de consistencia transversal y temporal, como se estudiará en el Capítulo 4. Ésta es la serie final, PIB de Galicia en índice y corregido, con la que se trabaja y es publicada por el IGE con un desfase de 50 días. El INE también publica esta serie para España y hace un reparto anual en función de las comunidades autónomas; por lo que publica un PIB anual para Galicia.

1.2. ABANCA y el seguimiento de la economía gallega

Desde el departamento de Planificación Estratégica y PMO de ABANCA se lleva a cabo un análisis y seguimiento semanal del entorno macroeconómico, siendo uno de las entradas básicas del proceso de planificación estratégica desarrollado en la Entidad. Con carácter permanente, se efectúa un estrecho seguimiento de las principales variables de entorno enfocado en España y en Galicia, por ser esta última el área donde ABANCA es la empresa líder en el sector bancario.

[1]Los números índices son instrumentos estadísticos utilizados para describir caracteres que varían en el tiempo o en el espacio y estudiar su evolución.

Con relación a España y a Galicia, se efectúa un análisis de un amplio panel de variables publicadas por los distintos organismos oficiales, fundamentalmente el Instituto Nacional de Estadística (INE), Banco de España e Instituto Galego de Estatística (IGE). Se trata de un panel de indicadores que permite un seguimiento de la actividad general a través de la evolución de áreas de especial interés para la actividad financiera. En particular, el Consumo Privado (con variables como matriculaciones de turismos, ventas en comercio minorista o empleo), el sector Industrial (Índice de Producción, Pedidos, etc.), Construcción (a través de hipotecas, ventas, visados, etc.), Servicios (cifras de negocio, pernoctas o empleo) y el Comercio Exterior. Seguimiento mensual que es complementado por otras variables de actividad como el propio PIB, IPC, sentimiento económico o el seguimiento diario de las principales magnitudes de mercados (bolsa, tipos de interés, primas de riesgo, petróleo, etc.).

Con carácter semanal se presenta en el Comité de Dirección un resumen con las principales implicaciones derivadas de los indicadores publicados en el periodo. Además, mensualmente se presenta un informe de entorno macro al Consejo de Administración, en el que se incorporan:

- Análisis de la coyuntura Mundial, de España y Galicia.
- Comportamiento de los mercados.
- Análisis de la relación entre el comportamiento mensual de las variables de entorno macro y la evolución del negocio financiero general del sector y de ABANCA en particular.

El seguimiento macroeconómico se complementa con el desarrollo de modelos estadísticos que permiten proyectar un conjunto de indicadores macroeconómicos. Estas proyecciones, una vez sometidas a revisión experta y contraste con otras previsiones publicadas por distintas casas de análisis y organismos internacionales (FUNCAS, FMI, Comisión Europea, Banco de España, AFI), constituyen el cuadro macroeconómico de referencia en la entidad. Así, estas proyecciones se utilizan como input en diversos procesos internos: desarrollo del Plan Estratégico, Presupuestación, ICAAP, RAF o Stress Test. Además, el cuadro macroeconómico generado se comunica al área de Riesgos, de manera que los modelos desarrollados para estimación de PD o LGD utilizan también como input esta proyección macroeconómica. Este planteamiento garantiza la consistencia de todos los procesos del banco, en cuanto a hipótesis macroeconómicas se refiere.

En el contexto anterior, surge la necesidad de un modelo para el PIB de Galicia motivado principalmente por dos razones. La primera, por ser el principal indicador de la situación económica de la región; y la segunda, por tener un desfase de publicación de aproximadamente 50 días desde el fin del trimestre. El IGE proporciona los datos de manera trimestral, mientras que el INE solamente proporciona el indicador desagregado por CCAA anualmente. En ocasiones, los valores anuales de estos dos Institutos difieren en gran medida, siendo esto un problema a la hora de analizar la situación económica. Por ejemplo, el IGE situa el crecimiento de la economía del año 2015 en un $2,4$, mientras que el INE da un crecimiento para este año de un $4,2$. En base a un criterio experto y dado que el IGE publica este agregado de manera trimestral y desglosado en función de los distintos enfoques, se ha elegido para trabajar la serie publicada por el IGE.

1.3. Conceptos previos

Definición 1.1. Un proceso estocástico es un conjunto de variables aleatorias $\{X_t\}_{t\in C}$ definidas sobre el mismo espacio de probabilidad.

El índice t que describe la sucesión de variables aleatorias no tiene por qué tener una interpretación concreta. En el contexto de series temporales, que es en el que se desarrolla este trabajo, el índice t da cuenta del periodo temporal al que pertenece la variable, $t \in \mathbb{Z}$. Por simplicidad, se empleará la notación $\{X_t\}$ para referirse a procesos estocásticos en el contexto de las series temporales $\{X_t\}_{t\in\mathbb{Z}}$.

Definición 1.2. Una serie temporal $\{x_1, x_2, \ldots, x_T\}$ es una realización parcial de un proceso estocástico $\{X_t\}$. Es decir, cada dato observado x_t de la serie temporal es una realización (una muestra de tamaño 1) de la variable aleatoria X_t.

Cuando se observan los valores de una serie de tiempo $\{x_1, \ldots, x_T\}$, lo que se está viendo son los valores muestrales del conjunto de variables aleatorias $X_1, \ldots, X_T$ que forman parte del proceso estocástico que genera los datos. En la mayoría de las ocasiones, es imposible determinar cuál es el proceso estocástico que está detrás de los datos obtenidos, siendo la tarea principal del investigador encontrar un proceso que sea compatible con los datos que se han obtenido.

Comentario 1.3. En la literatura, es frecuente que se emplee el término serie temporal tanto para referirse a los datos observados como para referirse al proceso del que provienen los datos. Esto no es un problema ya que, dependiendo del contexto, es sencillo saber a qué término se está haciendo referencia.

Asociado a cualquier proceso estocástico $\{X_t\}$ existen medidas estadísticas características para describirlo.

Definición 1.4. Sea $\{X_t\}$ un proceso estocástico, se definen:

- La función de medias del proceso: proporciona la esperanza de las distribuciones marginales de cada variable X_t:
$$\mu_t = \mathbb{E}(X_t).$$
Un proceso se dice que es estable en media si la función de medias es una constante, es decir, es independiente de t.

- La función de varianzas del proceso: proporciona la varianza de cada una de las variables que conforman el proceso:
$$\sigma_t^2 = Var(X_t) = \mathbb{E}((X_t - \mu_t)^2).$$
Un proceso se dice estable en varianza si la función de varianzas es una constante, es decir, es independiente de t.

- La funcion de autocovarianzas del proceso: proporciona la covarianza entre dos variables del proceso:
$$\gamma(s,t) = \mathrm{Cov}(X_s, X_t) = \mathbb{E}((X_s - \mu_s)(X_t - \mu_t)).$$
Mide el grado de dependencia lineal entre las variables X_s y X_t. En particular, se tiene que $\gamma(t,t) = \mathrm{Var}(X_t) = \sigma_t^2$.

- La función de autocorrelaciones simples: es una estandarización de la función de autocovarianzas:
$$\rho(s,t) = \frac{\gamma(s,t)}{\sigma_s \sigma_t}.$$
Se suele denotar ACF (en ingés, autocorrelation function), y es una medida de la capacidad de predecir X_t a partir de la variable X_s.

- La función de autocorrelaciones parciales entre dos variables: sirve para medir su dependencia lineal una vez que se ha extraído de cada una de ellas el efecto lineal de las variables que están comprendidas entre ellas:
$$\alpha(s,t) = \frac{Cov\left(X_s - \hat{X}_s^{(s,t)}, X_t - \hat{X}_t^{(s,t)}\right)}{\sqrt{\mathrm{Var}\left(X_s - \hat{X}_s^{(s,t)}\right)\mathrm{Var}\left(X_t - \hat{X}_t^{(s,t)}\right)}},$$
donde $\hat{X}_j^{(s,t)}$ es el mejor predictor lineal construído a partir de las variables comprendidas entre los instantes s y t. Se suele denotar PACF (en ingés, partial autocorrelation function).

Se definen a continuación los operadores habituales que se emplearán en el Capítulo 2.

Definición 1.5. Se define el operador retardo de orden k, denotado por B^k, como:

$$B^k X_t = X_{t-k}.$$

Definición 1.6. Se define el operador diferencia regular de orden d, denotado por ∇^d, como:

$$\nabla^d = (1-B)^d.$$

Definición 1.7. Se define el operador diferencia estacional de orden D y periodo estacional s, denotado por ∇_s^D, como:

$$\nabla_s^D = (1-B^s)^D.$$

Definición 1.8. Un proceso $\{X_t\}$ se dice que tiene periodo estacional s si el valor medio no es constante pero sigue un patrón cíclico de periodo s, $\mathbb{E}(X_t) = \mathbb{E}(X_{t+s})$.

Por ejemplo, en la serie mensual de ventas de helados, cabe esperar que la media de enero no sea igual a la de agosto; pero no es descabellado pensar que las ventas medias de los meses de agosto sean muy parecidas.

Procesos estacionarios

Como se ha visto, las series temporales que se observan y estudian son realizaciones de un proceso estocástico y, aunque hay ocasiones en las que el proceso estocástico podría repetirse, esto no ocurre en la mayoría de las áreas de estudio. En resultados generados en laboratorio, el experimento que da cuenta del proceso podría repetirse y de esta manera se obtendrían varias muestras del proceso. Esto no ocurre en las series económicas ni, en general, con las series temporales reales; sea $\{X_t\}$ el proceso estocástico que da cuenta del PIB de Galicia. La única muestra de la que se dispone es la serie observada desde 1995; siendo imposible volver atrás en el tiempo y generar otra muestra. Observada la serie, se sabe que hay un proceso estocástico por detrás que la ha generado, pero no se puede obtener más información de este proceso.

Dado el proceso estocástico $\{X_t\}$, en la mayoría de los casos sólo se dispondrá de una realización $\{x_1, x_2, \ldots, x_T\}$ de una pequeña parte del proceso, $\{X_t\}_{t\in\{1\ldots T\}}$ (ni siquiera de todo el proceso). Esto es equiparable, en el contexto de la inferencia clásica, a una situación en la que se tienen n variables independientes con distintas distribuciones y sólo un valor muestral de cada una de ellas; en esta situación resultaría imposible estimar algún parámetro. Por este motivo, para poder estimar las funciones descritas en la sección anterior, es necesario imponer alguna condición al proceso estocástico: la estacionariedad. La idea principal es que la ley de probabilidad que gobierna el proceso no varía con el tiempo, es decir, que el proceso se encuentra en algún tipo de equilibrio a lo largo del tiempo.

Definición 1.9. Una serie temporal $\{X_t\}$ es estacionaria en sentido estricto si la distribución conjunta de cualquier colección de variables que lo forman

$$\{X_{t_1}, X_{t_2}, \ldots, X_{t_k}\}$$

es idéntica a la de la colección

$$\{X_{t_1+h}, X_{t_2+h}, \ldots, X_{t_k+h}\},$$

para cualquier colección de instantes $t_1, \ldots, t_k$ y cualquier retardo $h \in \mathbb{Z}$.

Para el caso $k = 1$, esto se traduce en que las variables aleatorias X's son marginalmente idénticamente distribuidas y, por lo tanto, tienen la misma media y la misma varianza constante en el tiempo. Si el proceso estocástico a estudiar fuera la temperatura de cada hora del día, esta condición impondría que la probabilidad de que haya una temperatura negativa es la misma a las 4 de la mañana que a las 12 del mediodía.

La condición de estacionariedad estricta de la Definición 1.9 es demasiado fuerte para la mayoría de los contextos de estudio; además de ser muy complicada de verificar a partir de un único conjunto de datos. Por este motivo, en lugar de imponer condiciones a todas las distribuciones marginales posibles, se emplea una versión menos restrictiva que impone condiciones solo sobre los dos primeros momentos del proceso.

Definición 1.10. Una proceso estocástico $\{X_t\}$ es estacionario en sentido débil si:

- La función de media μ_t es constante (no depende del tiempo): $\mu_t = \mu, \forall t$.
- La función de varianza σ_t^2 es constante (no depende del tiempo): $\sigma_t^2 = \sigma^2, \forall t$.
- La función de autocovarianzas entre dos variables solo depende de la separación entre las variables, es decir, $\gamma(t, t+k) = \gamma(k)$.

En este caso, la función de autocovarianzas sólo depende de la separación entre las variables, convirtiéndose en una función definida para cada retardo k, $\gamma(k)$. Esto se hereda en las funciones de autocorrelaciones simples y parciales, definidas para cada retardo:

$$\rho(k) = \rho(t, t+k)$$
$$\alpha(k) = \alpha(t, t+k),$$

denominándose coeficientes de estas funciones a los valores que toman en función de los retardos.

A lo largo de este trabajo se emplea el término estacionario para referirse a la estacionariedad en sentido débil, siendo un proceso estacionario aquel con función de medias y varianza estables y cuya función de covarianzas depende únicamente de la separación entre las variables.

Un proceso estocástico particular es el de ruido blanco, que es un conjunto de variables aleatorias incorreladas para todos los retardos, de media 0 y con varianza finita σ_w^2, definido como sigue:

Definición 1.11. Un proceso estocástico $\{w_t\}$ se dice ruido blanco si

$$\mu_t = 0 \ , \ \sigma_t^2 = \sigma_w^2 \text{ y } \gamma(s,t) = \begin{cases} \sigma_w^2 \text{ si } s = t \\ 0 \text{ si } s = t \end{cases} \quad \forall s, t.$$

De acuerdo con Wei (2006, p.16), un proceso estacionario $\{X_t\}$ está caracterizado por la media del proceso, μ, la varianza σ^2 y las funciones de autocorrelaciones simples y parciales, ρ_k y α_k. En esta situación, es posible aprovechar la información de una realización muestral $\{x_1, \ldots, x_T\}$ para estimar todos estos parámetros:

Definición 1.12. : A continuación, se definen:

- Media muestral :
$$\bar{x} = \frac{\sum_{t=1}^{T} x_t}{T}.$$
- Funcion de autocovarianzas muestrales:
$$\hat{\gamma}(k) = \frac{\sum_{t=1}^{T-k} (x_t - \bar{x})(x_{t+k} - \bar{x})}{T}.$$
- Función de autocorrelaciones simples muestrales:
$$\hat{\rho}(k) = \frac{\hat{\gamma}_k}{\hat{\gamma}_0}.$$

- Función de autocorrelaciones parciales muestrales: $\hat{\alpha}_k = \hat{\alpha}_{kk}$, siendo $\hat{\alpha}_{kk}$ el estimador mínimo cuadrático en la regresión:

$$x_t = \alpha_{k0} + \alpha_{k1}x_{t-1} + \cdots + \alpha_{kk}x_{t-k} + \epsilon.$$

A menudo se representan gráficamente la funciones de autocorrelación simple y parcial muestrales en función de los retardos. En el caso de que el proceso sea ruido blanco, la siguiente propiedad proporciona, bajo ciertas condiciones, la distribución de los coeficientes de autocorrelación muestrales.

Propiedad 1.13. Bajo ciertas condiciones, si el proceso $\{X_t\}$ es ruido blanco, para un tamaño muestral T grande se verifica:

$$\hat{\rho}(k) \sim N\left(0, \frac{1}{\sqrt{n}}\right).$$
$$\hat{\alpha}(k) \sim N\left(0, \frac{1}{\sqrt{n}}\right).$$

En una serie que ha sido generada por un proceso de ruido blanco aproximadamente el 95 % de los coeficientes deberían estar en el intervalo $(\frac{-1.96}{\sqrt{n}}, \frac{1.96}{\sqrt{n}})$. Esta propiedad es realmente útil para contrastar si una serie de tiempo proviene de un proceso de ruido blanco.

Formas de representar los procesos

En el contexto de series de tiempo, existen diversas representaciones para describir los procesos estocásticos. A continuación se describen tres de ellas.

Definición 1.14. Una serie de tiempo $\{X_t\}$ se dice lineal si puede ser escrita como una "combinación lineal" de ruido blanco:

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

Toda serie lineal es estacionaria, por ser una combinación de un proceso estacionario.

Definición 1.15. Una serie de tiempo $\{X_t\}$ se dice causal si puede ser escrita como:

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

Definición 1.16. Una serie de tiempo $\{X_t\}$ se dice invertible si puede ser escrita como:

$$X_t = c + a_t + \pi_1 X_{t-1} + \pi_2 X_{t-2} + \ldots, \text{ con } \sum_{i=0}^{\infty} |\pi_i| < \infty.$$

El Teorema de descomposición de Wold (Wold (1938)), demuestra que cualquier proceso estacionario que no contenga componentes deterministas puede escribirse como un proceso lineal, es decir, como combinación lineal de infinitos errores aleatorios de media cero, misma varianza finita e incorrelados. Por tanto, la clase de los procesos lineales puede ser considerada un marco general para el estudio de los procesos estacionarios. Sin embargo, esta representación tiene un problema que la hace poco operativa: se necesitan infinitos parámetros. Muchos de los modelos incluidos en la metodología Box–Jenkins (Box y Jenkins (1970)) son casos particulares de esta representación general con pocos parámetros, y se estudiarán en el Capítulo 2.

La notación empleada a lo largo de este trabajo es la siguiente:

- $\{X_t\}$ es el proceso generador de la serie de tiempo.
- $x_1, \ldots, x_T$ es la serie observada.
- $\{w_t\}$ es el proceso de ruido blanco. Se supondrá siempre que w_t es independiente de $X_{t-1}, X_{t-2,\ldots}$.
- $A - Qi$ denotará el trimestre i con $i \in \{1, 2, 3, 4\}$ del año A (2007-Q1 es el primer trimestre del año 2007).
- $A - Mi$ denotará el mes i con $i \in \{1, \ldots, 12\}$ del año A (2007-M1 es enero del año 2007).

Capítulo 2

Metodología Box-Jenkins

La particularidad de las series de tiempo es la estructura de dependencia intrínseca que gobierna su evolución y que será lo que se trate de modelizar. Conocer cómo se comporta el proceso será clave a la hora de predecir valores futuros, siendo ésta la tarea principal del investigador.

En este capítulo se va a describir la metodología Box-Jenkins, en la que se engloba una clase de procesos a partir de los cuales se puede modelizar la evolución de gran cantidad de series de tiempo. Esta metodología consta de 4 etapas. La primera de ellas es la identificación; en las Secciones 2.1 y 2.2 se describen una clase de procesos susceptibles de haber generado la serie observada. Mediante las particularidades de cada uno de estos procesos, que se detallan en la Sección 2.3, es posible identificarlos a partir de las funciones de autocorrelación simple y parcial muestrales. En la Sección 2.4 se estudian dos métodos para estimar el modelo seleccionado. Una vez ajustado, es necesario verificar que se cumplen las hipótesis de especificación del modelo, que se detallan en la Sección 2.5. En ocasiones existe más de un modelo válido como candidato a ser generador de la serie; por este motivo en la Sección 2.6 se define un criterio para elegir qué modelo es mejor. Por último, si el modelo es válido, lo deseable será hacer predicciones, como se estudia en la Sección 2.7.

Las referencias bibliográficas empleadas para construir este capítulo han sido Peña (2010), Shumway y Stoffer (2011), Cryer y Chan (2008), Tsay (2005) y Aneiros (2016).

2.1. Modelos para series estacionarias

Definición 2.1. Un modelo autorregresivo de orden p, AR(p), tiene la forma:

$$X_t = \alpha + \phi_1 X_{t-1} + \phi_2 X_{t-1} + \cdots + \phi_p X_{t-p} + w_t, \text{con } \alpha = \mu(1 - \phi_1 - \phi_2 - \cdots - \phi_p),$$

donde $\{X_t\}$ es un proceso estacionario de media μ y $\phi_1, \phi_2, \ldots, \phi_p$ son constantes ($\phi_p \neq 0$). Además, $\{w_t\}$ es un proceso de ruido blanco con varianza σ_w^2.

Estos modelos generalizan los modelos de regresión lineal para modelizar la dependencia entre el valor presente de la serie y los valores en los instantes anteriores.

Definición 2.2. Dado un proceso AR(p) se definen:

- Operador autorregresivo: $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$, donde B es el operador retardo.
- Polinomio AR(p) característico: $\phi(z) = 1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p$.

A partir del operador autorregresivo se puede escribir la ecuación de un proceso $AR(p)$ de una manera más compacta:

$$\phi(B)X_t = \alpha + w_t.$$

Nótese que la estacionariedad de un proceso $AR(p)$ es equivalente a que las raíces del polinomio característico tengan módulo distinto de uno. De acuerdo con la Definición 1.16, se tiene que un proceso $AR(p)$ es siempre invertible sin más que tomar $\pi_i = \phi_i$ con $i = 1, \ldots, p$ y nulos los demás. Además, una condición suficiente para que sea causal es que las raíces del polinomio $AR(p)$ característico estén fuera del círculo de unidad.

La función de autocorrelación simple (ACF) de un proceso AR(p) tienen una estructura muy compleja y no es tarea sencilla identificar el proceso a través de ella. En cambio, la función de autocorrelación parcial tiene una estructura que lo caracteriza: el último coeficiente de autocorrelación parcial no nulo es el $p-$ésimo, que además coincide con el parámetro ϕ_p. Los cálculos necesarios para obtener estas conclusiones pueden consultarse en Peña (2010, p.130 y pp.134-135).

Definición 2.3. Un modelo de medias móviles de orden q, $MA(q)$, está definido como:

$$X_t = \mu + w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + \cdots + \theta_q w_{t-q},$$

donde μ es la media del proceso $\{X_t\}$ y $\theta_1, \theta_2, \ldots, \theta_q$ son constantes ($\theta_q \neq 0$). Se supone que $\{w_t\}$ es un proceso de ruido blanco con varianza σ_w^2.

Definición 2.4. Dado un proceso $MA(q)$ se definen:

- Operador de medias móviles: $\theta(B) = 1+\theta_1 B+\theta_2 B^2+\cdots+\theta_q B^q$, donde B es el operador retardo.
- Polinomio MA(q) característico: $\theta(z) = 1 + \theta_1 z + \theta_2 z^2 + \cdots + \theta_q z^q$.

A partir del operador de medias móviles se puede escribir la ecuación de un modelo $MA(q)$ de una manera más compacta como:

$$X_t = \mu + \theta(B) w_t.$$

Un proceso $MA(q)$ siempre es siempre lineal y causal, basta tomar en las definiciones 1.14 y 1.15 $\psi_i = \theta_i$ con $i = 1, \ldots, q$ y nulos los demás. Además, una condición suficiente para que sea invertible es que las raíces del polinomio $MA(q)$ característico estén fuera del círculo unidad.

En un modelo $MA(q)$ el último coeficiente de autocorrelación simple no nulo es el $q-$ésimo, permitiendo esta particularidad identificar el orden de un proceso de medias móviles. Por el contrario, la función de autocorrelación parcial tiene una estructura compleja, muy similar a la que tenía la función de autocorrelaciones simples del proceso $AR(p)$, existiendo una dualidad entre estos dos tipos de modelos. Esto puede verse con más detalle en Peña (2010, pp.155-160).

Los modelos que se describen a continuación son el resultado de combinar los modelos anteriormente descritos y pueden ser interpretados como un $AR(p)$ que se completa con un $MA(q)$ o al revés.

Definición 2.5. Un proceso ARMA de órdenes p y q, $ARMA(p, q)$, tiene la forma:

$$X_t = \alpha + \phi_1 X_{t-1} + \phi_2 X_{t-1} + \cdots + \phi_p X_{t-p} + w_t + \theta_1 w_{t-1} + \theta_2 w_{t-1} + \cdots + \theta_q w_{t-q},$$

con $\alpha = \mu(1-\phi_1-\phi_2-\cdots-\phi_p)$, donde $\{X_t\}$ es un proceso estacionario con media μ, $\phi_p \neq 0$ y $\theta_q \neq 0$. Se supone que $\{w_t\}$ es un proceso de ruido blanco con varianza σ_w^2.

Es claro que $ARMA(p, 0) = AR(p)$ y que $ARMA(0, q) = MA(q)$. Además, empleando los operadores autorregresivos y de medias móviles se puede escribir el modelo $ARMA(p, q)$ de una forma compacta como:

$$\phi(B) X_t = \alpha + \theta(B) w_t.$$

La condición de estacionariedad de un proceso $ARMA(p, q)$ es equivalente a que las raíces del polinomio $AR(p)$ definido con los coeficientes autoregresivos del modelo $ARMA(p, q)$ no tenga raíces de módulo uno. Las condiciones de causalidad e invertibilidad de un proceso $ARMA(p, q)$ son las mismas que las de los procesos $AR(p)$ y $MA(q)$, respectivamente. Además, las funciones de autocorrelación simple y parcial son el resultado de superponer las de los dos procesos; motivo por el cual es una tarea complicada identificar un proceso $ARMA(p, q)$ a partir de estas funciones.

Los modelos $ARMA(p,q)$ estudiados forman parte de un amplio abanico de modelos empleados para modelizar la dependencia de un instante con los instantes inmediatamente anteriores (dependencia regular). Son comunes los procesos en los que lo esperado es que haya una dependencia entre observaciones de tiempo separadas un múltiplo del periodo estacional s. Una primera opción para modelizar este tipo de dependencia podría ser aumentar los órdenes p y q hasta el retardo deseado (un múltiplo del periodo estacional). Sin embargo, esto exigiría incluir una cantidad innecesaria de parámetros en el modelo[1]. Como alternativa, los modelos anteriormente estudiados se modifican para modelizar la dependencia estacional; en vez de tomar retardos k, con $k = 1, \ldots, p$ ó q, se toman retardos múltiplos del periodo estacional, $k = s, \ldots, ps$ ó qs. La construcción de los modelos es muy similar y por este motivo sólo se expone la ecuación en forma compacta de cada uno de ellos.

Definición 2.6. Un modelo autorregresivo estacional de orden P y periodo estacional s, $AR(P)_s$, tiene la forma:

$$\Phi(B^s)X_t = \alpha + w_t, \text{ con } \Phi(B^s) = (1 - \Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_P B^{sP}),$$

donde $\alpha = \mu(1-\Phi_1-\Phi_2-\cdots-\Phi_P)$. Siendo $\{X_t\}$ un proceso estacionario de media μ y $\Phi_1, \Phi_2, \ldots, \Phi_P$ constantes ($\Phi_P \neq 0$). Además, $\{w_t\}$ es un proceso de ruido blanco con varianza σ_w^2

Definición 2.7. Un modelo de medias móviles estacional de orden Q y periodo estacional s, $MA(Q)_s$, está definido como:

$$X_t = \mu + \Theta(B^s)w_t, \text{ con } \Theta(B^s) = (1 + \Theta_1 B^s + \Theta_2 B^{2s} + \cdots + \Theta_Q B^{sQ}),$$

donde μ es la media del proceso $\{X_t\}$ y $\Theta_1, \Theta_2, \ldots, \Theta_Q$ son constantes ($\Theta_Q \neq 0$). Se supone que $\{w_t\}$ es un proceso de ruido blanco con varianza σ_w^2.

Las condiciones de causalidad e invertibilidad de los procesos $AR(p)_s$ y $MA(q)_s$ coincididen con las de los $AR(p)$ y $MA(q)$ respectivamente. Con respecto a la estructura de las funciones de autocorrelación simple y parcial de cada proceso, también hay una analogía con los modelos regulares:

- Para un proceso $AR(P)_s$, el último coeficiente de autocorrelación parcial de orden un múltiplo de s no nulo es el $P-$ésimo. Los valores de los retardos no múltiplos del periodo estacional son nulos.
- Para un proceso $MA(Q)_s$, el último coeficiente de autocorrelación simple de orden un múltiplo de s no nulo es el $Q-$ésimo. Los valores de los retardos no múltiplos del periodo estacional son nulos.

Los modelos $AR(P)_s$ y $MA(Q)_s$ se combinan dando lugar a los modelos $ARMA(P,Q)_s$

Definición 2.8. Una serie de tiempo $\{X_t\}$ es un proceso $ARMA$ estacional de órdenes P y Q, y periodo estacional s, $ARMA(P,Q)_s$, si:

$$\Phi(B^s)X_t = \alpha + \Theta(B^s)w_t, \text{ con } \alpha = \mu(1 - \Phi_1 - \Phi_2 - \cdots - \Phi_P),$$

donde $\Phi(B^s)$ y $\Theta(B^s)$ son los operadores definidos anteriormente. Se supone que $\{X_t\}$ es un proceso estacionario de media μ y que $\{w_t\}$ es un proceso de ruido blanco con varianza σ_w^2.

Es posible combinar y modelizar a la vez tanto dependencia regular como estacional, dando lugar a los modelos multiplicativos.

Definición 2.9. Un modelo autoregresivo multiplicativo de periodo estacional s y órdenes p, q, P y Q, $ARMA(p,q) \times (P,Q)_s$ tiene la forma:

$$\phi(B)\Phi(B^s)X_t = \alpha + \theta(B)\Theta(B^s)w_t, \text{ con } \alpha = \mu(1 - \phi_1 - \phi_2 - \cdots - \phi_P)(1 - \Phi_1 - \Phi_2 - \cdots - \Phi_P),$$

donde $\{X_t\}$ es un proceso estacionario de media μ y $\phi(B)$, $\theta(B)$, $\Phi(B^S)$ y $\Theta(B^s)$ son los operadores ya definidos anteriormente. $\{w_t\}$ es un proceso de ruido blanco con varianza σ_w^2.

[1] Para añadir un retardo de dos periodos estacionales es necesario incorporar $2s$ parámetros en el modelo.

La estructura de la función de autocorrelación simple y parcial de este tipo de procesos se describe a continuación:

- En los retardos bajos $(1, 2, \ldots, [\frac{s}{2}])$ se observarán los coeficientes de autocorrelación de la parte regular.
- En los retardos múltiplos del periodo estacional (s,2s,...) se observarán los coeficientes de autocorrelación de la parte estacional.

En el Cuadro 2.1 re recopila la estructura de las funciones de autocorrelaciones simples y parciales de los modelos estudiados hasta ahora.

Proceso	ACF $\rho(k)$	PACF $\alpha(k)$
$AR(p)$	Muchos coeficientes no nulos	Último coeficiente no nulo el $p-$ésimo
$MA(q)$	Último coeficiente no nulo el $q-$ésimo	Muchos coeficientes no nulos
$ARMA(p,q)$	Muchos coeficientes no nulos	Muchos coeficientes no nulos*
$AR(P)_s$	Muchos coeficientes no nulos*	Último coeficiente no nulo el $sP-$ésimo*
$MA(Q)_s$	Último coeficiente no nulo el $sQ-$ésimo*	Muchos coeficientes no nulos*
$ARMA(P,Q)_s$	Muchos coeficientes no nulos*	Muchos coeficientes no nulos*

Cuadro 2.1: Resumen de los procesos en base a la ACF y PACF. Con * se indica que los coeficientes no múltiplos de P (para el AR), de Q (para el MA) o de P (parte AR del ARMA) o Q (parte MA del ARMA), son nulos

2.2. Modelos para series no estacionarias

En la sección anterior se ha estudiado un abanico muy amplio de procesos candidatos a ser generadores de series estacionarias. Sin embargo, es habitual que las series con las que se trabaja no sean estacionarias, invalidando las hipótesis de los modelos anteriores. Los motivos más usuales de falta de estacionariedad en series temporales son:

1. Heterocedasticidad: ocurre cuando la variabilidad de la serie no es constante con el tiempo.
2. Tendencia: ocurre cuando el nivel de la serie no es estable en el tiempo.
3. Componente estacional: cuando la media de las observaciones no es constante pero tiene un patrón cíclico.

La heterocedasticidad se detecta en el gráfico de la serie y, bajo ciertas condiciones, puede ser corregida mediante las llamadas transformaciones Box-Cox (Box y Cox (1964)), definidas como:

$$\begin{cases} \frac{x^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log(x) & \lambda = 0. \end{cases}$$

Véase Cryer y Chan (2008, p.101) para más detalle acerca del parámetro λ y de esta transformación.
Para corregir la tendencia y/o estacionalidad existen dos maneras de proceder:

Método 1: La descomposición clásica, en la que la tendencia y/o componente estacional son estimadas y extraídas.

Método 2: Muchos procesos no estacionarios se convierten en estacionarios mediante la aplicación de diferencias regulares de orden d o de diferencias estacionales de periodo estacional s y de orden D, corrigiendo de esta manera la tendencia o la estacionalidad.

En el Capítulo 4 se detalla la metodología TRAMO-SEATS para emplear el Método 1. Una vez que se han extraído de la serie estas componentes, si la serie es estacionaria es susceptible de haber sido generada por alguno de los modelos definidos en la Sección 2.1.

En caso de aplicar el Método 2, la metodología detallada en la Sección 2.1 se amplía dejando hueco a aquellos procesos que no son estacionarias pero que, mediante la aplicación de diferencias regulares y estacionales de órdenes pertinentes, pasan a serlo.

En el caso de que la serie presente tendencia y sea necesario aplicarle una diferencia regular de orden d para que sea estacionaria, se definen los modelos $ARIMA(p,d,q)$, que no es otra cosa que un modelo $ARMA(p,q)$ sobre la serie diferenciada.

Definición 2.10. Un proceso $\{X_t\}$ se dice que es un $ARIMA(p,d,q)$ si

$$\nabla^d X_t = (1-B)^d X_t \text{ es un proceso } ARMA(p,q).$$

El modelo se puede escribir como

$$\phi(B)(1-B)^d X_t = \alpha + \theta(B)w_t, \text{ con } \alpha = \mu(1-\phi_1-\cdots-\phi_p), \text{ siendo } \mu = \mathbb{E}(\nabla^d X_t).$$

En el caso de que la serie presente estacionalidad y sea necesario aplicarle una diferencia estacional de periodo estacional s y de orden D para que sea estacionario, surgen los procesos $ARIMA(P,D,Q)$.

Definición 2.11. Un proceso $\{X_t\}$ se dice que es un $ARIMA(P,D,Q)_s$ si:

$$\nabla_s^D X_t = (1-B^s)^D X_t \text{ es un proceso } ARMA(P,Q)_s.$$

El proceso puede escribirse como:

$$\Phi(B^s)(1-B^s)^D X_t = \alpha + \Theta(B^s)w_t, \text{ con } \alpha = \mu(1-\Phi_1-\cdots-\Phi_P), \text{ siendo } \mu = \mathbb{E}(\nabla_s^D X_t).$$

En caso de que la serie presente tendencia y estacionalidad, será necesario aplicarle d diferencias regulares y D diferencias estacionales. Surgen así los procesos $ARIMA$ multiplicativos estacionales.

Definición 2.12. Un proceso $\{X_t\}$ se dice que es un $ARIMA(p,d,q)\times(P,D,Q)_s$ si

$$\nabla^d\nabla_s^D X_t = (1-B)^d(1-B^s)^D X_t \text{ es un proceso } ARMA(p,q)\times(P,Q)_s.$$

El modelo se puede escribir como

$$\phi(B)\Phi(B^s)(1-B)^d(1-B^s)^D X_t = \alpha + \theta(B)\Theta(B^s)w_t,$$

con $\alpha = \mu(1-\phi_1-\cdots-\phi_p)(1-\Phi_1-\cdots-\Phi_P)$, donde $\mu = \mathbb{E}(\nabla^d\nabla_s^D X_t)$

2.3. Identificación

Observada una serie temporal, la tarea del investigador es comprobar si ha podido ser generada por alguno de los modelos anteriormente descritos.

En caso de que la que la serie no sea estacionaria, el primer paso será transformarla para que lo sea, bien por el método 2, eligiendo los órdenes d y D; o por el método 1, en cuyo caso se tendrá una nueva serie $z_1,\ldots,z_T$ y se toma $d=0$ y $D=0$.

Siguiendo las indicaciones del Cuadro 2.1 sería posible identificar el proceso en base a los coeficientes de correlación simple y parcial de la serie diferenciada. Sin embargo, esto nunca se va a poder hacer ya que jamás se dispondrá de los coeficientes teóricos de una serie observada. Habrá que emplear los coeficientes muestrales y la siguiente proposición, que puede verse con más detalle en Cryer y Chan (2008, Cap.6).

Proposición 2.13. *Sea $\{x_1, \ldots, x_T\}$, entonces, bajo ciertas condiciones generales y asumiendo que T es grande:*

- *si proviene de un proceso $AR(p)$:*

$$\hat{\alpha}_k \approx N\left(0, \frac{1}{\sqrt{T}}\right), \forall k > p.$$

- *si proviene de un proceso $MA(q)$:*

$$\hat{\rho}_k \approx N\left(0, \frac{\sqrt{1 + 2(\rho_1^2 + \cdots + \rho_q^2)}}{\sqrt{T}}\right), \forall k > q,$$

En la práctica se sustituirá en el numerador cada ρ_i por $\hat{\rho}_i$.

Esta proposición proporciona la distribución muestral de los coeficientes de autocorrelación muestral bajo procesos autorregresivos o de medias móviles y será empleada, en un formato gráfico con las zonas de aceptación y rechazo de que los coeficientes sean nulos, para la identificación:

- En el caso de modelos de dependencia regular, se hace uso de esta proposición y se elegirá un modelo $ARMA(p, 0)$ o $ARMA(0, q)$ para la serie diferenciada, en función de cuál sea el último retardo no nulo.
- En el caso de modelos de dependencia estacional, se hace uso de esta proposición teniendo en cuenta la estructura de las funciones de autocorrelación simple y parcial de los modelos estacionales. Se ajustará entonces un modelo $ARMA(P, 0)_s$ o $ARMA(0, Q)_s$ a la serie diferenciada, en función de cuál sea el último retardo múltiplo del periodo estacional no nulo.
- En el caso de modelos multiplicativos, se hace uso de esta proposición teniendo en cuenta la estructura de las funciones de autocorrelación simple y parcial de los modelos multiplicativos. Se ajustará entonces un modelo $ARMA(p, 0) \times ARMA(P, 0)_s$, $ARMA(p, 0) \times ARMA(0, Q)_s$, $ARMA(0, q) \times ARMA(P, 0)_s$ o $ARMA(0, q) \times ARMA(0, Q)_s$ a la serie diferenciada.

2.4. Estimación

A efectos de simplicidad, esta sección se hará para un modelo $ARMA(p, q)$. Una vez elegido el modelo susceptible de haber generado la serie de tiempo objeto de estudio $\{x_1, \ldots, x_T\}$, es necesario estimar los parámetros del mismo.

Los parámetros a estimar en un modelo $ARMA(p, q)$ son $\boldsymbol{\beta} = (\alpha, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q, \sigma_w^2)$ y se estudiarán dos métodos para ello: mínimos cuadrados y máxima verosimilitud. En la práctica se emplea el último método tomando como valores iniciales las estimaciones del primero.

Los residuos asociados a las estimación $\tilde{\boldsymbol{\beta}} = (\tilde{\alpha}, \tilde{\phi}_1, \ldots, \tilde{\phi}_p, \tilde{\theta}_1, \ldots, \tilde{\theta}_q, \tilde{\sigma}_w^2)$ son:

$$\hat{w}_t = x_t - \tilde{\alpha} - \tilde{\phi}_1 x_{t-1} - \cdots - \tilde{\phi}_p x_{t-p} - \tilde{\theta}_1 \hat{w}_{t-1} - \cdots - \tilde{\theta}_q \hat{w}_{t-q}. \tag{2.1}$$

Mínimos cuadrados y mínimos cuadrados condicionados

Los estimadores de mínimos cuadrados se obtienen minimizando la suma de residuos al cuadrado:

$$(\hat{\alpha}, \hat{\phi}_1, \ldots, \hat{\phi}_p, \hat{\theta}_1, \ldots, \hat{\theta}_q, \hat{\sigma}_w^2) = \hat{\boldsymbol{\beta}} = \arg\min \sum_{t=1}^{T} (\hat{w}_t)^2.$$

A la hora de estimar los parámetros se plantean dos problemas:

1. Si $p \neq 0$, entonces los valores $\hat{w}_1, \ldots, \hat{w}_p$ dependen de los valores no observados $X_0, \ldots, X_{1-p}$.

   Esto se puede solucionar fijando a 0 los p primeros valores de $\hat{w}_t$.

2. Si $q \neq 0$, entonces los valores w_j dependerán de los valores $w_{j-1}, \ldots, w_{j-q}$.

   Si $p > 0$, conocidos los valores $\hat{w_k}$ con $k = p+1, \ldots, p+1-q$ es posible construir recursivamente todos los $\hat{w_r}$ con $r = p+1, \ldots, T$.

Se recurre entonces al método de mínimos cuadrados condicionados, resultando la función a minimizar:

$$\begin{cases} \arg\min \sum_{t=p+1}^{T} \hat{w}_t^2 \\ \text{condicionada a } \hat{w}_p = \hat{w}_{p+1} = \cdots = \hat{w}_{p+1-q} = 0. \end{cases}$$

Cuando $q = 0$, el problema se reduce a uno de regresión lineal y no es necesario ningún método iterativo. Si $q > 0$ se recurre a un método iterativo de estimación no lineal.

Máxima verosimilitud

La estimación por máxima verosimilitud se basa en la función de densidad conjunta[2], tomando como valores fijos los datos y maximizándola respecto a los parámetros.

Los parámetros estimados son:

$$\hat{\boldsymbol{\beta}} = \arg\max_{\tilde{\boldsymbol{\beta}}} L_{x_1,\ldots,x_T}(\tilde{\boldsymbol{\beta}}), \tag{2.2}$$

siendo L es la función de verosimilitud:

$$L_{x_1,\ldots,x_T}(\tilde{\boldsymbol{\beta}}) = f_{\tilde{\boldsymbol{\beta}}}(x_1, \ldots, x_T),$$

con $f_{\tilde{\boldsymbol{\beta}}}$ la función de densidad conjunta de un vector aleatorio $(\tilde{X}_1, \ldots, \tilde{X}_T)$ asociado a un proceso $ARMA(p,q)$ de parámetros $\tilde{\boldsymbol{\beta}}$.

En el caso en que el proceso es gaussiano los estimadores resultantes son centrados, eficientes y siguen una distribución normal. En el caso en que los datos no son normales, los estimadores pierden la propiedad de eficiencia.

2.5. Validación

Ajustado un modelo, es necesario comprobar que se cumplen todas las hipótesis que se han especificado en la definición del mismo. Para los modelos estudiados en este capítulo esto implica comprobar que las innovaciones $\{w_t\}$ son un ruido blanco gaussiano, es decir:

1. $\mu_t = 0$.
2. $\sigma_t^2 = \sigma_w^2$.
3. $\gamma_w(s,t) = \begin{cases} \sigma_w^2 \text{ si } s = t \\ 0 \text{ si } s = t. \end{cases}$
4. $w_t \in N(0, \sigma_w)$.

[2]Para calcular la función de densidad conjunta, se está en una situación en la que los datos no son independientes. Para ello, se emplea de manera recursiva la propiedad de que $f(x,y) = f(x)f(y|x)$. Entonces se llega a que $f(x_1, \ldots, x_T) = f(x_1)f(x_2|x_1)\ldots f(x_T|x_{T-1}, \ldots, x_1)$, en vez de obtener una descomposición de densidades marginales como se tendría en un caso de independencia. Un desarrollo más detallado puede verse en Peña (2010, Sec.10.2).

Las 3 primeras condiciones son las más importantes, ya que de no verificarse alguna se descartaría que el proceso que genera las innovaciones fuese ruido blanco invalidando el modelo. La hipótesis de normalidad es deseable, ya que bajo normalidad la incorrelación implica independencia. Además los estimadores de máxima verosimilitud gozarán de mejores propiedades y se podrán construir intervalos de predicción.

Lo cierto es que estas condiciones no se pueden comprobar en las innovaciones ya que éstas son inobservables; disponiendo únicamente de los residuos del modelo. No obstante, lo esperado es que si el modelo ajustado es apropiado entonces los residuos cumplan estas hipótesis.

Métodos gráficos

Lo esperado es que los residuos sean estacionarios; por este motivo en el gráfico de los residuos a lo largo del tiempo no se deberían observar tendencias, componente estacional o una variabilidad no constante. La presencia de alguna de estas características invalidaría alguna de las hipótesis.

Los contrastes acerca de si los residuos provienen de un proceso de ruido blanco mediante las funciones de autocorrelación simple y parcial muestrales se estudian en un formato gráfico usando la Propiedad 1.13.

Por último un gráfico cuantil-cuantil permitirá comprobar la normalidad de los residuos.

Contraste de incorrelación

En el contraste anterior de incorrelación sólo se observan las correlaciones de una en una, pudiendo ver si cada una de ellas es nula. El contraste de Ljung-Box permite constatar si los h primeros coeficientes de las autocorrelaciones son cero, empleando el siguiente estadístico:

$$Q(h) = T(T+2)\sum_{j=1}^{h} \frac{\hat{p}_j^2}{T-j}.$$

La distribución de $Q(h)$ es una X^2_{h-j} con j el números de parámetros estimados (véase Propiedad 1.12). Por ejemplo, si se ha ajustado un modelo $ARMA(p,q) \times (P,Q)$ con constante, y se desea contrastar si los h_1 primeros coeficientes de autocorrelación son nulos se tiene que $Q(h_1) \sim \chi^2_{h_1-(p+q+P+Q+1)}$.

Contraste de media cero

Para comprobar si la esperanza de las innovaciones es cero, se contrastará si los residuos del modelo tienen media cero. Para ello, se aplicará el clásico contraste de media cero, siendo importante, para que este contraste sea válido, que los residuos sean incorrelados. Para más detalle puede consultarse Peña (2010, p.326).

Contraste de normalidad

Para contrastar la normalidad se emplean dos contrastes. El contraste de Jarque-Bera (Jarque y Bera (1987)) comprueba si los datos provienen de una distribución normal en base a los coeficientes de asimetría y de kurtosis. Otro contraste que se emplea para chequear la normalidad de los residuos es el ya conocido contraste de Shapiro-Wilks (Shapiro y Wilk (1965)).

2.6. Criterios de selección de modelos

Una serie es susceptible de haber sido generada por diferentes modelos válidos. Por este motivo es necesario algún criterio para poder dictaminar con qué modelo es mejor y con cuál quedarse. No es posible elegir un modelo que a la vez maximice la verosimilitud y que sea lo más sencillo posible en cuanto a número de parámetros, siendo estas las características más deseables.

En la práctica, se busca un compromiso entre estos dos aspectos a través de distintas medidas. La medida que se va a emplear en este trabajo es el BIC, que penaliza el exceso de parámetros más que otras medidas y tiende a dar modelos más sencillos.

$$BIC = -2\log(L(\hat{\boldsymbol{\beta}})) + k\log(T_1),$$

con k el número de parámetros, T_1 el número de observaciones de la serie una vez que se ha diferenciado, es decir, $T - d - sD$, y $\hat{\boldsymbol{\beta}}$ el vector con las estimaciones (Ecuación (2.2)).

2.7. Predicción

Se estudia a continuación como es la predicción para un modelo AR(p) y para un modelo $MA(q)$, siendo la predicción de un modelo $ARMA(p, q)$ una combinación de ambas predicciones y la extensión al caso no estacional sencilla; al igual que también lo es la extensión a la situación de no estacionariedad de los modelos $ARIMA$.

Observada la serie $\{x_1, \ldots, x_T\}$, se pretende predecir el valor de la variable aleatoria X_{T+h}, denominada predicción a horizonte h. Dicha predicción se denotará por $\hat{x}_{T+h} = \hat{x}_T(h)$.

El predictor óptimo que minimiza el error cuadrático medio de predicción es la esperanza condicionada al conjunto de información. Los cálculos y desarrollos pueden verse en Peña (2010, pp. 224-226).

$$\hat{x}_{T+h} = \hat{x}_T(h) = \mathbb{E}(X_T(h)|X_T, X_{T-1}, X_{T-2}, \ldots, X_1).$$

Predicción puntual

Ajustado un modelo AR(p), la predicción a horizonte h será:

$$\hat{x}_{T+h} = \hat{x}_T(h) = \hat{\alpha} + \hat{\phi}_1\hat{x}_{T+h-1} + \cdots + \hat{\phi}_1\hat{x}_{T+h-p},$$

donde $\hat{x}_t = x_t$ para $t \leq T$.

En el caso de un modelo $MA(q)$ para la predicción a horizonte h se necesitan los valores de $w_{T+h}, \ldots, w_{T+h-q}$. Los valores $w_{T+1}, \ldots, w_{T+h}$ son desconocidos y se pueden fijar a cero (que es la media del proceso $\{w_t\}$). Serán también necesarios los valores $w_1, \ldots, w_T$, de los que la muestra contiene información y existen varias maneras de explotar esta información:

- Empleando la invertibilidad del proceso: $X_t = c + w_t + \pi_1 X_{t-1} + \pi_2 X_{t-2} + \ldots$

  $$w_T = -c + X_T - \pi_1 X_{T-1} - \cdots - \pi_{T-1} X_1 - \ldots$$

  se toma $X_t = x_t$ para $1 \leq t \leq T$ y $X_t = \mu$ para $t \leq 0$.
- Fijar un valor inicial para w_1 y construir de manera recursiva los valores w_j, con $j = 2, \ldots, T$.
- Predecir w_T a través de una combinación de los valores muestrales.

Para predecir para un modelo $ARMA(p, q)$ se combinarán los dos procesos anteriores.

Intervalos de predicción

Si se verifica que el tamaño muestral T es grande y que $\{w_t\}$ es un proceso de ruido blanco gaussiano, entonces la distribución muestral del error de predicción $e_T(k) = X_{T+k} - \hat{X}_T(k)$ es

$$e_T(k) \approx N\left(0, \sigma_w^2(1 + \psi_1^2 + \cdots + \psi_{k-1}^2)\right),$$

donde ψ_i son los coeficientes de la representación:

$$X_t = c + w_t + \psi_1 w_{t-1} + \psi_2 w_{t-2} + \ldots$$

Entonces, un intervalo de predicción al 95 % para X_{T+k} será:

$$\hat{X}_T(k) \pm 1.96\sqrt{\sigma_w^2(1 + \psi_1^2 + \cdots + \psi_{k-1}^2)}.$$

2.8. Conclusiones

Se ha definido una clase de procesos a partir de los cuales podrán ser modelizadas gran cantidad de series de tiempo. En la Figura 2.1 se resume el proceso que es necesario llevar a cabo:

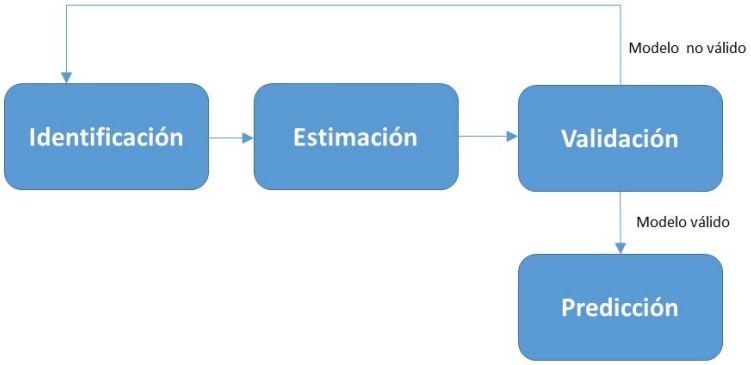

Figura 2.1: Proceso para ajustar un modelo Box-Jenkins.

Capítulo 3

Regresiones dinámicas

Los modelos de regresión dinámica sirven para modelizar la dependencia existente entre dos o más series temporales. Como se verá en este capítulo, la naturaleza de las series temporales hace que las técnicas clásicas empleadas en la regresión sean cuestionadas y haya que hacer alguna modificación.

En la Sección 3.1 se estudian medidas estadísticas para describir la dependencia lineal entre dos procesos estocásticos y el concepto de estacionariedad conjunta. En la Sección 3.2 se recuerda el modelo de regresión lineal para, a lo largo de la Sección 3.3 estudiarlo en el contexto de series temporales.

Es habitual que si se trata de aplicar las técnicas clásicas de regresión a series no estacionarias se lleguen a resultados engañosos, llamadas correlaciones espurias, que se detallan en la Sección 3.4. En la Sección 3.5 se describe una manera característica de trabajar con regresión de series temporales cuando estas no son estacionarias.

La teoría se ha desarrollado, por simplicidad y para una mejor compresión, para el modelo de regresión lineal simple, siendo sencilla la extensión al caso múltiple. Para elaborar este capítulo se han seguido las referencias bibliográficas Peña (2010) y Cryer y Chan (2008).

3.1. Conceptos

Definición 3.1. Dados dos procesos estacionarios $\{X_t\}$ e $\{Y_t\}$, se van a definir medidas estadísticas que dan cuenta de su dependencia lineal:

- La función covarianza cruzadas mide la relación lineal entre las variables X_s e Y_t, y depende del orden en el que se tomen las variables y los instantes temporales.

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

- Función de correlaciones cruzadas es una estandarización de la función anterior, y sirve para medir la magnitud y el sentido de la relación:

$$\rho_{X,Y}(s,t) = \frac{\gamma_{X,Y}(s,t)}{\sqrt{\gamma_X(0)}\sqrt{\gamma_Y(0)}},$$

  donde $\gamma_X(k)$ es la función de autocovarianza de un proceso estacionario $\{X_t\}$ (Definición 1.4).

Definición 3.2. Dos procesos $\{X_t\}$ y $\{Y_t\}$ se dirán conjuntamente estacionarios si:

1. Ambos son estacionarios.
2. Los coeficientes de covarianzas cruzadas solo dependen de la distancia entre las variables:

$$\gamma_{X,Y}(t, t+k) = \gamma_{X,Y}(k), \quad \forall k.$$

Para procesos conjuntamente estacionarios, la función de autocorrelaciones cruzadas se define para cada retardo, $\rho_{X,Y}(k)$. Esta función no solamente mide la fortaleza de la relación entre los procesos, sino que también indica la dirección. Esto se puede comprender mediante la Figura 3.1, donde se muestra la manera de interpretar los coeficientes de autocorrelación y de autocovarianza cruzados. Es importante darse cuenta de que se trata de funciones no simétricas respecto a cero, $\rho_{X,Y}(k) \neq \rho_{X,Y}(-k)$.

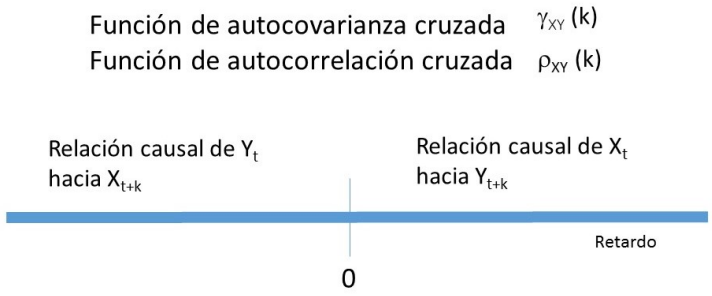

Figura 3.1: Interpretación de la función de correlaciones cruzadas.

En este capítulo se dirá que dos procesos $\{X_t\}$ e $\{Y_t\}$ son estacionarios para indicar que son mutuamente estacionarios y en este caso será posible estimar las funciones anteriores como:

- Función de covarianzas cruzadas muestral:

$$\hat{\gamma}_{X,Y}(k) = \frac{1}{T}\sum_{t=1}^{T-k}(x_t - \bar{x})(y_{t+k} - \bar{y}).$$

- Función de correlaciones cruzadas muestral:

$$\hat{\rho}_{X,Y}(k) = \frac{\hat{\gamma}_{X,Y}(k)}{\sqrt{\hat{\gamma}_X(0)}\sqrt{\hat{\gamma}_Y(0)}},$$

  con $\hat{\gamma}_X(k)$ la función de autocovarianza muestral del proceso estacionario $\{X_t\}$ (Definición 1.12).

Las funciones de correlaciones cruzadas muestrales a menudo se representan en un gráfico como el de la Figura 3.2. La interpretación de este gráfico se hace en base a lo detallado en la Figura 3.1.

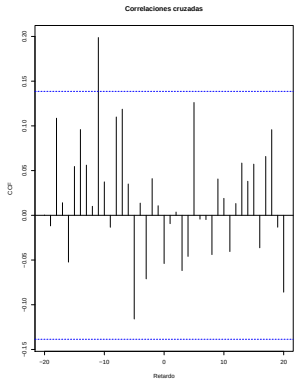

Figura 3.2: Correlaciones cruzadas muestrales.

3.2. Modelo de regresion lineal

Dadas las variables aleatorias X e Y, el modelo de regresión se define como:

$$Y = m(X) + \epsilon, \text{ donde } m(x) = \mathbb{E}(Y|X = x).$$

Las hipótesis básicas del modelo de regresión lineal simple son:

- Linealidad: la función de regresión es una recta: $m(X) = \beta_0 + \beta_1 X$.
- Homocedasticidad de los errores: $\text{var}(\epsilon|X = x) = \sigma^2$, $\forall x$.
- Normalidad de los errores: $\epsilon \in N(0, \sigma^2)$.

Además, en lo que al muestreo se refiere, se asume que los datos muestrales (X_i, Y_i) son independientes e idénticamente distribuidos (y por tanto, también lo son los correspondientes errores ϵ_i). Las hipótesis de homocedasticidad, normalidad e independencia simplifican las tareas de inferencia. En consecuencia, se puede reescribir el modelo como:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \text{ con } \epsilon_i \text{iid} N(0, \sigma^2), \text{ i =1,. . . ,n} \tag{3.1}$$

La estimación se lleva a cabo mediante mínimos cuadrados ordinarios (MCO, abreviadamente), obteniendo el siguiente estimador de la pendiente:

$$\hat{\beta}_1^{MCO} = \frac{S_{XY}}{S_X^2} \text{ con } \hat{\beta}_1^{MCO} \in N\left(\beta_1, \sqrt{\frac{\sigma^2}{nS_X^2}}\right), \tag{3.2}$$

con $S_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ y $S_X^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$.
Para más información acerca de los modelos de regresión lineal se puede consultar Faraway (2004, Cap.2).

3.3. El modelo de regresión lineal en el contexto temporal

En el caso de querer replicar el modelo (3.1) en el contexto de series temporales para los procesos $\{Y_t\}$ y $\{X_t\}$, tendría la siguiente forma:

$$Y_t = \beta_0 + \beta_1 X_t + \epsilon_t, \text{con } \{\epsilon_t\} \text{ ruido blanco gaussiano de varianza } \sigma_w^2. \tag{3.3}$$

Las estimaciones serían:

$$\hat{\beta}_1^{MCO} = \frac{\hat{\gamma}_{XY}(0)}{\hat{\gamma}_X(0)} \text{ con } \hat{\beta}_1^{MCO} \in N\left(\beta_1, \sqrt{\frac{\sigma_w^2}{T\hat{\gamma}_X(0)}}\right), \tag{3.4}$$

Si los procesos con los que se está trabajando $\{Y_t\}$ y $\{X_t\}$ son ruido blanco gaussiano, no se incumpliría ninguna de las hipótesis del modelo y podría emplearse la metodología clásica de modelos lineales de regresión. Sin embargo, las series con las que se trabaja no suelen ser estacionarias, ni mucho menos ruido blanco. Al tratar de aplicar este modelo (3.3) a variables que no son ruido blanco surgen dos inconvenientes con los que será necesario lidiar:

1. En el modelo (3.3) se supone que los errores son independientes, pero al trabajar con series temporales esta hipótesis es difícil que se verifique, siendo muy posible que los errores presenten autocorrelación.
2. El valor de Y_t puede depender no solo de manera contemporánea de la variable X_t, sino de la variable retardada k instantes de tiempo, X_{t-k}, y será la tarea decidir cuál es el retardo adecuado.

3.3.1. Errores autocorrelados

Se presenta ahora un modelo en el que los errores tienen una estructura dinámica:

$$Y_t = \beta_0 + \beta_1 X_t + n_t, \text{con } \{n_t\} \text{ un proceso estacionario.} \tag{3.5}$$

Si se estima el modelo (3.5) sin tener en cuenta la existente estructura de dependencia en los errores y, para contrastar la existencia de relación lineal se utiliza (erróneamente) la distribución normal mostrada en (3.4), entonces aparecerán falsas relaciones lineales entre los procesos, llamadas relaciones espurias (que se estudian con más detalle en la Sección 3.2).

En un contexto de dependencia en los errores como el del modelo (3.5), la varianza del estimador de mínimos cuadrados de la pendiente teniendo en cuenta la estructura de autocorrelación del error tiene la siguiente estructura (véase Peña (2010, pp. 533-535):

$$Var(\hat{\beta}_1) = \frac{\sigma_w^2}{T s_X^2} \left(1 + 2 \sum_{t=1}^{T-1} (T-t) \rho_n(t) \mathbb{E}(\hat{\rho}_X(t)) \right), \tag{3.6}$$

con σ_w^2 la varianza del error y $\rho_n(t)$ y $\hat{\rho}_X(t)$ los coeficientes de autocorrelación teóricos y muestrales del error y del proceso $\{X_t\}$, respectivamente.

Como consecuencia de la Ecuación (3.6), si se emplea el estimador de mínimos cuadrados ordinarios sin tener en cuenta la autocorrelación de los errores (3.4), este tendría una varianza menor que (3.6) y conduciría a rechazar falsamente la hipótesis nula de que la pendiente sea nula ($H_0 : \beta_1 = 0$), asumiendo una relación lineal inexistente.

La manera de proceder es encontrar una transformación del modelo (3.5), de manera que en vez de un error con autocorrelación aparezca un proceso de ruido blanco. Cuando el error es un proceso estacionario, esto se puede lograr mediante los procesos estudiados en el Capítulo 2. Si se empleara un modelo $ARMA(p, q)$ para modelizar el error, el modelo resultante sería:

$$\begin{aligned} Y_t &= \beta_0 + \beta_1 X_t + \epsilon_t \\ \phi(B)\epsilon_t &= \theta(B) w_t, \end{aligned}$$

donde $\{w_t\}$ es un proceso de ruido blanco con varianza σ_w^2.

Para estimar este modelo se puede emplear el método de máxima verosimilitud o mínimos cuadrados generalizados, que es una modificación del método de mínimos cuadrados ordinarios para la estimación de modelos de regresión para el caso de que los errores son heterocedásticos o presenten correlación. Para ver este procedimiento con más detalle se recomienda consultar Fox (2010).

Ajustado el modelo, es necesario verificar sus hipótesis, que son que el proceso $\{w_t\}$ sea un proceso de ruido blanco gaussiano. Como el error es inobservable, estas comprobaciones se harán con los residuos del modelo. Para ello se emplearán los contrastes y los métodos gráficos establecidos en la Sección 2.7.

3.3.2. Elección del retardo

En muchas ocasiones, la relación lineal entre los procesos no tiene por qué ser contemporánea, pudiendo estar la variable Y_t relacionada con la variable explicativa retardada r instantes temporales, X_{t-r}:

$$Y_t = \beta_0 + \beta X_{t-r} + \epsilon_t, \text{con } \epsilon_t \in N(0, \sigma^2). \tag{3.7}$$

En los modelos $AR(p)$ Y $MA(q)$ las funciones de autocorrelación parciales y simples servían para identificar los órdenes del modelos. Podría pensarse en emplear la función de autocorrelaciones cruzadas para elegir el retardo r. No obstante, como se verá en el siguiente ejemplo propuesto en Peña (2010, p.503) esto no siempre tiene sentido.

Ejemplo: Sean $\{X_t\}$ e $\{Y_t\}$ dos procesos estacionarios de media cero y con la siguiente relación:

$$Y_t = aX_t + bX_{t-1} + w_{t+k}. \tag{3.8}$$

La función de covarianzas cruzada sería:

$$\begin{aligned}\gamma_{XY}(k) &= \mathbb{E}(X_{t-k}Y_t) = \mathbb{E}(X_{t-k}(aX_t + bX_{t-1} + w_{t+k}) = a\mathbb{E}(X_{t-k}X_t) + b\mathbb{E}(X_{t-k}X_{t-1}) + \mathbb{E}(X_{t-k}w_{t+k}) \\ &= a\gamma_X(k) + b\gamma_X(k-1).\end{aligned}$$

La función de correlaciones cruzadas es una función de los coeficientes que relacionan los procesos y de la covarianza del proceso $\{X_t\}$. Si se supone que este sigue un proceso $AR(1)$, entonces $\gamma_x(k) \neq 0$. Por este motivo, si $b \neq 0$, entonces los coeficientes de correlación cruzados serán todos no nulos, indicando correlación entre X_{t-k} e Y_t para todo t. Sin embargo, esta es una conclusión errónea; recuérdese que esta relación solo existe para $k = 0$ o $k = 1$ (véase Ecuación (3.8)).

Con este ejemplo, queda claro que la función de correlaciones cruzadas es a veces inviable para identificar la dirección o el número de retardos de la relación.

Las propiedades muestrales de la función de correlación cruzada son complicadas. Sin embargo, si ambos procesos $\{X_t\}$ e $\{Y_t\}$ son ruido blanco, se tiene que:

$$\hat{\rho}_{X,Y}(k) \in N\left(0, \frac{1}{\sqrt{T}}\right).$$

Esta propiedad se verifica también en el caso de que uno de los procesos sea ruido blanco y el otro estacionario (veáse Cryer y Chan (2008, p.265)). Entonces, si se tienen dos series no estacionarias se puede aplicar un procedimiento para transformar una de ellas en ruido blanco y la otra en un proceso estacionario. Este procedimiento se llama preblanqueado. Observadas las series $\{y_1, \ldots, y_T\}$ y $\{x_1, \ldots, x_T\}$ se recomienda seguir los siguientes pasos:

Paso 1: Transformar el proceso $\{X_t\}$ en un proceso de ruido blanco $\{\widetilde{X}_t\}$: $\widetilde{X}_t = \pi(B)X_t$.

Paso 2: Aplicar la misma transformación al proceso $\{Y_t\}$, $\widetilde{Y}_t = \pi(B)Y_t$ suponiendo que $\{\widetilde{Y}_t\}$ es estacionario.

Paso 3: Proponer una relación lineal entre $\{X_t\}$ y $\{X_t\}$. Si para algún retardo se verifica:

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

entonces se propondrá el modelo $Y_t = \beta_0 + \beta_1 X_{t-k} + \epsilon_t$ [1].

Para un análisis preliminar rápido de este tipo es recomendable diferenciar regular y estacionalmente las variables si presentan tendencia o estacionalidad y ajustarles un modelo sencillo, del tipo $AR(p)$. Como desventaja de este procedimiento, cabe destacar que es complicado extenderlo a un modelo de más de dos variables.

3.3.3. Manera de proceder

A continuación se muestra el procedimiento que es recomendable seguir:

Paso 1: Elegir el orden de los retardos, bien por un proceso de preblanqueado o bien por una rutina más automática.

[1] Como $\pi(B)$ es un operador lineal, cualquier relación lineal entre las series transformadas se mantendrá entre las series originales.

Paso 2: Ajustar un modelo de regresión lineal entre las variables actuando como si los errores fueran incorrelados y observar los residuos. En el caso de que sean estacionarios, ir al Paso 3. Si los residuos no son estacionarios se recomienda diferenciar las series y volver a hacer el ajuste o volver al Paso 1.

Paso 3: Examinar los residuos para identificar un modelo $ARMA(p,q) \times (P,Q)$. Para ello se recomienda seguir la metodología descrita en el Capítulo 2.

Paso 4: Ajustar el modelo de manera conjunta, bien por mínimos cuadrados generalizados o por máxima verosimilitud.

Paso 5: Validación del modelo. En el caso de que no sea un modelo válido, se recomienda volver al Paso 3 y probar otro modelo para los residuos. Si no se consigue un modelo válido, volver al Paso 1.

3.4. Relaciones espurias

En Neyman (1952) se analiza la relación entre la tasa de nacimientos y la población de cigüeñas en varias regiones, encontrando un elevado coeficiente de correlación entre estas variables variables. Dos variables se dicen que tienen una relación espuria cuando aparentemente tienen una fuerte relación empírica pero no existe entre ellas ninguna conexión lógica. Que exista una fuerte correlación entre dos fenómenos A y B puede ser debido a cuatro factores:

- Que A cause B.
- Que B cause A.
- Que haya un tercer fenómeno, C, que provocara tanto A como B.
- Puro azar.

En base a lo anterior, se puede afirmar que la correlación no siempre implica causalidad. Existen situaciones en las que la correlación implica pura casualidad. En la página web creada por Tyler Vigen, `http://www.tylervigen.com/spurious-correlations` se encuentran una gran cantidad de fenómenos entre los que existe una relación espuria. Por ejemplo, el coeficiente de correlación entre el número de divorcios y el consumo de margarina en Maine es del 99.25 %.

Al estimar regresiones entre series no estacionarias es muy fácil que la relación sea espuria, basta con que ambas series tengan tendencias crecientes o decrecientes para que surja una aparente relación entre ellas. En la mayoría de las situaciones, las relaciones son de casualidad y no de causalidad. Al suprimir la tendencia, por ejemplo, diferenciando los datos, la relación espuria desaparece.

3.5. Cointegración

Como ya se ha mencionado, una solución para evitar las correlaciones espurias podría ser emplear el método de preblanqueado o bien diferenciar las series para así trabajar con errores estacionarios. Sin embargo, es posible que aunque dos series $\{X_t\}$ y $\{Y_t\}$ sean no estacionarias, una explique completamente el comportamiento no estacionariario de la otra, es decir:

$$Y_t = \beta_0 + \beta_1 X_t + \epsilon_t,$$

donde $\{\epsilon_t\}$ es estacionario. En este caso $\{X_t\}$ explica el comportamiento no estacionario de $\{Y_t\}$. Este fenómeno se llama cointegración y, de no haberse tenido en cuenta, se hubieran diferenciado las series y llegado a un modelo de regresión lineal en diferencias más complicado de interpretar.

Definición 3.3. Un proceso estocástico $\{X_t\}$ se dice que es integrado de orden d, y se representa por $I(d)$ si es necesario aplicarle una diferencia regular de orden d para transformarlo en un proceso estacionario.

Definición 3.4. Se dirá que dos procesos $\{X_t\}$ e $\{Y_t\}$ que son $I(d)$ están cointegrados si existe una combinación lineal de ellos que es de orden de integración menor que d. Es decir, si se puede construir un proceso $\{\epsilon_t\}$ que es $I(d_0)$, con $d_0 < d$ como sigue:

$$\epsilon_t = \alpha_1 Y_t + \alpha_2 Y_t.$$

Un caso de cointegración importante en la práctica es cuando las dos series son $I(1)$, pero existe una combinación lineal que se encontrará por medio de una regresión por mínimos cuadrados ordinarios que es estacionaria.

La manera de proceder se resume en los siguientes pasos.

Paso 1: Estimar por mínimos cuadrados una regresión entre las variables y construir la variable $\hat{y}_t$ como:

$$\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 x_t.$$

Paso 2: Obtener los residuos de la siguiente ecuación:

$$\hat{\epsilon}_t = y_t - \hat{\beta}_0 - \hat{\beta}_1 x_t.$$

Si son estacionarios, las variables están cointegradas e ir al paso siguiente. Si no son estacionarios, se recomienda desistir de esta metodología y seguir los pasos detallados en la Sección 3.3.2.

Paso 3: Existen varias opciones de modelos para esta situación (véase Peña (2010, pp. 545-547)), a continuación se muestra el llamado modelo de corrección de error:

$$\nabla y_t = \alpha(y_{t-1} - \beta_1 x_{t-1}) + \sum_{i=0}^{q} \omega_i \nabla x_{t-i} + \sum_{j=1}^{q} \phi_j y_{t-i} + w_t,$$

donde p y q se eligen para que los residuos sean ruido blanco.

Capítulo 4

Corrección de series temporales

Cuando se estudian series temporales de carácter socioeconómico es habitual trabajar con las series corregidas de calendario y estacionalidad; este es el caso del Producto Interior Bruto, objeto de estudio de este trabajo. Con la finalidad de entender en qué consiste esta transformación y el motivo por el que se lleva a cabo, se ha elaborado este capítulo.

En la Sección 4.1 se motiva el proceso de desestacionalizar una serie temporal y se introducen los conceptos necesarios. El programa estadístico que se emplea para este cometido es TRAMO SEATS, cuya metodología se describe en la Sección 4.2. Para comprender este procedimiento y su importancia, se propone el estudio de un caso real en la Sección 4.3. Por último, en la Sección 4.4 se analiza el caso particular del Producto Interior Bruto, así como el de los indicadores empleados a lo largo de este trabajo.

Para la elaboración de este capítulo se han seguido numerosos trabajos, entre los que destacan IGE (2016), INE (2016), INE (2002a), INE (2002b) y Caporello y Maravall (2004).

4.1. Motivación y conceptos previos

A la hora de estudiar el comportamiento de una variable temporal es habitual no tener en cuenta únicamente su evolución en niveles y, con el objetivo de poder detectar cambios a lo largo del tiempo, se emplean diferentes medidas para comparar distintos periodos, como las variaciones intertrimestrales e interanuales. Recurrir a variaciones permite cuantificar el crecimiento o decrecimiento que ha experimentado la variable respecto a un periodo de referencia. No obstante, tiene como desventaja el retraso que se produce a la hora de identificar los puntos de inflexión de la serie.

Cuando se compara una variable en diferentes instantes temporales es necesario tener en cuenta algunas consideraciones para no caer en resultados o conclusiones engañosas. El heladero que festeje haber tenido un incremento de ventas en julio respecto a enero es candidato a arruinarse (de Pablo (2013)). Para poder comparar las ventas de enero con las ventas de julio se debería extraer el efecto que el verano tiene en las ventas de helados. De igual modo, a la hora de analizar las variaciones interanuales de variables mensuales que cuantifiquen la producción o demanda, se debería tener en cuenta cuándo ha coincidido la Semana Santa o el número de días laborables del mes. Por un lado, si el primer año la Semana Santa coincidió en mayo y el siguiente año en abril, considerar la variación interanual del mes de abril no será correcto sin extraer del segundo año el efecto de la Semana Santa. Por otro lado, no se pueden esperar los mismos resultados en cuanto a producción o ventas en un mes que ha tenido tres días laborables más que otro.

Una serie temporal puede llegar a descomponerse en cinco componentes **ortogonales**: tendencia, ciclo, estacionalidad, irregularidad y efectos de calendario. Las cuatro primeras componentes son estocásticas y no observables, mientras que la última es determinista. Las descomposiciones que se van

a considerar en este trabajo son de de tipo aditivo o multiplicativo[1]:

$$\textbf{Modelo aditivo}: y_t = T_t + C_t + S_t + CALt + I_t.$$
$$\textbf{Modelo multiplicativo}: y_t = T_t \cdot C_t \cdot S_t \cdot CALt \cdot I_t.$$

Las componentes se pueden definir como sigue:

Tendencia (T_t): Movimientos de duración superior a 8 años que reflejan la evolución de la serie a largo plazo. Es el resultado de influencias como el crecimiento de la población, la inflación de los precios y los cambios económicos, tecnológicos e institucionales, entre otros.

Ciclo (C_t): También llamado ciclo económico, son oscilaciones cuya duración se sitúa entre 2 y 8 años en torno a la tendencia y que está generado por factores diferentes. Está compuesto principalmente por periodos alternos de expansión y contracción y la duración depende del mercado o industria que se considere. Como ejemplo, se puede pensar en los periodos en los que un determinado producto está de moda o en auge.

Estacionalidad (S_t): Se trata de un movimiento periódico o casi periódico de duración inferior a un año. Puede identificarse por la presencia de picos y valles regularmente espaciados que tienen una dirección constante y aproximadamente la misma o proporcional magnitud cada año.

Efectos de calendario ($CALt$): Son aquellos movimientos que están relacionados con la composición del calendario y el efecto que este tiene a lo largo de la evolución de la serie. Las fiestas móviles, la proporción de los distintos días de la semana dentro de cada mes y la ocurrencia de año bisiesto se engloban en esta componente. Las fiestas fijas como la Navidad se consideran que están en la componente estacional.

Irregularidad (I_t): Son movimientos erráticos y generalmente impredecibles (huelgas, guerras, elecciones, etcétera) que distorsionan la relación lineal entre la serie observada y sus componentes estructurales. La componente irregular recoge la incapacidad del resto de componentes para explicar totalmente el comportamiento de la serie.

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), dando lugar a la siguiente descomposición en el caso aditivo:

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

La estacionalidad y los efectos de calendario a menudo pueden enmascarar movimientos subyacentes de carácter no estacional de la serie. Se conoce como proceso de desestacionalización a la extracción de estas dos componentes de la serie temporal.

La serie desestacionalizada es aquella a la que se la han sustraído la componente estacional y los efectos de calendario. Esta nueva serie proporciona una estimación de lo nuevo (cambio de tendencia, ciclo y componente irregular). Además, es más fiable para medir la señal de tendencia y realizar comparaciones entre meses consecutivos y no consecutivos.

El manual por el que se guían las principales instituciones y países para este tratamiento de series temporales es el publicado por Eurostat, donde entre las principales ventajas de desestacionalizar se citan (veáse Eurostat (2009, p.7):

- La evolución de las series ajustadas de estacionalidad es más comprensible para el análisis económico.
- Facilita la comparación de movimientos de corto y largo plazo entre sectores y países.
- Suministra a los usuarios el input necesario para el análisis cíclico, detención de puntos de giro, descomposición ciclo-tendencia, etc.

[1] El modelo multiplicativo se convierte en el aditivo sin más que aplicar logaritmos a la serie original.

Como contrapunto, es también interesante mencionar alguno de los riesgos de publicar o trabajar con las series desestacionalizadas:

- Las series ajustadas dependen del método utilizado y de las hipótesis del modelo elegido, así como del sofware empleado en el proceso.
- Un ajuste estacional inapropiado o de baja calidad puede generar resultados erróneos y señales falsas.

Para desestacionalizar una serie temporal será necesario estimar la componente estacional y los efectos de calendario. Para ello, existen dos enfoques:

Enfoque no paramétrico: Permite estimar las componentes no observadas sin recurrir a la especificación de un modelo estadístico para la serie. La metodología de ajuste estacional más utilizada es la del programa X12 ARIMA y las componentes se estiman mediante la aplicación sucesiva de filtros lineales que son independientes de la serie con la que se está trabajando.

Enfoque paramétrico: Parte de la especificación explícita de un modelo estadístico para la serie de tiempo observada o bien para las componentes. La metodología más empleada es TRAMO SEATS, donde los filtros aplicados se adaptan a la estructura particular de la serie con la que se trabaja.

4.2. Corrección de estacionalidad y calendario mediante un enfoque paramétrico

El Instituto Galego de Estatística (IGE) sigue un enfoque paramétrico para corregir las principales series de coyuntura de Galicia mediante el programa TRAMO SEATS, que es la metodología que se detalla en esta sección. Puesto que la corrección de series temporales no es el objeto de estudio de este trabajo, no se entrará en profundidad.

La metodología seguida por TRAMO-SEATS para desestacionalizar una serie temporal sigue un procedimiento de dos etapas:

1. Para estimar los efectos de calendario se ajusta un modelo de regresión con variables ficticias cuyos residuos siguen un modelo autorregresivo, integrado y de medias móviles de tipo multiplicativo $(ARIMA(p,d,q) \times (P,D,Q)_s)$.
2. A la serie corregida de los efectos de calendario se le estiman las componentes estocásticas.

Los efectos de calendario que se van a corregir son:

- **Ciclo Semanal** (CS_t).
- **Semana Santa** (E_t).
- **Año bisiesto** (LY_t).

La cuantificación de los tres efectos anteriores se realiza mediante la identificación, estimación y diagnóstico de un modelo de regresión con variables ficticias y cuyos residuos siguen un modelo $ARIMA(p,d,q) \times (P,D,Q)_s$. Se detalla a continuación la manera en la que se construyen las variables que modelizan cada uno de los efectos anteriormente señalados.

La distinta composición de los meses puede afectar a variables que cuantifiquen la producción o demanda de un bien o servicio; de manera que un mes podría tener una mayor producción que otro debido únicamente a que hubo más días laborables. Particularidades como que el número de días laborables de cada mes sea distinto, que la actividad puede no ser la misma todos los días de la semana y que no todos los meses tienen la misma longitud, se engloban en el efecto de Ciclo Semanal. Existen muchas maneras de parametrizar este efecto; una de ellas es diferenciando dentro de cada

mes entre el número de días laborables (lunes-viernes) y el número de días no laborables (sábados y domingos). Para ello, se construye la siguiente variable regresora:

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

siendo $X_{i,t}, i \in \{1 \cdots 7\}$ el número de lunes, martes,... del mes t y $\frac{5}{2}$ es la constante que homogeneiza los dos elementos de D_t (veáse INE (2016, p. 49)). El efecto de Ciclo Semanal es:

$$CS_t = \beta D_t.$$

La Semana Santa es una fiesta móvil resultado de situar el Domingo de Pascua en el primer domingo posterior a la primera luna llena que sigue al equinocio de primavera, variando su localización temporal entre el 22 de marzo y el 25 de abril. Esta festividad tiene un gran impacto cuando el Domingo de Pascual coincide en meses diferentes dos años consecutivos.

Se construye la siguiente variable regresora, bajo el supuesto de que el impacto de la Semana Santa se percibe de manera idéntica, en los τ días anteriores al Domingo Santo [2] (veáse Gómez(1998, p.29)):

$$P(\tau)_t = \begin{cases} p_t(\tau) - \frac{1}{2} & \text{si t=marzo ó abril} \\ 0 & \text{en otro caso,} \end{cases}$$

donde $p_t(\tau)$ es la proporción de los τ días anteriores al domingo de Pascua que corresponde al mes/trimestre t. El efecto de Semana Santa es:

$$E_t = \gamma P(\tau)_t.$$

El hecho de que un año sea bisiesto y el mes de febrero cuente con un día más puede tener impacto en variables que, por ejemplo, miden la producción. La variable regresora que se emplea para parametrizar el efecto de Año Bisiesto se construye de la siguiente manera, (veáse INE (2016, p. 50)):

$$B_t = \begin{cases} 0.75 \text{ si t es febrero y el año es bisiesto.} \\ -0.25 \text{ si t es febrero y el año no es bisiesto.} \\ 0 \text{ si t no es febrero.} \end{cases}$$

Y el efecto del Año Bisiesto es:

$$LY_t = \alpha B_t.$$

Los efectos de calendario se modelizan como:

$$CAL_t = CS_t + E_t + LY_t = \beta D_t + \gamma P(\tau)_t + \alpha B_t,$$

y se cuantifica mediante el siguiente modelo:

$$y_t = \beta D_t + \gamma P(\tau)_t + \alpha B_t + \frac{\theta_q(B)\theta_Q(B^s)(1-B)^d(1-B)^D}{\phi_p(B)\phi_P(B^s)(1-B)^d(1-B)^D} a_t. \tag{4.1}$$

La presencia de efectos de calendario se determina mediante contrastes de significación estadística con hipótesis nulas $\gamma = 0$ (ausencia de efecto de Ciclo Semanal), $\beta = 0$ (ausencia de efecto de Semana Santa) y $\alpha = 0$ (ausencia de efecto de Año Bisiesto) en el modelo (4.1). La serie corregida de los efectos de calendario, N_t, se obtiene sustrayendo de la serie original los efectos significativos. Si los tres efectos lo son:

$$N_t = y_t - \hat{\gamma} P(\tau)_t - \hat{\beta} D_t - \hat{\alpha} B_t.$$

[2] Normalmente la significación más alta del coeficiente se da con $\tau = 6$.

La serie corregida de los efectos de calendario N_t se descompone en sus componentes estocásticas de ciclo-tendencia, estacionalidad e irregularidad. La metodología seguida considera que cada componente sigue un modelo autoregresivo integrado de medias móviles de tipo multiplicativo, teniendo que ser los modelos compatibles, en su conjunto, con el que caracteriza a la serie corregida de los efectos de calendario N_t.

Si se supone que N_t se desagrega en k componentes ortogonales entre sí:

$$N_t = \sum_{i=1}^{k} N_{i,t}.$$

Lo expuesto antes se traduce en:

$$N_{i,t} = \frac{\theta_i(B)}{\phi_i(B)} a_{i,t} = \psi_i a_{i,t} \text{ con } a_{i,t} \sim \text{iid} N(0, V_i) \qquad N_t = \frac{\theta(B)}{\phi(B)} a_t = \psi a_t \text{ con } a_t \sim \text{iid} N(0, V_a)$$

Como consecuencia, se llega a las siguientes condiciones que relacionan los operadores de la serie corregida con los operadores de las componentes inobservables $N_{i,t}$:

$$\sum_{i=1}^{k} \frac{\theta_i(B)}{\phi_i(B)} a_{i,t} = \frac{\theta(B)}{\phi(B)} a_t \rightarrow \begin{cases} \phi(B) = \prod_{i=1}^{k} \phi_i(B) \\ \theta(B) = \sum_{i=1}^{k} \phi_{(i)}(B)\theta_i(B)a_{i,t}, \text{ siendo } \phi_{(i)}(B) = \prod_{j=1, j\neq i}^{k} \phi_j(B) \end{cases}$$

Aunque los operadores de la forma reducida N_t fueron estimados, no se pueden calcular el resto de operadores a partir de ellos, ya que existe un problema de identificación. Existen infinitas formas de $\phi_i(B)$ compatibles con el modelo $\phi(B)$. Para afrontar el problema anterior se recurre al Teorema (4.1).

Teorema 4.1. *Teorema de Descomposición Canónico.*

$$N_{i,t} = N_{i,t}^s + \zeta_{i,t}$$
$$\zeta_{i,t} \approx iid N(0, \sigma_i)$$

implica $N_{i,t} = N_{i,t}^s$ *(sólo existe señal) ó* $N_{i,t} = \zeta_{i,t}$ *(sólo existe ruído).*

El Teorema 4.1 establece que la descomposición de cada componente como una señal más ruido blanco es imposible. Esto implica que cada componente o bien es señal pura (sin información redundante) o bien es ruido blanco. Se emplea este teorema imponiendo que todo el ruido blanco de las componentes se le asigna a la componente irregular. De este modo, la varianza de la componente irregular es máxima, siendo el resto de componentes lo más estables posible.

Una vez aplicado el teorema, es posible determinar los valores de $\zeta_i(B)$ a partir de $\zeta(B)$ mediante, por ejemplo, el método de los momentos.

Una vez definidos los modelos teóricos para las componentes, se estiman las componentes a partir de los valores observados N_t. Para ello se emplean los filtros de la familia de Wiener-Kolmogorov, diseñados para minimizar el error cuadrático entre el estimador y el componente teórico. Estos filtros se obtienen como solución del siguiente problema:

$$\min_{\hat{N}_{i,t}} \mathbb{E}(N_{i,t} - \hat{N}_{i,t})^2$$
$$\text{s.a. } N_{i,t} = \Psi(B) a_{i,t}$$

Y la solución conduce a:

$$\hat{N}_{i,t} = \frac{V_i}{V_a} \frac{\psi_i(B)\psi_i(F)}{\psi(B)\psi(F)} N_t, \text{ con } F = B^{-1}$$

4.3. El programa TRAMO-SEATS y un ejemplo

El programa TRAMO-SEATS está compuesto por los programas TRAMO y SEATS, desarrollados por Víctor Gómez y Agustín Maravall. Existe una versión para Windows (TSW) desarrollada por Gianluza Caporello y Agustín Maravall, cuya interfaz gráfica se muestra en la Figura 4.1.

El programa TRAMO ("Time Series Regression with Arima Nose, Missing Observations and Outliers") permite la identificación, estimación y diagnosis de modelos de regresión cuyos errores siguen un modelo $ARIMA(p,d,q) \times (P,D,Q)_s$. Permite tener los ajustes previos para poder realizar un ajuste estacional.

El programa SEATS ("Signal Extraction in ARIMA Time Series") descompone la serie temporal corregida de los efectos deterministas, en sus componentes estocásticos no observables.

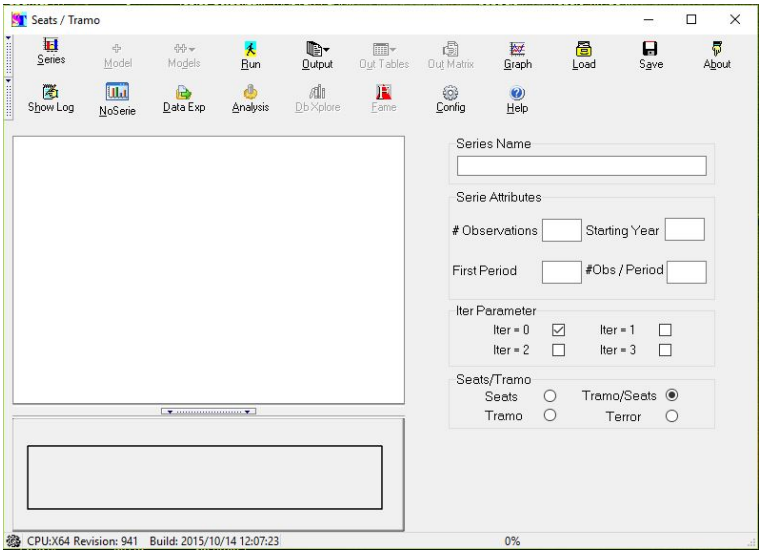

Figura 4.1: Interfaz del programa TSW.

En el Apéndice A se detallan las principales funciones y entradas del programa de manera que, con la ayuda de este capítulo, el lector pueda entender y ser capaz de hacer un ajuste sencillo.

Para comprender en qué consiste y la importancia de desestacionalizar una serie temporal se propone a continuación un ejemplo con la variable del consumo de gasolina en Galicia, medida en miles de toneladas, X_t. El ajuste se hace con un modelo de tipo multiplicativo, y la serie en logaritmos se corrige de los tres efectos de calendario estudiados. El IGE publica esta serie corregida de calendario y estacionalidad y los parámetros de entrada de TRAMO-SEATS pueden consultarse en IGE (2017, p.11).

La Figura 4.2 se representa la evolución del consumo de gasolina desde enero de 1993 hasta julio de 2017, pudiendo observarse una tendencia negativa. En los gráficos de abajo a la izquierda y abajo centro se muestra el consumo de cada mes a lo largo del periodo de estudio. Se observa que julio y agosto son los meses en los que hay un mayor consumo; mientras que el menor consumo de gasolina se produce en los meses de febrero y noviembre. Se puede apreciar, tanto en el gráfico de niveles como en el que muestra el consumo durante los años 2007, 2008 y 2009 (abajo a la derecha), que hay un patrón de picos y valles que se repite año tras año, entre los que se encuentran los repuntes de los meses citados. Este patrón repetitivo se corresponde con la componente estacional.

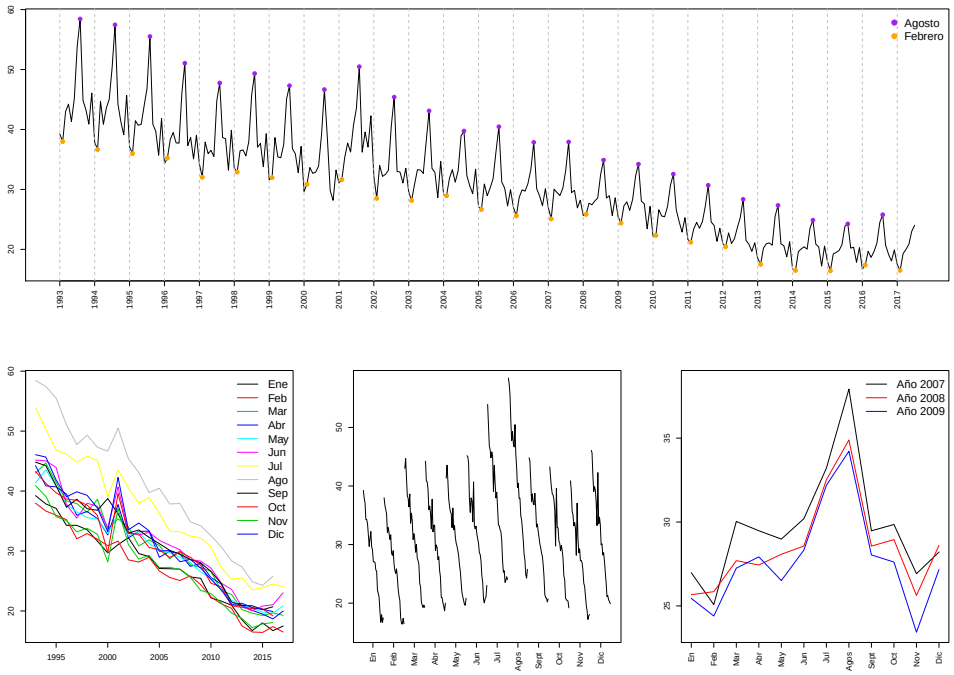

Figura 4.2: Evolución del consumo de gasolina en niveles (arriba). Series del consumo por meses (abajo izquierda y abajo centro). Consumo de gasolina para los años 2007, 2008 y 2009 (abajo derecha).

Con el objetivo de comprender los efectos de calendario se va a hacer un estudio detallado para los años 2008 y 2009. Las características de estos años pueden consultarse en el Cuadro 4.1, donde se aprecia que el 2008 fue un año bisiesto en el que la Semana Santa coincidió en marzo, mientras que el 2009 no fue bisiesto y la Semana Santa ocurrió en abril.

Año	Ene	Feb	Mar	Abr	May	Jun	Jul	Ago	Sep	Oct	Nov	Dic	Bis.	D. Santo
2008	23|8	21|8	21|10	22|8	22|9	21|9	23|9	21|10	22|8	23|8	20|10	23|8	Sí	23-03
2009	22|9	20|8	22|9	22|8	21|10	22|8	23|8	21|10	22|8	22|9	21|9	23|8	No	12-04

Cuadro 4.1: Número de días laborables y festivos (laborable|festivo) de cada mes, indicador de si el año fue bisiesto (Bis) y fecha del Domingo Santo (D. Santo) para los años 2008 y 2009.

$\hat{\beta}$ (Ciclo Semanal)	$\hat{\alpha}$ (Año Bisiesto)	$\hat{\gamma}$ (Semana Santa con $\tau = 6$)
0.004	0.051	0.036

Cuadro 4.2: Coeficientes obtenidos del modelo para corregir el efecto calendario de la serie de consumo de gasolina.

Las estimaciones de los coeficientes de los efectos de calendario se muestran en la Cuadro 4.2. Como el modelo empleado es el propuesto por el IGE, no se tendrá en cuenta la significación de los

parámetros en este momento, se asume que eran significativos en el momento que se decidió corregir la serie de estos tres efectos.

En el Cuadro 4.3 se muestran las variables regresoras construidas, los efectos estimados y el resultado final para una selección de meses del año 2008 a modo de ejemplo.

$$D_{2008-\text{M3}} = 21 - \frac{5}{2}10 = -4, \quad P(6)_{2008\text{-M3}} = (\frac{6}{6} - \frac{1}{2}) = 0.5, \quad B_{\text{M3-2008}} = 0$$

$$N_{2008\text{-M3}} = \exp\{\ln(3.321) - 0.004 \cdot (-4) - 0.036 \cdot \frac{1}{2} - 0.051 \cdot 0\} = 27.67$$

t	X_t	$ln(X_t)$	D_t	B_t	$P(6)_t$	βD_t	αB_t	$\gamma P(6)_t$	TD_c	LY_c	$EAST_c$	N_t
2008-M1	25, 67	3, 245	3	0	0	0, 0128	0	0	25, 344	25, 67	25, 67	25, 344
2008-M2	25.85	3.252	1	0.75	0	0.004	0.038	0	25.742	24.878	25.852	24.773
2008-M3	27.70	3.321	−4	0	0.5	−0.017	0	0.018	28.174	27.698	27.206	27.673
2008-M4	27.44	3.312	2	0	−0.5	0.009	0	−0.018	27.207	27.44	27.936	27.699
2008-M5	28.08	3.335	−0.5	0	0	−0.002	0	0	28.14	28.08	28.08	28.14

Cuadro 4.3: Desglose del modelo de efectos de calendario para una selección de meses del año 2008.

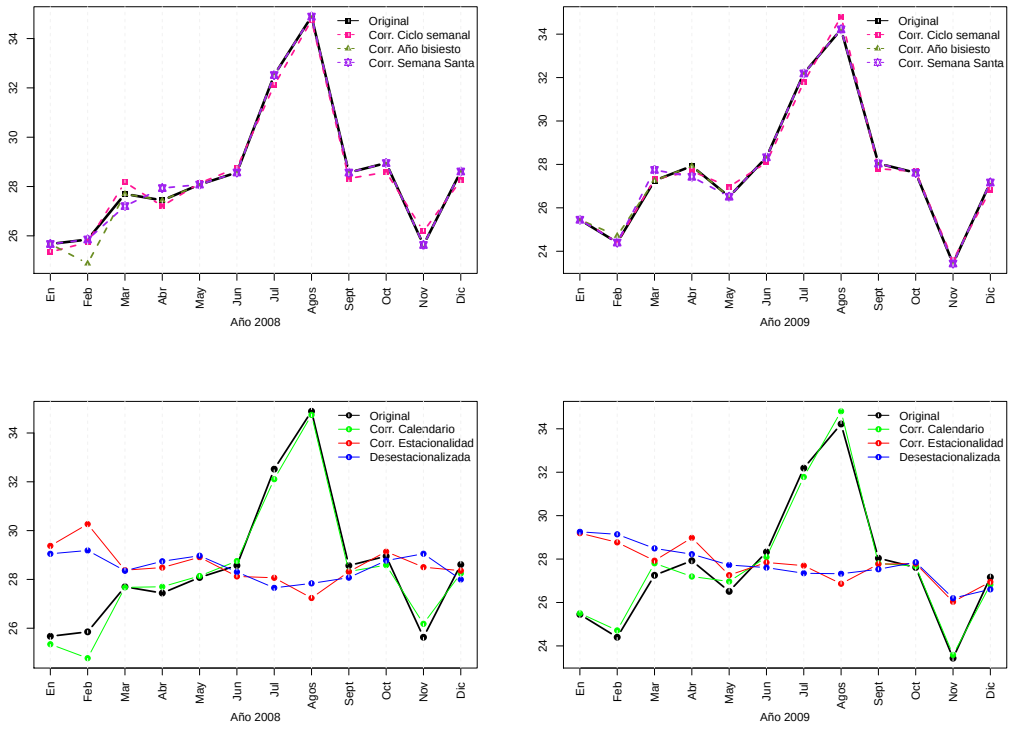

Figura 4.3: Distintas correcciones de la serie de consumo de gasolina para los años 2008 y 2009.

En la Figura 4.3 se muestran los datos de los años 2008 y 2009 corregidos de cada uno de los efectos de calendario, del efecto total de calendario, de estacionalidad y desestacionalizadas. En los gráficos

de arriba puede verse la serie corregida del efecto de año bisiesto en color verde oscuro; de manera que el nivel de la serie corregida de este efecto baja en febrero de 2008, mientras que para el año 2009 esto no ocurre. La Semana Santa coincidió en marzo en el 2008 y en abril en el 2009, y se aprecia como el nivel de la serie corregida de este efecto, línea morada, baja en marzo y sube en abril para el año 2008 y viceversa para el año 2009. La serie corregida del efecto de Ciclo Semanal es la línea roja, y vemos como es distinta cada mes a la serie original. En los gráficos de abajo de la Figura 4.3 se representa la series corregida de estacionalidad para los años 2008 y 2099 en líneas rojas; ya no se aprecian los altos repuntes de los meses de julio y agosto, ni las caídas de febrero o noviembre, que era parte del patrón repetitivo que se observaba año tras año en la Figura 4.2. La línea azul es la serie desestacionalizada, fruto de sustraer de la serie original los efectos de calendario y la componente estacional ajustados.

Se ha comentado la importancia de corregir las serie a la hora de comparar periodos temporales y en el Cuadro 4.4 esto se aprecia claramente. Cuando se considera la serie sin corregir, la variación interanual de marzo de 2009 es de $-1.6\,\%$, lo que se traduce en que en marzo de 2009 hubo un menor consumo de gasolina que en marzo del 2008. Al considerar estos datos en bruto, no se está teniendo en cuenta que la semana santa sólo coincidió en marzo en el año 2008. Cuando se estudian los datos después de extraerles el efecto de la Semana Santa, esta cifra se convierte en un $0.48\,\%$, lo que conduce a una conclusión opuesta a la primera: hubo un incremento del consumo de gasolina en marzo de 2009 respecto a marzo del 2008. Una conclusión similar se obtiene al considerar los datos desestacionalizados.

Fecha	Original	Corr. Calendario	Corr. Estacionalidad	Desestacionalizada
$\frac{y_{2009\text{-}M1}}{y_{2008\text{-}M1}} - 1$	-0.855	0.633	-0.777	0.712
$\frac{y_{2009\text{-}M2}}{y_{2008\text{-}M2}} - 1$	-5.615	-0.236	-5.545	-0.162
$\frac{y_{2009\text{-}M3}}{y_{2008\text{-}M3}} - 1$	-1.604	0.478	-1.601	0.480
$\frac{y_{2009\text{-}M4}}{y_{2008\text{-}M4}} - 1$	1.764	-1.818	1.774	-1.809
$\frac{y_{2009\text{-}M5}}{y_{2008\text{-}M5}} - 1$	-5.582	-4.165	-5.710	-4.295

Cuadro 4.4: Estudio en variaciones de la serie corregida de los efectos de calendario, la componente estacional y ambos.

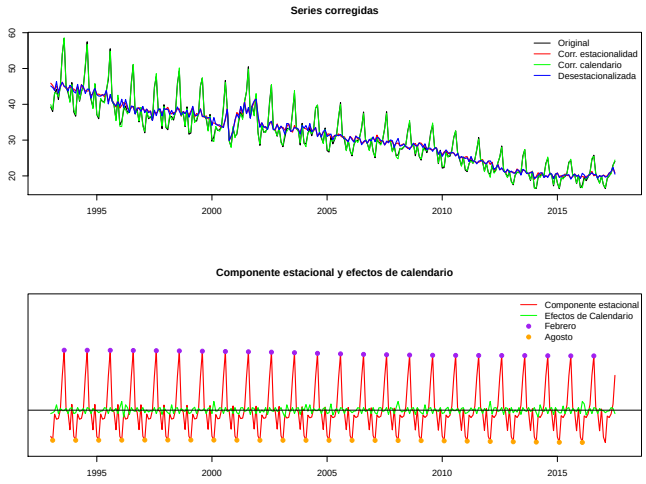

Figura 4.4: Serie del consumo de gasolina corregida de estacionalidad, calendario y desestacionalizada (arriba). Componentes de los efectos de calendario y de estacionalidad (abajo).

En la Figura 4.4 se representa la serie desestacionalizada. De nuevo, se aprecia como el patrón de picos y valles de cada año ha desaparecido. En la gráfica de abajo se muestran las componentes estacionales y los efectos de calendario ajustados. Al tratarse de un modelo multiplicativo, para corregir de estacionalidad se dividirá la serie en niveles entre la componente estacional. Los picos se corresponden con los meses de agosto y tendrán como consecuencia una reducción del nivel en estos meses de la serie corregida de estacionalidad, mientras que los picos que se corresponden con los meses de febrero generarán un aumento del nivel en la serie corregida de estacionalidad.

4.4. Corrección del PIB y de los principales indicadores

El IGE proporciona corregidas de calendario y estacionalidad las llamadas series de coyuntura de Galicia, que son los principales indicadores macroeconómicos. Los parámetros de entrada para ajustar los modelos en TSW pueden consultarse en IGE (2016, p.10).

Cada vez que se dispone de una nueva observación los modelos se reestiman, cambiando el valor de los coeficientes y, consecuencia de ello, se modifican también todos los valores de la serie corregida. Es por este motivo que a las series desestacionalizadas se les denomina "series vivas", ya que varían a medida que la serie va evolucionando en el tiempo.

Las series con las que habitualmente se trabaja son series estáticas; conforme se avanza en el tiempo la serie cuenta con más observaciones pero las observaciones pasadas no varían. Si la serie observada en tiempo T es $\{x_1, \ldots, x_T\}$, transcurridos dos instantes temporales la serie seguirá siendo la misma pero contará con dos observaciones más, $\{x_1, \ldots, x_T, x_{T+1}, x_{T+2}\}$. Por este motivo, la serie histórica empleada para elegir un modelo es siempre la misma. Escogido un modelo, se cumple que los datos con los que fue estimado y seleccionado no cambian. Esto no ocurre con las llamadas series vivas; si en tiempo T la serie observada es $\{x_1, \ldots, x_T\}$, dos instantes temporales después la serie será $\{\hat{x}_1, \ldots, \hat{x}_T, \hat{x}_{T+1}, \hat{x}_{T+2}\}$, dificultando la elección de un modelo.

La serie del PIB con la que se trabaja está corregida de calendario y estacionalidad y además está sometida a un tratamiento para garantizar su consistencia temporal y transversal. La consistencia temporal sirve para asegurar que la serie corregida sea consistente con los datos anuales; esto quiere decir que aunque la serie cambia cada vez que se corrige, esta corrección se ajusta para que el dato anual se mantenga. Es decir, si la media del año 2015 fue un 2.4, cada vez que se corrija la serie los datos del año 2015 cambiarán pero esta variación se mantendrá. Cuando se publica la serie del PIB también se publica un desglose de los distintos componentes en función de las tres perspectivas estudiadas en el Capítulo 1, también corregidas de calendario y estacionalidad. La consistencia transversal busca que al final los componentes corregidos en los tres desgloses cuadren con el total corregido.

Capítulo 5

Creación de una aplicación y automatización en la obtención de datos

La programación necesaria para ajustar los modelos estudiados en los capítulos anteriores se ha desarrollado en `R` (R Core Team (2015)), software estadístico libre y de código abierto cada vez más empleado en el mundo empresarial. Uno de los problemas que se nos planteó fue el hecho de que los expertos en economía de ABANCA que van a hacer uso de estos modelos no son conocedores del lenguaje ni del programa. Como consecuencia de esto se generaba una dependencia de alguien que sí fuera conocedor, restando valor a la utilidad que la herramienta pudiera tener en una plataforma conocida por todos.

Como solución a este inconveniente se ha decidido crear una aplicación; de esta manera cualquier usuario no conocedor de `R` puede acceder a la herramienta en el momento que desee, ajustar los modelos e incluso interactuar con los indicadores. Esto ha sido posible gracias a la librería `Shiny` (Chang et al. (2015)), cuyo funcionamiento se describe a lo largo de la Sección 5.1.

Otro aspecto en el que se ha trabajado es la automatización del proceso de obtención de datos mediante las APIs que ofrece el IGE, como se detalla en la Sección 5.2.

5.1. R Shiny

Shiny es una librería de R creada para desarrollar aplicaciones que permiten al usuario manipular los datos y las distintas funciones de R sin necesidad de manejar código. Para construir estas aplicaciones no son necesarios conocimientos de HTML, CSS o Javascript, principales lenguajes de programación[1].

El objetivo de esta sección no es que el lector aprenda a crear una aplicación con Shiny, pero sí que entienda la arquitectura y el esquema de construcción de este tipo de aplicaciones; paso previo fundamental a la hora de poder desarrollar una. En la Sección 5.1.1 se presenta la librería y se describe el formato de cualquier aplicación construida en Shiny. Los diferentes objetos y el diálogo que hay por detrás entre ellos y que hace que la aplicación funcione se detalla en la Sección 5.1.2. Son numerosas las referencias que se han seguido tanto para elaborar este capítulo como para aprender el lenguaje de Shiny y lograr crear una aplicación. Destacan, entre otras, Grolemund (2015), Mora (2017) y Carmona y Subirana (2015).

Por último, mencionar que existen librerías de R en las que existen plantillas para crear una aplicación, más se ha optado por construir una desde cero, ya que así se puede adaptar al gusto y

[1]No obstante, de cara a personalizar la aplicación sí que son de gran utilidad estos lenguajes, siendo la aplicación creada en Shiny compatible con todos ellos.

modificarla en función de las necesidades que vayan surgiendo.

5.1.1. Introducción y arquitectura

Shiny se basa en un esquema de programación reactiva[2], donde se vinculan de forma dinámica los objetos de entrada (inputs) con los de salida (outputs). De esta manera, cada vez que el usuario hace una modificación todo se actualiza.

Toda aplicación Shiny para ser operativa necesita de dos patas; una página web que muestra la aplicación y un ordenador que esté ejecutando R. Este último puede ser el propio ordenador, en local, o algún servicio de hosting especializado para Shiny (ShinyApps.io, por ejemplo).

El esquema de funcionamiento se basa en dos partes que interactuan entre ellas, `UI` y `server`. Diferenciarlas es fundamental a la hora de entender el funcionamiento de una aplicación.

UI (interfaz gráfica): En esta parte se describe la interfaz gráfica de la aplicación, tanto su aspecto como su diseño. Se reciben los inputs y se muestran los outputs, de los cuales se gestiona la salida visual.

Server (código que alimenta la aplicación): Es la parte en la que se hace el "trabajo" necesario para que la aplicación funcione. Se puede ejecutar cualquier código que se ejecutaría en un script de R con la diferencia de que los parámetros o variables que se emplean pueden ser definidos por el usuario (inputs) a través de widgets de entrada en la parte `UI`; como resultado se obtienen elementos reactivos[3]. Los elementos reactivos que se quieren presentar en la aplicación se denominan outputs y se mandan a la parte `UI`. Se puede decir que en el `server` se realizan los cálculos necesarios en R para, a partir de los inputs, obtener los outputs.

El esquema de la programación dinámica así como el diálogo entre las partes, se muestra en la Figura 5.1.

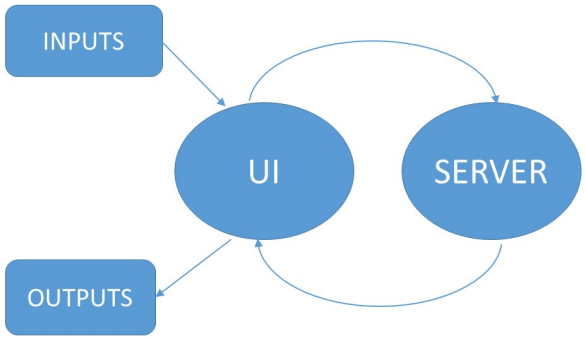

Figura 5.1: Diagrama de la reactividad de cualquier aplicación Shiny.

[2]Este tipo de programación se caracteriza por valores que cambian en el tiempo y expresiones que registran esos cambios.

[3]Se emplea el término reactivo debido a que si los inputs cambian, estos elementos se actualizan y también cambian.

Aunque toda aplicación de Shiny consta de las dos partes anteriormente señaladas, existen dos formas de construirla:

Opción 1: En un archivo `App.R` se recogen las dos partes y mediante la función `shinyApp` se convierten en una aplicación.

Opción 2: Se separa cada parte en un archivo diferente. La aplicación estará compuesta por un archivo ui.R y otro archivo server.R. En este caso, es innecesario emplear la función `shinyApp`.

En la Figura 5.2 se muestra cómo es la construcción en cada una de las dos opciones.

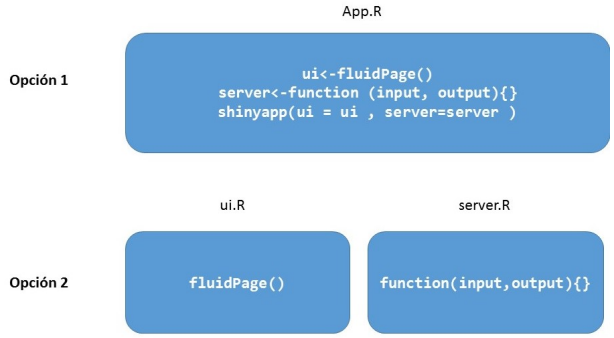

Figura 5.2: Construcción de una aplicación en Shiny.

A mayores, se suele incluir una carpeta llamada `www` donde se guardan diferentes recursos que se emplean en la aplicación, como imágenes o scripts.

5.1.2. Esquema de funcionamiento

El esquema de funcionamiento de una aplicación se muestra en la Figura 5.3. Primero se leen los inputs que el usuario introduce a través de la interfaz. Estos inputs se emplean en el Server para generar objetos reactivos. Por último, algunos de estos objetos reactivos (outputs) se envían a la interfaz de la aplicación. Cada vez que el usuario modifica un input, los outputs que dependen de este input se actualizan, de acuerdo con el esquema de reactividad que se mostró en la Figura 5.1.

Figura 5.3: Funcionamiento de una aplicación en Shiny.

Cada una de las etapas anteriores está caracterizada por una sintaxis y una arquitectura propia que se explican a continuación. Para poder comprender esto, se recurre a un ejemplo muy sencillo. El objetivo será diseñar una aplicación en la que el usuario introduzca un número natural comprendido entre 1 y 50 y se le devuelva este número más 5.

Lectura de Inputs

Shiny dispone de una gran variedad de widgets para introducir los inputs; algunos se muestran en la Figura 5.4. Cualquiera de ellos tiene los siguientes argumentos:

- Un nombre para acceder al valor del widget que se define en el campo `inputId`.
- Una etiqueta que aparecerá en el widget que se define en el campo `label`.
- Otros argumentos que dependen del widget seleccionado.

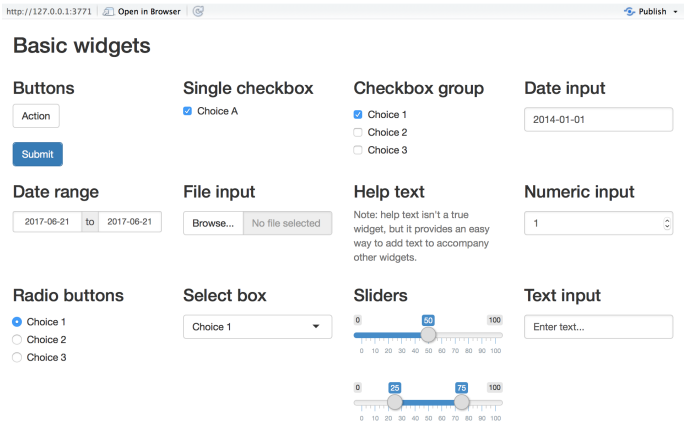

Figura 5.4: Algunos de los widgets de Shiny.

Para llamar desde la parte `server` a los inputs se emplea el comando `input$inputId`, siendo `inputId` el nombre del input.

Ejemplo: Para que el usuario pueda introducir un número natural se emplea el widget selector `numericInput(inputId="n_entrada",label="Introduzca n",value=2,min=1,max=50)`. Por defecto, el número que aparece seleccionado es el 2 y se pueden introducir valores entre 1 y 50, tal y como se pretendía. Para llamar a este valor introducido por el usuario desde la parte `server` se emplea el comando `input$n_entrada`.

Construcción de elementos reactivos

Un elemento reactivo es un elemento que se ha generado en la parte `server` mediante funciones de R y con información que proviene de inputs. Su construcción requiere de una sintaxis especial:

- Aquellos elementos reactivos que se deseen enviar a la parte `ui` y mostrar en la aplicación se denominan outputs. Los outputs se definen mediante la sintaxis `output$nombre_del_output` y se construyen a través de una función `render`* bajo el siguiente esquema.

  ```
  ouput$nombre_del_output<-render*({
  código R necesario para crear el objeto, en el que se pueden incluir
   inputs como input$inputId
  })
  ```

  Algunas de las clases de funciones `render`* se muestran en el Cuadro 5.1.

- Si se quiere crear un objeto reactivo que no se vaya a mostrar en la aplicación se emplea la función `reactive`. Se define el objeto como se definen los elementos en R; la única diferencia con la sintaxis habitual es que para llamar a este objeto se empleará el nombre con el que se definió seguido de ().

```
output$objeto_reactivo<-reactive{(
código R necesario para crear el objeto, en el que se pueden incluir
 inputs como input$input_Id
)},
```

y se llamará como `objeto_reactivo()`.

Función	Resultado
`renderPlot`	Define un gráfico reactivo
`renderText`	Convierte un vector de caracteres en un elemento reactivo
`renderTable`	Define una tabla reactiva

Cuadro 5.1: Algunas funciones del tipo `render`*.

Ejemplo: Dado el input `n_entrada`, se desea generar un output que devuelva el valor `n_entrada+5`; para ello es necesario construir un objeto de salida que se llamará `valor_salida`. Como lo que se quiere mostrar es un texto, un valor numérico, se emplea la función `renderText`.

```
ouput$valor_salida<-renderText{(
input$n_entrada+5
)}
```

Visualización de los outputs

Una vez que se han construido outputs en la parte `server`, el objetivo es que estos se devuelvan a la parte `UI` para que se muestra en la interfaz. Para ello es necesario emplear una función del tipo `*Output("nombre_del_output")`. Esta función depende de la función `Render*` que se empleó para definir el output. En el Cuadro 5.2 se muestran algunas de ellas.

Función	Función a la que acompaña	Resultado
`plotOutput`	`renderPlot`	Crea un gráfico como elemento de salida
`textOutput`	`renderPrint`	Crea un texto como elemento de salida
`tableOutput`	`renderTable`	Crea una tabla como elemento de salida

Cuadro 5.2: Algunas funciones del tipo `*output` con las funciones `render*` a las que acompañan.

Ejemplo: Generado el output `valor_salida` ahora se quiere que aparezca en la pantalla de la aplicación. Como el objeto se construyó con una función `renderPrint`, para mostrarlo en la interfaz será necesario una función `textOutput`.

```
textOutput("salida")
```

Aplicación resultante del ejemplo

En la Figura 5.5 se muestra la aplicación que se ha construido a modo de ejemplo. Para construirla se ha generado un único archivo en el que se alojan las dos partes. El número que introduce el usuario es un dato de entrada, y de él depende la salida (número que se muestra debajo). Se aprecia claramente

el diálogo que hay entre las dos partes, la reactividad. Al cambiar el input (de 2 a 4) , el output también cambia (de 7 a 9).

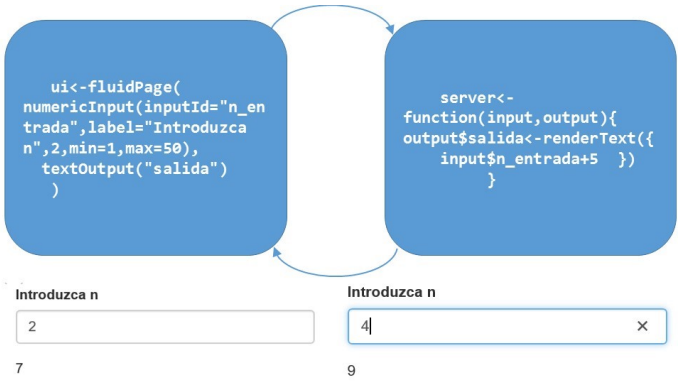

Figura 5.5: Código del ejemplo construido y aplicación.

R-Shiny utiliza de una manera simplificada los estándares del lenguaje HTML. En el siguiente ejemplo se muestra las equivalencias entre la función de Shiny y el código HTML necesarios para generar un párrafo de texto en negrita.

```
 # Lenguaje HTML                              #Funciones de Shiny
 <p>
<strong> Texto en negrita </strong>          p(strong("Texto en negrita"))
 </p>
```

De cara a diseñar la página existen muchas funciones. Se suele emplear `fluidPage`, que es un diseño de página fluida que consiste en filas que a su vez se dividen en columnas. Existen muchas opciones para estructurar la página, como las que se muestran en la Figura 5.6.

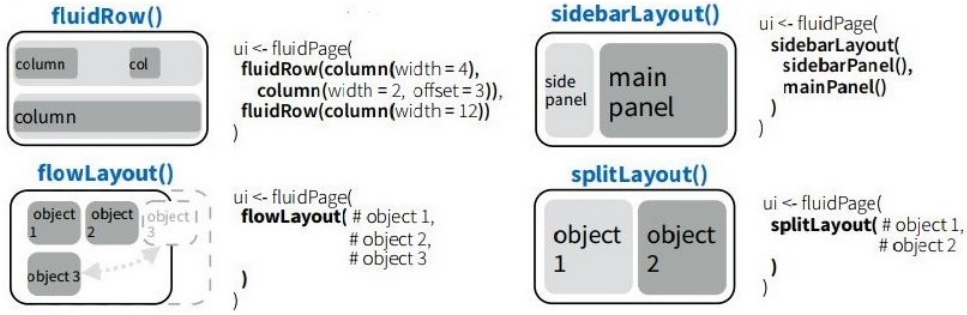

Figura 5.6: Distintos diseños de página.

5.2. Obtención de los datos

Para llevar a cabo toda la modelización descrita, se necesitan datos tanto de la variable objeto de estudio, el PIB, como de las candidatas a ser variables explicativas. Todos los datos que se han empleado están disponibles en la página del IGE y se ha implementado una rutina para consultarlos de manera automática. En la Sección 5.2.1 se describen las APIs de las dispone el IGE y cuál es

su funcionamiento, junto con algunos ejemplos. Con estas APIs se puede acceder a los datos de una manera mucho más ágil y dinámica respecto a como se venía haciendo en la Entidad, como se relata a lo largo de la Sección 5.2.2. Las referencias seguidas han sido las indicaciones que el propio IGE proporciona y la experiencia personal adquirida como usuarios de la herramienta.

5.2.1. Proceso de automatización

El IGE dispone de tres APIs [4] mediante las que facilita el acceso a sus datos. Estas herramientas permiten acceder mediante direcciones url a toda la información estadística disponible en la base de datos del IGE. El uso de estas APIs conlleva la aceptación de las condiciones de uso que se especifican en la propia página web, a la que se accede de la siguiente manera:

`http://www.ige.eu` ⇒ `Servicios` ⇒ `Descarga de información. APIs`

Para acceder a la información estadística es necesaria una url y basta con copiarla para acceder a los datos de la consulta. Mediante pequeñas modificaciones de esta url se pueden extraer datos similares pero con alguna modificación (series en variaciones o corregidas de calendario y estacionalidad, por ejemplo). Los formatos de descarga que se ofrecen son csv y json.

La sintaxis de la url es `http://www.ige.eu/igebdt/igeapi/csv/datosserc/codigoserie`, en el caso de que la extensión deseada sea csv. ASS es el código de la serie de afiliaciones a último día de mes; si se emplea la url `http://www.ige.eu/igebdt/igeapi/csv/datosserc/ASS`, la consulta resultante es:

```
"serie","codtempo","tempo","dato"
"ASS","199001","1990/Xaneiro",755065
"ASS","199002","1990/Febreiro",756661
              ...
```

Otra forma de acceder a la información es mediante **Descarga de tablas**, en cuyo caso la url de consulta tiene la siguiente sintaxis `http://www.ige.eu/igebdt/igeapi/datos/código de la tabla/[parámetros de selección]`. A las tablas se puede acceder desde el panel de **Información Estadística por Temas** de la página principal del IGE y seleccionando las series deseadas. En el caso de la serie del PIB, la ruta para acceder es:

`http://www.ige.eu` ⇒ `Economía` ⇒ `Cuentas económicas trimestrales` ⇒
⇒ `Índices de volumen.Referencia año 2010=100.Unidad:índices 2010=100` ⇒ `Máis datos.`

Una vez que se ha seleccionado una tabla, basta con ir a `Copiar URL de descarga` para que en el panel superior de la pantalla aparezca un cuadro en el que genera la url de la consulta. En el caso de la serie del PIB en niveles, corregida de calendario y estacionalidad y para todos los años en los que se disponen datos, características seleccionadas por defecto, la url resultante es:

```
http://www.ige.eu/igebdt/igeapi/datos/7436/0:199513:199514:199515:199516:199613:199614:199615:
199616:199713:199714:199715:199716:199813:199814:199815:199816:199913:199914:199915:199916:
200013:200014:200015:200016:200113:200114:200115:200116:200213:200214:200215:200216:200313:
200314:200315:200316:200413:200414:200415:200416:200513:200514:200515:200516:200613:200614:
200615:200616:200713:200714:200715:200716:200813:200814:200815:200816:200913:200914:200915:
200916:201013:201014:201015:201016:201113:201114:201115:201116:201213:201214:201215:201216:
201313:201314:201315:201316:201413:201414:201415:201416:201513:201514:201515:201516:201613:
201614:201615:201616:201713:201714:201715,1:1,2:1,3:0,4:0
```

[4] API son las siglas de Application Programming Interfaces (Interfaces de Programación de Aplicaciones). Una API es una interfaz para dar un acceso limitado a la base de datos de un servicio web, evitando que se conozca o acceda al propio código fuente de la aplicación original. Ejemplos de API's son Google Maps o Twitter. Para más información puede consultarse `https://www.internetya.co/que-es-y-para-que-sirve-una-api/`.

Obtener las serie en variaciones o sin corregir es sencillo; basta con seleccionar las características deseadas y la url que se genera en la parte superior de la pantalla se modificará. Por ejemplo, para obtener la serie en variaciones respecto al trimestre anterior solo hay que sustituir los últimos dígitos de la url `4:0` por `4:1`. A continuación, se muestra la consulta generada con la url descrita.

```
"CodTempo","Tempo","Prezos","Tratamento do dato","Compoñentes","Dato","CodEspazo","Espazo","DatoN","DatoT"
199513,"1995/I","Índices de volume encadeados, referencia ano 2010=100",
"Datos corrixidos de estacionalidade e calendario","PIBpm",
"Niveis","12","12 Galicia",64.65,"64,65"
199514,"1995/II","Índices de volume encadeados, referencia ano 2010=100",
"Datos corrixidos de estacionalidade e calendario","PIBpm",
"Niveis","12","12 Galicia",65.12,"65,12"
```

La estructura de las consultas empleando esta metodología es la siguiente:

- La primera línea contiene los nombres de las variables y las siguientes los clasificadores.
- El separador utilizado es la coma.
- Las variables alfanuméricas aparecen entre comillas.
- Las dos últimas columnas contienen los datos en formato numérico y alfanumérico, respectivamente.
- El campo DatoN se corresponde con los datos en formato numérico. Cuando no esté disponible aparecerá en blanco y el campo DatoT muestra la explicación de que este dato no se encuentra disponible.
- El campo DatoT se corresponde con los datos en formato texto. Las opciones son las siguientes:
  - *: secreto estadístico.
  - .. : sin datos o non consta.
  - - : no procede incluir datos.

5.2.2. Utilidad y principales ventajas

Como se ha detallado en el Capítulo 1, desde el Departamento de Planificación Estratégica y PMO de Abanca se hace un seguimiento de la evolución económica de Galicia. Para esta tarea, se dispone de un archivo Excel, a modo de base de datos, en el que se encuentran todos los indicadores y series de coyuntura de las que se hace seguimiento. A partir de estos datos, los expertos en economía estudian el comportamiento de los indicadores (aceleraciones, cambios de tendencia, tensionamientos, etcétera).

El procedimiento que se estaba empleando hasta el momento para alimentar esta base de datos era bastante tedioso. Con la publicación de cada nuevo dato la persona encargada de hacer el seguimiento debía acceder a la página web del IGE, buscar la tabla correspondiente, copiar el dato y, por último, pegarlo en la columna adecuada de la tabla de datos. Como principales complicaciones cabe resaltar:

- No todos los indicadores se actualizan el mismo día.
- Los últimos datos publicados la mayoría de las veces son provisionales y están sujetos a correcciones, por lo que es necesario comprobar que coinciden y que no ha habido modificaciones en las sucesivas actualizaciones.
- Existen indicadores que no tienen una fecha fija de publicación, por lo que es necesario estar pendiente de su publicación.

El descubrimiento de estas APIs agiliza de manera significativa el procedimiento anteriormente descrito y evita las complicaciones señaladas. Se ha desarrollado una rutina desde `R` mediante la cual se pueden leer directamente los datos que se deseen y guardarlos en un archivo Excel para posteriormente hacer el estudio macroeconómico y llevar a cabo el seguimiento.

En lo relativo a la modelización llevada a cabo a lo largo de este trabajo también se han simplificado las cosas, ya que leer los datos de archivos y estar pendiente de que estos estén actualizados pasan a ser tareas innecesarias.

Parte II

Parte práctica

Capítulo 6

Descripción de las variables

En este capítulo se describen las variables con las que se va a trabajar. Como ya se indicó en el Capítulo 5, se ha creado una aplicación y se ha implementado una rutina para que las series se descarguen automáticamente. De esta manera, se tiene la seguridad de que siempre se trabaja con datos actualizados.

En la Sección 6.1 se analiza la serie del PIB y se muestra el apartado creado en la aplicación para trabajar con ella. Los indicadores con los que se hace el seguimiento macroeconómico de Galicia, y que son susceptibles de ser variables explicativas a la hora de modelizar el PIB se describen en la Sección 6.2. A modo ilustrativo, se muestra la pestaña que se ha creado en la aplicación para realizar el seguimiento macroeconómico de Galicia, además de algunas otras pestañas que permiten ver el aspecto de la aplicación desarrollada.

6.1. La serie del PIB

En este trabajo, siempre que se emplea el término serie del PIB se hace referencia al índice del PIB real corregido de calendario y estacionalidad y sujeto a un tratamiento de consistencia temporal y transversal (véase Capítulo 1). Esta serie comienza el primer trimestre del año 1995 (en adelante, 1995-Q1), tiene una frecuencia trimestral y se publica con un desfase de aproximadamente 50 días.

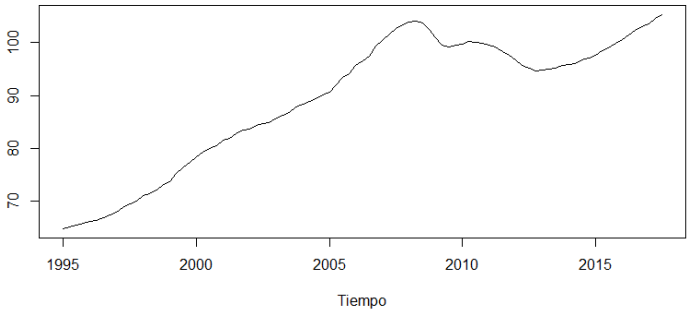

Figura 6.1: Serie del PIB que se quiere modelizar.

En la Figura 6.1 se muestra la serie objeto de estudio: El PIB corregido de estacionalidad y calendario, y con el ajuste de consistencia temporal y transversal. Como medidor de la economía que es, se

pueden apreciar el crecimiento de los años de la burbuja económica, seguido de los años de crisis. Se puede observar que ahora se está en un periodo de crecimiento económico.

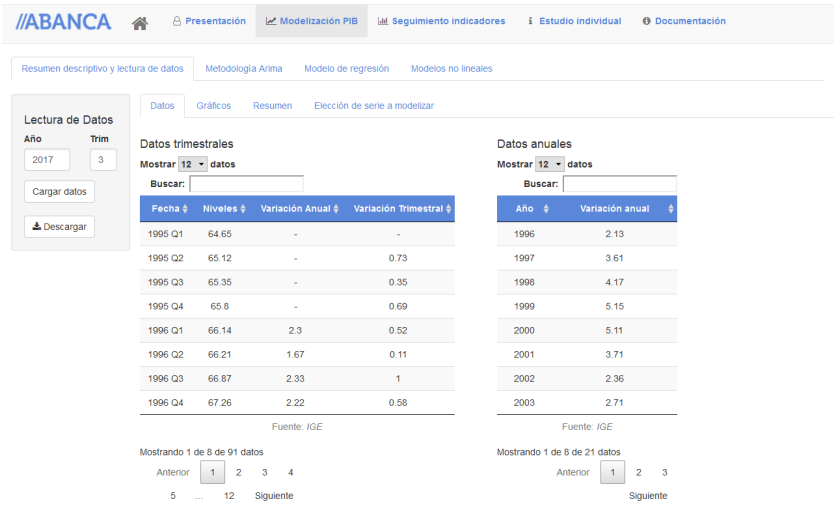

Figura 6.2: Aspecto de la pestaña de estudio de la serie del PIB en la aplicación creada.

Los estudios sobre esta serie no sólo se centran en analizar su valor; se estudian también las tasas intertrimestrales e interanuales, ya que muchas veces lo interesante para describir la situación económica es una comparativa respecto a un periodo de referencia. Esto se puede analizar en una de las pestañas creadas en la aplicación, cuyo aspecto se muestra en la Figura 6.2. Las tablas creadas permiten al usuario buscar una fecha o un dato y ordenar las columnas, entre otras tareas.

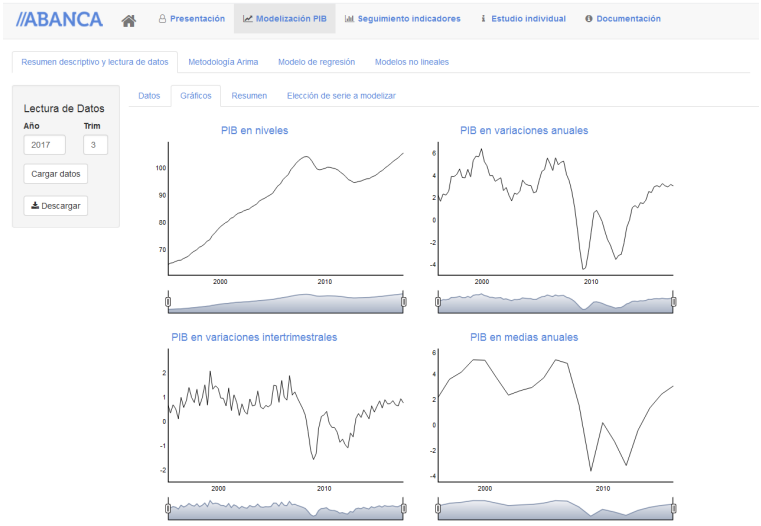

Figura 6.3: Aspecto de la pestaña de estudio gráfico de la serie del PIB en la aplicación creada.

En la Figura 6.3 se muestra la parte de la aplicación en la que se puede estudiar gráficamente la serie del PIB. En esa pestaña se muestra la evolución en diferentes formatos (en niveles, variaciones

trimestrales intertrimestrales e interanuales y variaciones anuales) mediante gráficos dinámicos, con los que el usuario puede interactuar (al pulsar sobre un punto se muestra el valor y es posible hacer zoom y ver sólo un periodo de tiempo). A mayores, también es posible descargar la serie en un formato `.csv`.

6.2. Indicadores macroeconómicos

Se describen en esta sección las variables que se han empleado a lo largo de este trabajo y que han sido seleccionadas mediante alguno de los siguientes criterios:

- Forman parte del seguimiento económico que se hace en ABANCA.
- Eran variables que de manera recurrente aparecían en la revisión bibliográfica que se hizo acerca de modelos para explicar el PIB.
- Se han elegido por el corto desfase de publicación.

A continuación se muestran cuáles son estas variables, la nomenclatura y la unidad en las que se miden.

$\{X_t^1\}$**:** Media mensual de trabajadores afiliados en alta laboral en la Seguridad Social [Unidad: nº de afiliaciones].

$\{X_t^2\}$**:** Trabajadores afiliados en alta laboral en la seguridad social el último día del mes [Unidad: nº de afiliaciones].

$\{X_t^3\}$**:** Paro registrado (obtenida del SPEE) [Unidad: Personas].

$\{X_t^4\}$**:** Matriculación vehículos. Turismos [Unidad: Unidades].

$\{X_t^5\}$**:** Ventas de combustibles líquidos. Gasóleos [Unidad: Miles TM].

$\{X_t^6\}$**:** Índice de entrada de pedidos en la industria [Unidad: Índice].

$\{X_t^7\}$**:** Índice de producción industrial [Unidad: Índice].

$\{X_t^8\}$**:** Importaciones totales con destino Galicia [Unidad: Miles de Euros].

$\{X_t^9\}$**:** Exportaciones totales con origen Galicia [Unidad: Miles de Euros].

$\{X_t^{10}\}$**:** Compraventa de viviendas. Total [Unidad: número].

$\{X_t^{11}\}$**:** Ocupación en los establecimientos hosteleros. Pernoctaciones [Unidad: Personas].

$\{X_t^{12}\}$**:** Transporte total aéreo. Pasajeros [Unidad: Personas].

$\{X_t^{13}\}$**:** Transporte marítimo. Mercancías cargadas, descargadas y transbordadas. Total [Unidad: Miles TM].

$\{X_t^{14}\}$**:** Índice general de la cifra de negocios del sector servicios (CNAE-2009). Base 2010 [Unidad: Sin unidad].

$\{X_t^{15}\}$**:** IPC base 2016 [Unidad: Sin unidad].

Estas variables se han clasificado en 5 grandes grupos: Mercado laboral (ML), Construcción y vivienda (VIV), Energía e industria (IND), Comercio Exterior (CE) y Servicios (SERV). En el Cuadro 6.1 se muestra un resumen de estas variables: el primer dato disponible, la frecuencia y el desfase en la publicación. Además, se indica la fuente y si están corregidas de calendario y estacionalidad, así como el código que se emplea en la aplicación para referirse a ellas.

Serie	Frec.	Inicio	Publicación	Correg.	Fuente	Sector	Código
$\{X_t^1\}$	12	2004-M1	t+4 días	Sí	IGE	ML	AMSS
$\{X_t^2\}$	12	1990-M1	t+ 14 días	No	IGE	ML	ASS
$\{X_t^3\}$	12	2001-M2	t+ 3 días	Sí	IGE	ML	PAROREG
$\{X_t^4\}$	12	1980-M1	t+ 20 días	Sí	IGE	IND	MATTUR
$\{X_t^5\}$	12	1993-M1	-	Sí	IGE	IND	CPPGASOL
$\{X_t^6\}$	12	2010-M1	t+ 50 días	No	IGE	IND	IEP_GA
$\{X_t^7\}$	12	2002-M1	t+ 50 días	Sí	IGE	IND	IPI10XE
$\{X_t^8\}$	12	1988-M1	t+ 50 días	Sí	IGE	CE	CXX
$\{X_t^9\}$	12	1988-M1	t+ 50 días	Sí	IGE	CE	CXM
$\{X_t^{10}\}$	12	2007-M1	t+ 40 días	No	IGE	VIV	GAL_CVIV_TOTAL
$\{X_t^{11}\}$	12	1999-M1	t+ 23 días	Sí	IGE	SERV	EOHP_04
$\{X_t^{12}\}$	12	1980-M1	-	Sí	IGE	SERV	TAP
$\{X_t^{13}\}$	12	1980-M1	-	Sí	IGE	SERV	TMM
$\{X_t^{14}\}$	12	2005-M1	t+ 50 días	No	IGE	SERV	SERV_CN_2010
$\{X_t^{15}\}$	12	2002-M1	t+ 15 días	No	IGE	SERV	IPC1600general

Cuadro 6.1: Resumen de los indicadores empleados

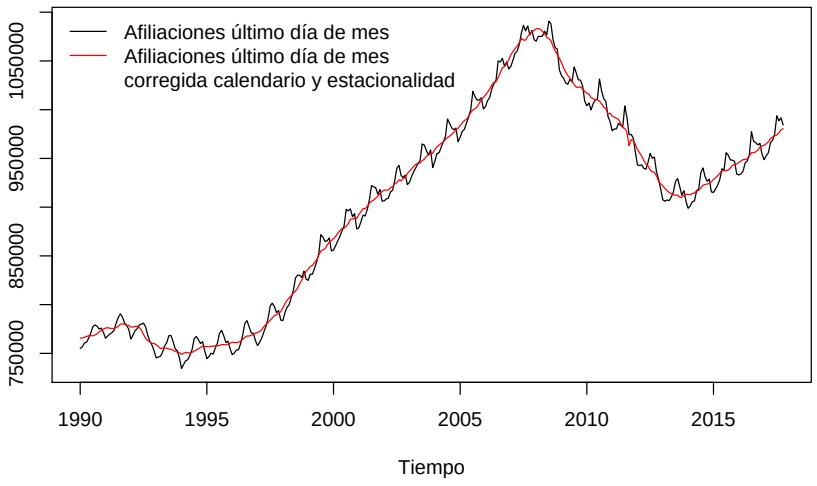

Figura 6.4: Afiliaciones a la S.Social en niveles y corregida de calendario y estacionalidad.

La notación empleada es $\{X_t^j\}$ para hacer referencia a la serie en niveles y $\{Z_t^j\}$ para referirse a la serie corregida de calendario y estacionalidad. Como se verá en el Capítulo 7, ha sido necesario corregir de calendario y estacionalidad la serie $\{X_t^2\}$ de afiliados a la Seguridad Social el último día de mes por no encontrarse corregida en la base de datos del IGE. En la Figura 6.4 se muestra un gráfico de la serie en niveles y la serie corregida.

Motivado por el seguimiento económico que se hace en ABANCA (véase Capítulo 1), se ha incorporado una pestaña dentro de la aplicación en la que se puede llevar a cabo esta tarea de una manera menos laboriosa y más rápida a como se estaba haciendo en la Entidad. Incluir estas variables en esta herramienta de estudio del PIB es una buena opción, ya que muchas veces no es suficiente únicamente el PIB para comprender la situación económica, sino que es importante estudiar cómo se están comportando el resto de indicadores.

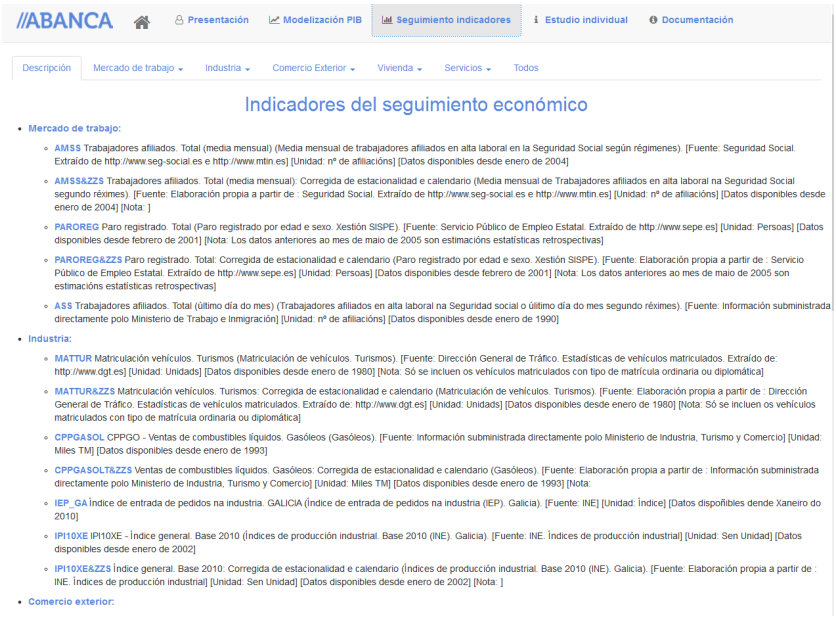

Figura 6.5: Aspecto de la pestaña donde se hace la descripción de los indicadores de los que se hace seguimiento y están disponibles en la aplicación creada.

Se han dividido los indicadores en cinco pestañas en función del grupo al que pertenecen y se muestran, tanto los datos en niveles como en variaciones, en gráficos dinámicos y en tablas. Para aquellas series que se publican corregidas, se muestra de manera conjunta la serie sin corregir y la serie corregida, ya que a ojos del analista esto facilita el estudio. En la Figura 6.5 se muestra la pestaña en la que detallan las variables que se estudian.

En la Figura 6.6 se muestra el seguimiento de los indicadores en niveles del grupo Comercio Exterior; las variables de exportaciones e importaciones.

Toda la programación se ha hecho con el programa estadístico R (R Core Team (2015)) y se han empleado principalmente las librerías `Shiny` (Chang et al. (2015)), `TSA` (Chan y Ripley (2015)) y `forecast` (Hyndman et al. (2017)).

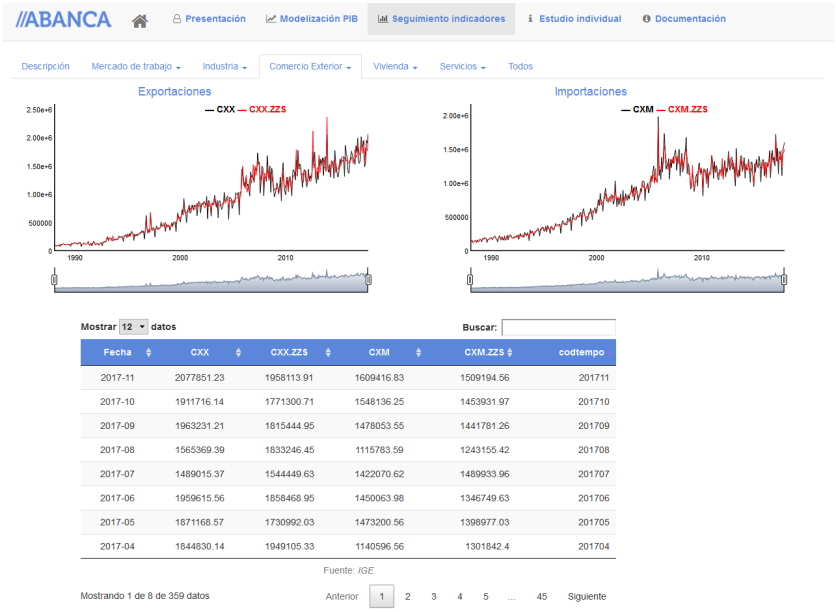

Figura 6.6: Aspecto de la pestaña de seguimiento de los indicadores, en niveles, del grupo Comercio Exterior.

Capítulo 7

Modelos ARIMA

En este capítulo se aplica la metodología Box-Jenkins estudiada en el Capítulo 2, siendo el objetivo ajustar un modelo $ARIMA(p,d,q) \times (P,D,Q)_4$ a la serie del PIB descrita en el capítulo anterior. El procedimiento llevado a cabo se describe a lo largo de la Sección 7.1. En la Sección 7.2 se muestra el aspecto y se describe la parte de la aplicación en la que se ha implementado esta metodología.

7.1. Procedimiento seguido

Uno de los principales problemas a la hora de tratar de modelizar la serie del PIB es que no es una serie temporal con un tratamiento clásico (conviene recordar que es un índice corregido de calendario y estacionalidad y sometido a un tratamiento de consistencia temporal y transversal). Como se ha visto en el Capítulo 4, con la incorporación de cada nuevo dato la serie histórica se modifica.

Para reflejar esto, en el gráfico de la izquierda de la Figura 7.1 se muestra la serie del PIB a fecha de marzo de 2017 (línea negra) y a fecha de noviembre del mismo año (línea roja). Si bien ambas series no son distintas, las discrepancias son más acusadas en los últimos instantes de la serie.

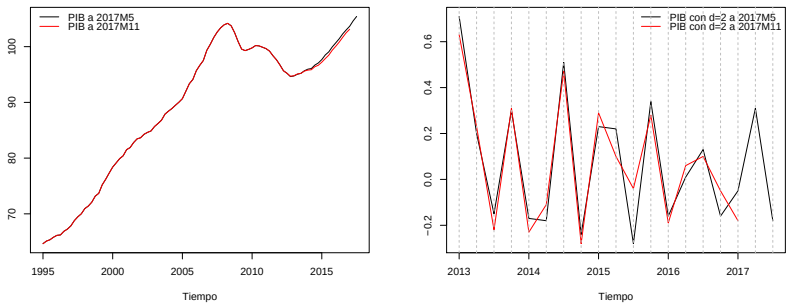

Figura 7.1: Serie del PIB publicada en diferentes fechas.

Como se verá a continuación, la serie que se va a modelizar es la del PIB una vez que se le ha aplicado una diferencia regular de orden dos ($d = 2$). En el gráfico de la derecha de la Figura 7.1 se aprecia como esta serie diferenciada cambia en función de la fecha de publicación.

El primer paso para ajustar un modelo Box-Jenkins es tener una serie estacionaria [1]. La serie del PIB con la que se trabaja no presenta estacionalidad, ya que está corregida de esta componente; prueba

[1] Entre las principales causas de falta de estacionariedad se encuentran la presencia de tendencia, heterocedasticidad o estacionalidad (véase Capítulo 2).

de ello es que en la Figura 7.1 no se aprecia ningún patrón repetitivo. En la Figura 7.2 se muestra el comportamiento de cada trimestre y, al contrario de lo que pasaba en el caso del consumo de gasolina del Capítulo 4 (Figura 4.2), no se aprecia ningún comportamiento característico en los trimestres.

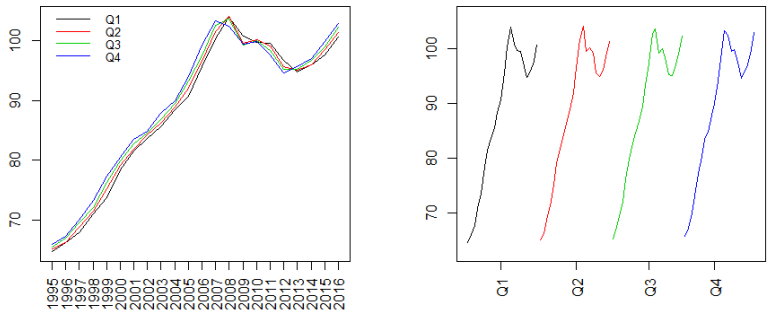

Figura 7.2: Serie del PIB por trimestres.

Por el contrario, en la Figura 7.1 sí se observa que la serie presenta una fuerte tendencia. Entre los años 2002 y 2009 tuvo lugar un fuerte crecimiento económico, que desencadenó entre los años 2009 y 2012 una fuerte crisis de la que la economía se está recuperando. La presencia de tendencia se confirma en la Figura 7.3 con las autocorrelaciones simples muestrales de la serie sin diferenciar (d=0), pues son muy elevadas y tardan en disminuir.

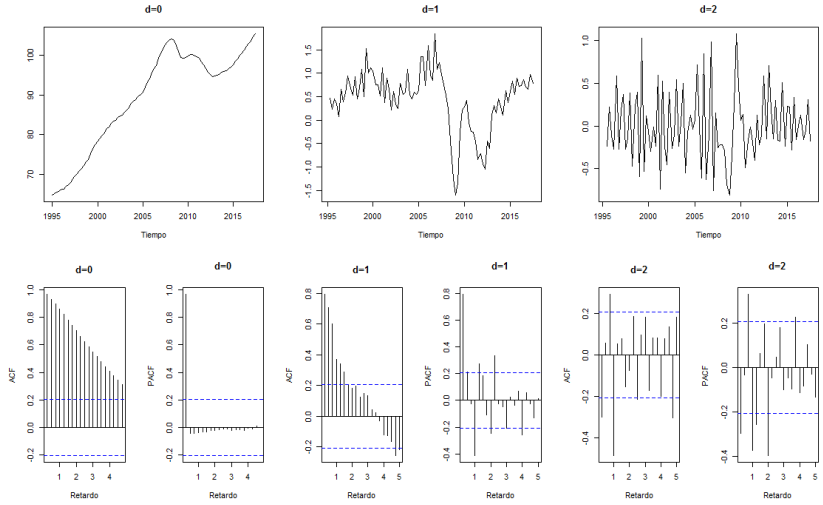

Figura 7.3: Serie en niveles y funciones de autocorrelaciones simples y parciales muestrales para la serie del PIB con diferencias regulares de distinto orden (d=0,1,2).

En los gráficos de la Figura 7.3 se muestra la serie del PIB tras aplicarle diferencias regulares de distintos órdenes. Si bien es cierto que con una diferencia no se consigue todavía una serie estacionaria, gráficamente todo parece indicar que dos diferencias regulares son suficientes.

Mediante el contraste de estacionariedad de Dickey–Fuller aumentado se confirma que son necesarias dos diferencias regulares para que la serie sea estacionaria, como se puede ver en el Cuadro 7.1.

Serie	Estadístico	Nivel crítico
d=0	-1.5836	0.7477
d=1	-2.5302	0.3579
d=2	-6.0338	0.01

Cuadro 7.1: Contraste de Dickey-Fuller aumentado para la serie del PIB tras aplicarle diferencias regulares de distintos órdenes.

Una vez que se ha establecido que son necesarias dos diferencias regulares para que la serie sea estacionaria, el siguiente paso es la identificación de un modelo $ARIMA(p,2,q)\times(P,0,Q)_4$.

Si se estuviera trabajando con una serie convencional, se habría optado por elegir un modelo $ARIMA(p_1,2,q_1)\times(P_1,0,Q_1)_s$. Si el modelo fuera válido, se conservaría en el sentido de que, con la incorporación de cada nuevo dato se reestimarían los parámetros, pero el modelo sería el mismo.

Sin embargo, debido a las circunstancias particulares de esta serie ya comentadas, sumadas a que es susceptible de ser revisada y modificada en cualquier momento, se ha elegido otra opción. En vez de elegir un modelo $ARIMA(p,2,q)\times(P,0,Q)_4$ y quedarse con él, se ha optado por elegir el mejor modelo válido en cada momento, es decir, se ha optado por un modelo ad-hoc. De esta manera, se tiene la seguridad de que por mucho que cambie la serie, se está eligiendo el mejor modelo. Además, con esta forma de proceder, se evita elegir un modelo que quede obsoleto en los próximos trimestres, haciendo la herramienta más funcional y con más vida útil. En ocasiones es complicado elegir un mejor modelo válido, por ello se ha seleccionado una pequeña batería de mejores modelos válidos.

Se ha construido una función llamada `selector1(serie,d_BIC)` que se aplica a la serie del PIB, pero que es operativa para cualquier otra serie temporal y cuyas etapas se describen a continuación.

Motor de ajuste de modelos

Dada la serie, se seleccionan los órdenes de las diferencias regulares y estacionales que son necesarias para que sea estacionaria [2]. Seleccionados los órdenes d y D, se inicia el motor de ajuste de modelos. Para cada p, q, P, y Q se ajusta un modelo $ARIMA(p,d,q)\times(P,D,Q)_s$, siendo s la frecuencia de la serie (en el caso del PIB, s=4). Cada modelo ajustado se valida, exigiendo que los residuos sean incorrelados y tengan media cero. Para contrastar la incorrelación se emplea el contraste de Ljung–Box con $h=1,\ldots,[0.25\dot{T}]$. Se ha implementado una función `Ind_autoc` que devuelve un 1 si se superan los $[0.25T]$ contrastes; en caso contrario se devuelve un 0.

En el momento que uno de ellos no se verifique, se descarta ese modelo y se ajusta el siguiente. Si el modelo supera los contrastes, se calcula el BIC y se guarda. El output de esta parte es un mensaje en el que se informa de cuántos modelos se han ajustado y cuántos han pasado los contrastes.

```
Se han ajustado 255 modelos, de los cuales 88 fueron válidos.
```

Selector de modelos

Con los modelos válidos del paso anterior, se selecciona aquel que tiene menor BIC y aquellos cuyo BIC diste de este menos de `d_BIC` unidades. En el caso del PIB, seleccionado `d_BIC=2`, resultan los modelos que se muestran en el Cuadro 7.2.

[2]Se emplean para ello las funciones `ndiffs` y `nsdiff` de la librería `forecast`(Hyndman et al. (2017)).

Modelo	BIC	p	d	q	P	D	Q
Modelo 1	65.697	0	2	3	0	0	2
Modelo 2	65.979	0	2	4	0	0	1
Modelo 3	65.589	1	2	2	0	0	1
Modelo 4	64.859	1	2	2	0	0	2
Modelo 5	64.197	2	2	1	0	0	1

Cuadro 7.2: Órdenes del modelo con el menor BIC y aquellos cuyo BIC dista menos de dos unidades de este.

Significación de parámetros

A cada uno de los modelos del paso anterior se les aplica una función para que resulten modelos con todos los parámetros significativos. Esta función itera los siguientes pasos hasta que todos los parámetros del modelo son significativos.

Paso 1: Se ajusta el modelo.
Paso 2: Se analizan los niveles críticos del contraste de que cada uno de los coeficientes del modelo sean cero. Si todos estos niveles críticos son menores que 0.05, se finaliza. Si por el contrario, algún nivel crítico es mayor que 0.05, se va al Paso 3.
Paso 3: Se detecta qué parámetro tiene un nivel crítico más grande, se fija a cero y se vuelve al Paso 1.

Es posible que al aplicar esta función a los modelos, algunos de ellos se conviertan en el mismo modelo. La función programada detecta esta coincidencia y elimina aquellos modelos repetidos. También es posible que alguno de los modelos resultantes de la revisión de los parámetros ya no sea válido, por este motivo se les vuelve a pasar una validación, descartando aquellos que dejen de ser válidos. En este paso se devuelven modelos son todos los parámetros significativos y válidos.

Para el caso del PIB, en el Cuadro 7.3 se muestran los modelos del Cuadro 7.2 una vez que se ha fijado a cero algunos y resultan todos significativos.

Modelo 1			Modelo 2			Modelo 3			Modelo 4			Modelo 5		
$\hat{\theta_1}$	-0.232	(0.087)	$\hat{\theta_1}$	-0.232	(0.087)	$\hat{\phi_1}$	0		$\hat{\phi_1}$	0.844	(0.106)	$\hat{\phi_1}$	0.422	(0.169)
$\hat{\theta_2}$	0.499	(0.104)	$\hat{\theta_2}$	0.499	(0.104)	$\hat{\theta_1}$	-0.232	(0.087)	$\hat{\theta_1}$	-0.989	(0.127)	$\hat{\phi_2}$	0.396	(0.102)
$\hat{\theta_4}$	0		$\hat{\theta_3}$	0.000		$\hat{\theta_2}$	0.499	(0.104)	$\hat{\theta_2}$	0.455	(0.099)	$\hat{\theta_1}$	-0.586	(0.163)
$\hat{\Theta_1}$	-1	(0.116)	$\hat{\theta_4}$	0.000		$\hat{\Theta_1}$	-1.000	(0.116)	$\hat{\Theta_1}$	-1.304	(0.167)	$\hat{\theta_2}$	-1.000	(0.068)
$\hat{\Theta_2}$	0		$\hat{\Theta_1}$	-1.000	(0.116)				$\hat{\Theta_2}$	0.304	(0.137)			

Cuadro 7.3: Estimación de los modelos con todos los parámetros significativos; entre paréntesis la desviación típica de la estimación.

Se observa que los modelos 1, 2 y 3 a pesar de ser en un principio modelos diferentes, se convierten en el mismo modelo al revisar la significación de los parámetros. Se descartan dos de ellos, obteniendo los 3 modelos que se muestran en el Cuadro 7.4. Estos modelos se vuelven a validar, devolviendo una tabla con el nivel crítico del contraste de media cero y el resultado de la función `Ind_Autoc` descrita

anteriormente. Aunque la normalidad no es necesaria, es deseable que los errores sean normales; por este motivo también se aplican los contrastes de normalidad.

Modelo 1			Modelo 4			Modelo 5		
$\hat{\theta}_1$	-0.232	(0.087)	$\hat{\phi}_1$	0.844	(0.106)	$\hat{\phi}_1$	0.422	(0.169)
$\hat{\theta}_2$	0.499	(0.104)	$\hat{\theta}_1$	-0.989	(0.127)	$\hat{\phi}_2$	0.396	(0.102)
$\hat{\theta}_3$	0		$\hat{\theta}_2$	0.455	(0.099)	$\hat{\theta}_1$	-0.586	(0.163)
$\hat{\Theta}_1$	-1	(0.116)	$\hat{\Theta}_1$	-1.304	(0.167)	$\hat{\theta}_2$	-1.000	(0.068)
$\hat{\Theta}_2$	0		$\hat{\Theta}_2$	0.304	(0.137)			

Cuadro 7.4: Estimación de los modelos resultantes.

Contraste	Modelo 1	Modelo 2	Modelo 3
Media Cero	0.514	0.754	0.831
Jarque Bera	0.002	0.000	0.334
Shapiro	0.037	0.013	0.525
Autocorrelación	1	1	1

Cuadro 7.5: Validación de los modelos.

Para el ejemplo con el que se está trabajando, esto se muestra en el Cuadro 7.5, donde se puede ver que todos los modelos son válidos, aunque no todos verifican la hipótesis de normalidad.

Predicciones

Dados los modelos válidos y con todos los parámetros significativos, es hora de hacer la predicción. Se ha optado por dar una predicción de cada modelo y una predicción conjunta en la que se pondera la predicción de cada modelo por el inverso del peso de los 5 últimos residuos al cuadrado. Es decir, si existen 3 modelos y $\hat{y}_{T+1}^{M1}$, $\hat{y}_{T+1}^{M2}$ y $\hat{y}_{T+1}^{M3}$ son las predicciones para cada uno de ellos; la predicción conjunta será:

$$\hat{y}_{T+1}^{Comb} = p_{M1}\hat{y}_{T+1}^{M1} + p_{M2}\hat{y}_{T+1}^{M2} + p_{M3}\hat{y}_{T+1}^{M3},$$

siendo $p_{Mj} = \frac{r^{Mj}}{\sum_j r^{Mj}}$, con $r^{Mj} = \sum_{t=T}^{T-4} \frac{1}{(\hat{w}_t^{Mj})^2}$.

Puesto que el valor puntual de la predicción sin tener los valores históricos de la serie es difícil de interpretar, la función devuelve los valores de la serie de los dos últimos años más las predicciones. Para el caso del PIB, el resultado se muestra en el Cuadro 7.6.

Fecha	Datos reales	Modelo1	Modelo2	Modelo3	Comb.
2016-Q1	100.680	-	-	-	-
2016-Q2	101.420	-	-	-	-
2016-Q3	102.290	-	-	-	-
2016-Q4	103.000	-	-	-	-
2017-Q1	103.660	-	-	-	-
2017-Q2	104.630	-	-	-	-
2017-Q3	105.420	-	-	-	-
2017-Q4	-	106.287	106.252	106.290	106.275
2018-Q1	-	107.269	107.279	107.265	107.271

Cuadro 7.6: Últimos datos reales y predicciones.

En la Figura 7.4 se esquematiza la rutina de la función `selector1(serie,d_BIC)` anteriormente descrita:

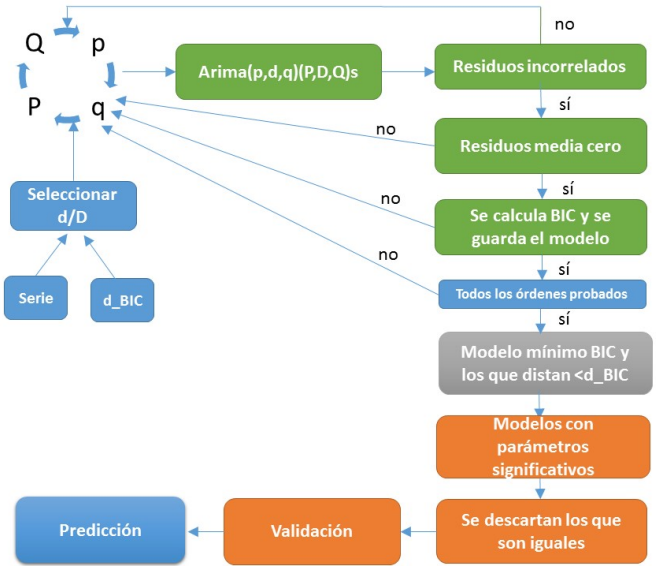

Figura 7.4: Proceso que se lleva a cabo para elegir de manera automática un modelo autoregresivo.

7.2. Aplicación

Este proceso automático se ha implementado en la aplicación creada, además de incluir la posibilidad de hacer un ajuste manual. Para que todo quedara más amigable, las salidas de R que son código (por ejemplo la función `summary`) se han convertido en tablas. Se ha organizado esta parte en 4 pestañas:

Portada: Se detalla lo que se hace en cada una de las siguientes pestañas.

Estudio preliminar: Se muestra la serie a modelizar, junto con su función de autocorrelaciones simples y parciales muestrales. En la pestaña Gráficos serie diferenciada se muestran el orden de las diferencias regulares y estacionales recomendadas para que la serie sea estacionaria. Además, en función de las diferencias d y D que desee el usuario, se muestran tres gráficas de la serie con estas diferencias: en niveles y las funciones de autocorrelaciones simples y parciales muestrales.

Ajuste manual: En esta pestaña el usuario puede introducir los órdenes del modelo $ARIMA(p,d,q)\times(P,D,Q)_s$ que desea ajustar. El programa ajusta este modelo y devuelve la estimación de los parámetros. Además, devuelve el modelo una vez que se han fijado a cero los parámetros no significativos mediante la rutina descrita. Se muestra también la validación del modelo y las predicciones para el horizonte elegido, tanto en formato gráfico como en una tabla.

Ajuste automático: Es aquí donde está implementada la función `selector` que se ha descrito anteriormente. Se muestran los órdenes de los modelos ajustados y los modelos que surgen de estos al hacer todos los parámetros significativos. Además, para estos modelos finales se proporciona también una validación, de forma que el usuario puede ver si los residuos verifican la normalidad o no. Por último, se devuelven las predicciones en niveles, variaciones trimestrales interanuales, variaciones intertrimestrales y variaciones anuales.

En la Figura 8.3 se muestra el aspecto de la parte de la aplicación en la que se encuentra todo lo detallado anteriormente.

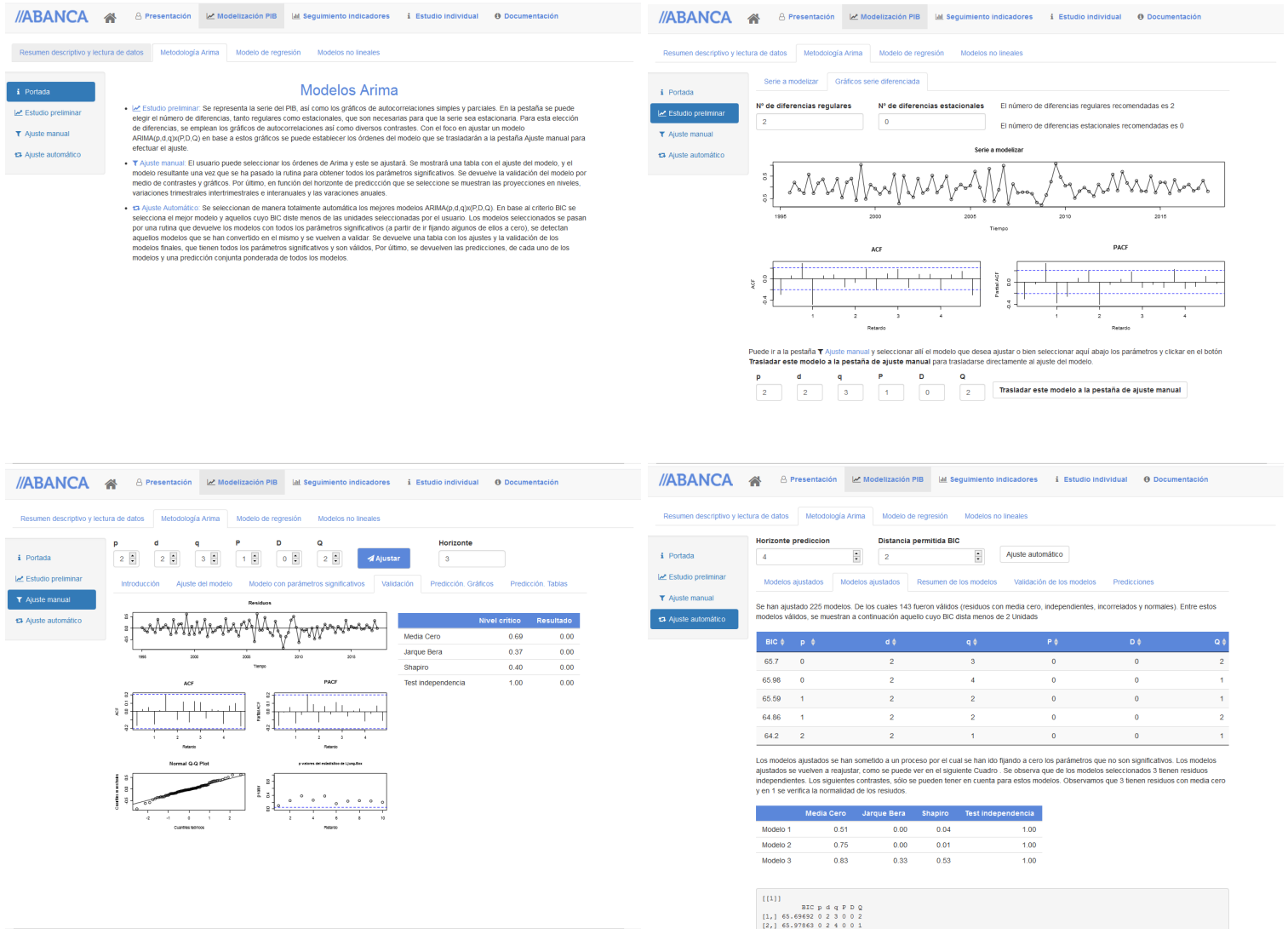

Figura 7.5: Aspecto de las pestañas de la aplicación en las que se ha implementado la metodología Box-Jenkins.

Capítulo 8

Modelos con variables explicativas

En esta capítulo se describen los modelos de regresión dinámica empleados para modelizar el PIB. Estos modelos se han elegido y ajustado con el objetivo de poder anticiparse a lo que está aconteciendo en la economía antes de que se publique la cifra del PIB (recuérdese que se publica con un desfase aproximado de 50 días tras el final del trimestre). En la Sección 8.1 se motiva el uso de este tipo de modelos y se detalla la rutina que se ha implementado para finalmente elegir tres modelos. Los modelos seleccionados se describen en la Sección 8.2, donde se muestra cuándo se emplea cada uno de ellos, el ajuste y la validación.

Al igual que se hizo con los modelos Box–Jenkins, se ha incorporado esta metodología en la aplicación creada, y en la Sección 8.3 se describe su aspecto, su utilidad y su funcionamiento.

8.1. Motivación y selección de los modelos

Normalmente, una de las principales motivaciones u objetivos a la hora de emplear modelos dinámicos es que las variables explicativas del modelo entren de manera retardada. De este modo, a la hora de predecir la variable respuesta no será necesario contar con un modelo a mayores para obtener proyecciones de la variable explicativa.

Sin embargo, la finalidad con la que se emplean los modelos de regresión dinámica en este trabajo es otra. Nos encontramos en una situación en la que la variable respuesta es una variable trimestral con un desfase de publicación de en torno a los 50 días; mientras que los indicadores son mensuales con un desfase de publicación que va desde los 5 hasta los 50 días. Con el foco puesto en predecir el dato del trimestre del PIB que está pendiente de publicar, lo que se busca con esta clase de modelos es adelantarse a los movimientos y comportamiento de la economía del trimestre en cuestión, mediante los indicadores que se van publicando esos meses. Por este motivo, lo deseable es que los indicadores entren de manera contemporánea. Lo que se persigue es que lo que está pasando en distintos sectores económicos (a través de los indicadores mensuales) proporcione información acerca de cuál será el valor del PIB. Esto tiene sentido ya que, al fin y al cabo, el PIB es una medida resumen de todos los indicadores.

Con el dato de 2017-Q3 publicado, el dato de 2017-Q4 se publicará a finales de febrero. La idea será emplear y aprovechar la información macroeconómica del cuarto trimestre de 2017 mediante los indicadores mensuales que se irán publicando desde octubre hasta febrero, pues estas variables darán cuenta de la situación de la economía durante los meses del trimestre en cuestión. Si ocurriera algo que provocara un desequilibrio en algún momento del trimestre, estos indicadores se harían eco y esto se podría tener en consideración a la hora de obtener una predicción del PIB.

Elección de variables explicativas

Como ya se ha comentado en repetidas ocasiones a lo largo de esta memoria, la serie del PIB con la que se trabaja está corregida de calendario y estacionalidad. Por este motivo lo esperado es que las variables que sean candidatas a explicarlo estén también corregidas de estos efectos; pues de no ser así se metería ruido en el modelo. De entre todos los indicadores que se muestran en el Cuadro 6.1, a la hora de buscar candidatos para ser variables explicativas del PIB se ha priorizado en función de los siguientes criterios:

- Que exista una correlación elevada con el PIB, no sólo entre las series en niveles sino también entre las series diferenciadas para evitar así correlaciones espurias.
- Que la publicación del dato no se produzca con mucho retraso.
- Que la serie sea lo suficientemente larga.

En base a lo expuesto anteriormente, se han preseleccionado como posibles indicadoras: las afiliaciones a la seguridad social el último día del mes (una vez que se ha corregido, véase Capítulo 6), las matriculaciones, las ventas de combustibles, las exportaciones, las importaciones, las pernoctaciones, el tráfico aéreo de pasajeros y el transporte marítimo de mercancías. Es decir, $\{Z_t^j\}$, con $j \in \{2, 4, 5, 8, 9, 11, 12, 13\}$.

Todos los indicadores son variables mensuales, mientras que el PIB es trimestral. El procedimiento más habitual es convertir los indicadores mensuales a trimestrales calculando el promedio de los 3 meses de cada trimestre. Es aquí donde se ha dado un paso más con el objetivo de poder aprovechar la información que se va publicando. Con cada indicador mensual $\{Z_t^j\}$ se han construido 5 indicadores trimestrales como sigue:

- $\{^{T_1}Z_t^j\}$: El indicador mensual se hace trimestral tomando como valor del trimestre el valor correspondiente al primer mes del trimestre.
- $\{^{T_2}Z_t^j\}$: El indicador mensual se hace trimestral tomando como valor del trimestre el valor correspondiente al segundo mes del trimestre.
- $\{^{T_3}Z_t^j\}$: El indicador mensual se hace trimestral tomando como valor del trimestre el valor correspondiente al último mes del trimestre.
- $\{^{T_{12}}Z_t^j\}$: El indicador mensual se hace trimestral tomando como valor del trimestre el valor correspondiente a la media de los valores que corresponden al primer y segundo mes del trimestre.
- $\{^{T_{123}}Z_t^j\}$: El indicador mensual se hace trimestral tomando como valor del trimestre el valor correspondiente a la media de los tres valores mensuales del trimestre.

En el Cuadro 8.1, para la serie $\{Z_t^4\}$ de matriculaciones se muestra la construcción de los 5 indicadores trimestrales a partir del indicador mensual.

Cod.Mens	2007-M1	2007-M2	2007-M3	2007-M4	2007-M5	2007-M6	Cod.Trim	2007-Q1	2007-Q2
Z_t^4	7447.790	6790.410	6949.015	6518.837	7022.939	7028.158			
							$^{T_1}Z_t^4$	7447.790	6518.837
							$^{T_2}Z_t^4$	6790.410	7022.939
							$^{T_3}Z_t^4$	6949.015	7028.158
							$^{T_{12}}Z_t^4$	7119.100	6770.888
							$^{T_{123}}Z_t^4$	7062.405	6856.645

Cuadro 8.1: Ejemplo de las transformaciones para pasar un indicador mensual a trimestral.

Por simplicidad, se usará para ambas series el subíndice t para referirse al tiempo ya que, con la nomenclatura empleada no es complicado distinguir cuando se trata de la serie en trimestres o meses. Tras una revisión de todas estas variables se ha decidido descartar las variables ${}^{T_2}Z_t^j$ y ${}^{T_3}Z_t^j$ ya que, si bien es cierto que dan información acerca de los segundos y terceros meses del trimestre; serán más adecuadas las variables ${}^{T_{12}}Z_t^j$ y ${}^{T_{123}}Z_t^j$ puesto que el PIB es una medida resumen del trimestre[1].

Es posible que los indicadores no sólo entren de manera contemporánea, sino que entren también de manera retardada. Se denotará a X_t a la variable en tiempo contemporánea y por X_{t-1} a la variable retardada un periodo temporal. Para el caso del PIB, esto se detalla en el Cuadro 8.2.

Fecha	Y_t	Y_{t-1}	Y_{t-4}
2006-Q4	99.34		
2007-Q1	100.42	99.34	
2007-Q2	101.65	100.42	
2007-Q3	102.63	101.65	
2007-Q4	103.39	102.63	99.34
2008-Q1	103.39	103.39	100.42

Cuadro 8.2: Variables retardadas, ejemplo para el caso del PIB.

Selección de los modelos

Es importante hacer hincapié en que la idea que ha primado ha sido la de ajustar varios modelos con el objetivo de ir aprovechando la información de la que se va disponiendo a medida que avanza el trimestre. Por ejemplo, es lógico pensar que un primer modelo debería incorporar las variables construidas con los primeros meses de cada trimestre, pues serán de las que primero se tengan datos, $\{{}^{T_1}Z_t^j\}$.

Tras distintos experimentos y por las características de los indicadores, se ha decidido trabajar tanto con la variable respuesta como con las variables explicativas con la transformación logarítmica (para hacer que estén todas en la misma escala) y con una diferencia regular de orden uno (con el fin de que los errores sean estacionarios y evitar correlaciones espurias, ya que todas las series presentaban tendencias). Se denotará por $\tilde{Y}_t$ la transformación anterior aplicada sobre la variable Y_t; es decir $\tilde{Y}_t = \log(Y_t) - \log(Y_{t-1})$, con t=2, . . . , n.

Una vez se han hecho trimestrales los indicadores, se han descartado algunos en base a sus fechas de publicación. Por ejemplo, para las exportaciones el dato del tercer mes del trimestre no estará disponible hasta que ya se haya publicado el PIB; por este motivo ya no se ha considerado la variables ${}^{T_{123}}\tilde{Z}_t^8$. Además, se ha tratado de evitar aquellas series de las que no se tiene fecha de publicación. Así, y con recomendaciones expertas recibidas, se llega a la siguiente lista de variables candidatas a ser explicativas (a efectos de simplificar la lectura se han obviado los corchetes de serie temporal):

Afiliaciones último día de mes: ${}^{T_1}\tilde{Z}_t^2$, ${}^{T_{12}}\tilde{Z}_t^2$, ${}^{T_{123}}\tilde{Z}_t^2$, ${}^{T_{123}}\tilde{Z}_{t-1}^2$, ${}^{T_{123}}\tilde{Z}_{t-2}^2$.

Matriculaciones: ${}^{T_1}\tilde{Z}_t^4$, ${}^{T_{12}}\tilde{Z}_t^4$, ${}^{T_{123}}\tilde{Z}_t^4$, ${}^{T_{123}}\tilde{Z}_{t-1}^4$, ${}^{T_{123}}\tilde{Z}_{t-2}^4$.

Consumo gasóleo: ${}^{T_1}\tilde{Z}_t^5$, ${}^{T_{123}}\tilde{Z}_{t-1}^5$, ${}^{T_{123}}\tilde{Z}_{t-2}^5$.

Exportaciones: ${}^{T_1}\tilde{Z}_t^8$, ${}^{T_{12}}\tilde{Z}_t^8$, ${}^{T_{123}}\tilde{Z}_{t-1}^8$, ${}^{T_{123}}\tilde{Z}_{t-2}^8$.

Importaciones: ${}^{T_1}\tilde{Z}_t^9$, ${}^{T_{12}}\tilde{Z}_t^9$, ${}^{T_{123}}\tilde{Z}_{t-1}^9$, ${}^{T_{123}}\tilde{Z}_{t-2}^9$.

[1] También es cierto que podrían entrar a la vez ${}^{T_1}Z_t^j$ y ${}^{T_2}Z_t^j$ y dar un resumen de los dos primeros meses , pero se ha tomado esta decisión para no tener un exceso de parámetros en el modelo.

Pernoctaciones: ${}^{T_1}\tilde{Z}_t^{11}$, ${}^{T_{12}}\tilde{Z}_t^{11}$, ${}^{T_{123}}\tilde{Z}_t^{11}$, ${}^{T_{123}}\tilde{Z}_{t-1}^{11}$, ${}^{T_{123}}\tilde{Z}_{t-2}^{11}$.

Tráfico aéreo: ${}^{T_1}\tilde{Z}_t^{12}$, ${}^{T_{123}}\tilde{Z}_{t-1}^{12}$, ${}^{T_{123}}\tilde{Z}_{t-2}^{12}$.

Transporte marítimo: ${}^{T_1}\tilde{Z}_t^{13}$, ${}^{T_{123}}\tilde{Z}_{t-1}^{13}$, ${}^{T_{123}}\tilde{Z}_{t-2}^{13}$.

PIB en instantes pasados: Y_{t-1} , Y_{t-2}, Y_{t-3} e Y_{t-4}.

Se ha programado una función para ajustar y poder elegir modelos en función de su BIC. Después, a la hora de seleccionar, se priorizará siempre por aquellos modelos que empleen variables con información más actualizada.

Rutina para ajustar los modelos

Seleccionado el conjunto de variables explicativas que se quiere que entren en el modelo (supongamos que ${}^{T12}\tilde{Z}_t^2$, ${}^{T1}\tilde{Z}_t^{13}$ y $\tilde{Y}_{t-2}$), el siguiente paso es ajustar un modelo de regresión con los errores convertidos en ruido blanco mediante un proceso $ARMA(p,q) \times (P,Q)$. Si los residuos de la regresión lineal no son estacionarios, se descarta el modelo; en caso contrario se inicia el siguiente proceso. Seleccionados los órdenes p_1, q_1, P_1 y Q_1, se ajusta el modelo de manera conjunta[2]:

$$Y_t = a + b^{T_{12}}\tilde{Z}_t^2 + c^{T_1}\tilde{Z}_t^{13} + d\tilde{Y}_{t-2} + \epsilon_t$$
$$\phi(B)_{p_1}\Phi_{P_1}(B^4)\epsilon_t = \theta(B)_{q_1}\Theta_{Q_1}(B^4)w_t.$$

Si el modelo no es válido o si las variables explicativas no son significativas, se descarta este modelo y se prueba con otros órdenes p_2, q_2, P_2, Q_2. Si el modelo es válido se calcula su BIC, se guarda y se ajusta el siguiente modelo para los residuos. Esto se repite hasta que se hayan probado todos los posibles modelos $ARMA(p,q) \times ARMA(P,Q)$. Es entonces cuando se selecciona aquel que tiene el mínimo BIC y ese es el modelo que se propone para las variables explicativas seleccionadas.

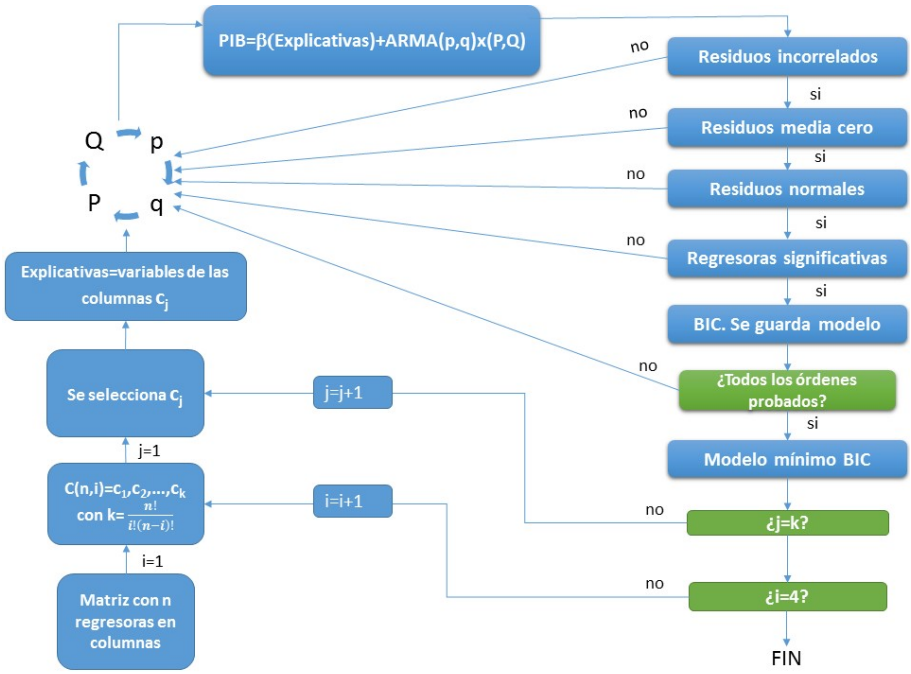

Figura 8.1: Rutina empleada para ajustar los modelos de regresión dinámica.

[2]En los operadores autorregresivos y de medias móviles el subíndice incluido indica el orden del operador.

La rutina anterior se lleva a cabo para una selección fijada de variables explicativas pero, ¿cómo elegir esta selección?. Para ello se ha partido de la lista con todas las variables candidatas a ser explicativas expuesta (un total de 36 variables) y se han ajustado tantos modelos de regresión como combinaciones se puedan hacer de 1, 2, 3 y 4 variables explicativas. Para cada combinación de variables se ejecuta la rutina anterior y de esta manera se tendrá, cuando sea posible, el mejor modelo en base al criterio BIC para cada combinación.

El esquema de funcionamiento y los distintos pasos se muestran en la Figura 8.1. El resultado es un total de 66711 combinaciones. En el Cuadro 8.3 se muestran los mejores y los peores modelos para las combinaciones de 3 variables explicativas.

Variables	BIC	p	d	q	P	D	Q
$\tilde{Y}_{t-2}$–$\tilde{Y}_{t-4}$–$^{T_{12}}\tilde{Z}_t^2$	-776.9785	0	0	0	0	0	1
$\tilde{Y}_{t-2}$–$\tilde{Y}_{t-4}$–$^{T_{123}}\tilde{Z}_t^2$	-776.5647	0	0	0	0	0	1
$\tilde{Y}_{t-1}$–$\tilde{Y}_{t-4}$–$^{T_{123}}\tilde{Z}_t^2$	-774.2285	1	0	0	0	0	1
$\tilde{Y}_{t-2}$–$\tilde{Y}_{t-4}$–$^{T_1}\tilde{Z}_t^2$	-774.0128	0	0	0	0	0	1
$\tilde{Y}_{t-1}$–$\tilde{Y}_{t-2}$–$^{T_{12}}\tilde{Z}_t^2$	-772.8388	0	0	2	0	0	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
$^{T_1}\tilde{Z}_t^4$_$^{T_{12}}\tilde{Z}_t^5$_$^{T_{123}}\tilde{Z}_t^5$	-273.0509	1	0	1	0	0	1
$^{T_1}\tilde{Z}_t^2$_$^{T_1}\tilde{Z}_t^9$_$^{T_{123}}\tilde{Z}_t^5$	-272.1728	1	0	0	0	0	0
$^{T_1}\tilde{Z}_t^4$_$^{T_1}\tilde{Z}_t^5$_$^{T_{123}}\tilde{Z}_t^5$	-271.2700	0	0	1	0	0	1
$^{T_1}\tilde{Z}_t^4$_$^{T_1}\tilde{Z}_t^5$_$^{T_{123}}\tilde{Z}_t^5$	-266.5972	0	0	1	1	0	1
$^{T_1}\tilde{Z}_t^9$_$^{T_1}\tilde{Z}_t^5$_$^{T_{123}}\tilde{Z}_t^5$	-265.9263	0	0	1	1	0	1

Cuadro 8.3: Mejores y peores modelos de combinaciones de 3 variables explicativas.

8.2. Modelos seleccionados

En base al criterio BIC, los modelos que mejor resultaban eran los que incluían las Afiliaciones a último día de mes y el PIB retardado, existiendo una gran diferencia en cuanto a BIC con modelos que incorporaban otros indicadores. Debido a que el desfase de publicación de esta variable es sólo de 14 días, parece muy apropiado quedarse con los 3 modelos que se detallan a continuación.

Modelo 1 (M1): Aprovecha la información de los primeros meses. Este modelo se empleará desde la fecha en la que se publique el dato de un trimestre del PIB hasta el día 14 del siguiente mes, que ya se publicará el dato de la variable explicativa para el segundo mes del trimestre. Incorpora también como explicativas el PIB retardado dos y cuatro instantes temporales.

$$\tilde{Y}_t = a_{M1} + b_{M1}\tilde{Y}_{t-2} + c_{M1}\tilde{Y}_{t-4} + d_{M1}{}^{T_1}\tilde{Z}_t^2 + \epsilon_t$$
$$\epsilon_t = \Phi_{M1} w_{t-4}.$$

Modelo 2 (M2): Este modelo incorpora la información de los dos primeros meses de cada trimestre, en media. Se emplea desde el día 14 del segundo mes del trimestre en cuestión hasta el día 14 del mes siguiente. Además entra el PIB retardado dos y cuatro instantes temporales.

$$\tilde{Y}_t = a_{M2} + b_{M2}\tilde{Y}_{t-2} + c_{M2}\tilde{Y}_{t-4} + d_{M2}{}^{T_{12}}\tilde{Z}_t^2 + \epsilon_t$$
$$\epsilon_t = \Phi_{M2} w_{t-4}.$$

Modelo 3 (M3): En este modelo la variable explicativa da información de lo que ha pasado a lo largo del trimestre ya que es la media de los tres meses. Se incluye también el PIB retardado dos y cuatro instantes temporales.

$$\tilde{Y}_t = a_{M3} + b_{M3}\tilde{Y}_{t-2} + c_{M3}\tilde{Y}_{t-4} + d_{M3}{}^{T_{123}}\tilde{Z}_t^2 + \epsilon_t$$
$$\epsilon_t = \Phi_{M3} w_{t-4}.$$

Por las particularidades de ambas series se ha optado por elegir un modelo sin fijar los coeficientes, estimándose cada vez que se ajuste el modelo.

Para comprender mejor la idea que hay detrás de estos tres modelos, en la Figura 8.2 se muestra las fechas de publicación del PIB y afiliaciones en función de los trimestres (por colores) y los modelos que se van aplicando (M1, M2 y M3). Se puede observar como con el uso de estos modelos se va aprovechando la información conforme se va publicando.

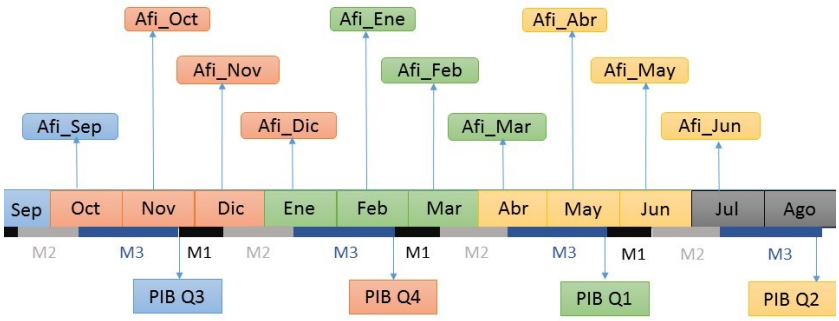

Figura 8.2: Esquema para comprender los modelos.

Predicción

Ajustado los modelos y comprobado que son válidos, el objetivo es obtener proyecciones del PIB a horizonte h. Estos modelos se han diseñado para que la primera predicción que se haga del PIB no requiera de proyecciones de las variables explicativas, pero sí para un horizonte mayor. Para la serie de afiliaciones, se emplean las proyecciones que proporciona el programa TSW con la corrección de calendario y estacionalidad, pues son predicciones que, tras varios ensayos, se ha comprobado que funcionan bastante bien.

De cara a predecir, se ha implementado una función que se describe para el modelo M1, pero que es operativa para cualquier modelo de regresión dinámica. Sean los parámetros ajustados $(\hat{a}_{M1}, \hat{b}_{M1}, \hat{c}_{M1}, \hat{d}_{M1}, \hat{\Phi}_{M1})$. Los pasos para obtener las predicciones se muestran a continuación:

Paso 1: Se ajusta la regresión lineal y se guardan los residuos:

$$\hat{\epsilon}_t = \tilde{y}_t - \hat{a}_{M1} - \hat{b}_{M1}\tilde{y}_{t-2} - \hat{c}_{M1}\tilde{y}_{t-4} - \hat{d}_{M1}{}^{T_1}\tilde{z}_t^2.$$

Paso 2: A los residuos se les ajuste un modelo $MA(1)_4$ con coeficiente $\hat{\Phi}_{M1}$.

$$\hat{\epsilon}_t = w_t + \hat{\Phi}_{M1}\hat{\epsilon}_{t-4}.$$

Paso 3: Se predicen h valores de los residuos con el modelo anterior (véase Capítulo 2):

$$(\hat{\epsilon}_t(1), \ldots, \hat{\epsilon}_t(h)).$$

Paso 4: Con el modelo de regresión del primer paso se predicen h valores iniciales de la respuesta, $(\tilde{y}_t(1), \ldots, \tilde{y}_t(h))$ como sigue (por comodidad se denota por $\tilde{y}(h)$ a la predicción en vez de emplear $\hat{\tilde{y}}(h)$):

$$\tilde{y}_t(h) = \hat{a}_{M1} + \hat{b}_{M1}\tilde{y}_{t-2}(h) + \hat{c}_{M1}\tilde{y}_{t-4}(h) + \hat{d}_{M1}{}^{T_1}\tilde{z}_t^2(h). \tag{8.1}$$

Paso 5: Las h predicciones finales son:

$$(\tilde{y}_t(1) + \hat{\epsilon}_t(1), \ldots, \tilde{y}_t(h) + \hat{\epsilon}_t(h))$$

Nótese que, aunque del lado derecho de la Ecuación (8.1) se haya empleado la notación de predicción a horizonte h, no todos los valores serán predicciones. Por ejemplo, el dato ${}^{T_1}z_t^2(1)$ siempre será real y para la variable Y_{t-2} se tiene que los valores $y_{t-2}(1)$ y $y_{t-2}(2)$ también serán reales. Además, al entrar como explicativas la variable respuesta retardada, será necesario ir alimentando las proyecciones de las explicativas con las proyecciones de la respuesta (en este caso cuando $h > 2$). Por último, conviene recordar que las variables se han transformado con un logaritmo y una diferencia regular, por lo que será necesario deshacer estas transformaciones de las proyecciones.

Todo esto se ha implementado en una función que resuelve todos estos inconvenientes y que devuelve ya las proyecciones en niveles. Además, para el ajuste del modelo se ha desarrollado otra rutina de manera que el usuario tiene que introducir la serie del PIB, la serie de afiliaciones reales y la de proyectadas. El programa selecciona automáticamente qué modelo emplear y construye también la serie de afiliaciones para proyectar (siendo el primer dato siempre real).

Ajuste y validación de estos modelos

En el momento en el que se hizo el ajuste el último dato disponible del PIB era 2017-Q3 y de las afiliaciones 2017–$M10$. En el Cuadro 8.4 se muestra la validación de los modelos a través de los contrastes estudiados en el Capítulo 2 (la función que detecta autocorrelación se estudió en el capítulo anterior). Se puede ver que todos son modelos válidos y el ajuste de estos modelos se muestra en el Cuadro 8.5, siendo todos los parámetros significativos a los niveles de significación usuales.

Contraste	M1	M2	M3
Media Cero	0.921	0.990	0.965
Jarque Bera	0.361	0.563	0.993
Shapiro	0.513	0.511	0.899
Test autocorrelación	1	1	1

Cuadro 8.4: Validación de los modelos de regresión (niveles críticos de los contrastes y salida de la función para detectar la autocorrelación).

$\hat{\Phi}_{M1}$	$\hat{a}_{M1}$	$\hat{b}_{M1}$	$\hat{c}_{M1}$	$\hat{d}_{M1}$	$\hat{\Phi}_{M2}$	$\hat{a}_{M2}$	$\hat{b}_{M2}$	$\hat{c}_{M2}$	$\hat{d}_{M2}$	$\hat{\Phi}_{M3}$	$\hat{a}_{M3}$	$\hat{b}_{M3}$	$\hat{c}_{M3}$	$\hat{d}_{M3}$
-0.875	0.002	0.490	-0.308	0.702	-0.834	0.002	0.454	-0.272	0.705	-0.778	0.002	0.427	-0.238	0.702
(0.079)	(0.000)	(0.090)	(0.061)	(0.060)	(0.081)	(0.000)	(0.089)	(0.059)	(0.058)	(0.085)	(0.000)	(0.091)	(0.059	(0.059)

Cuadro 8.5: Ajuste de los tres modelos tomando como último dato de las afiliaciones 2017–$M10$ y del PIB el 2017–Q3. Entre paréntesis están las desviaciones típicas.

Nótese que en esta situación el modelo adecuado sería M1, ya que con el dato de 2017–$M10$ se daría una proyección del trimestre en cuestión. En el Cuadro 8.6 se muestran los datos de las afiliaciones que son necesarios para obtener proyecciones del PIB a horizonte dos. Se puede observar como el primer valor necesario para las predicciones es real.

2017 Q3 (real)	2017-Q4 (real)	2018-Q1 (predicción)
973972.6	980487.0	984803.3

Cuadro 8.6: Valores empleados para obtener la proyección.

En el Cuadro 8.7 se muestran las predicciones obtenidas para el PIB por medio del modelo M1.

Fecha	Datos reales	M1
2016-Q1	100.680	-
2016-Q2	101.420	-
2016-Q3	102.290	–
2016-Q4	103.000	–
2017-Q1	103.660	–
2017-Q2	104.630	–
2017-Q3	105.420	–
2017-Q4	-	106.305
2018-Q1	-	107.1682

Cuadro 8.7: Últimos datos reales y predicciones del modelo M1.

8.3. Aplicación

Esta metodología se ha implementado en la aplicación creada. Como la variable explicativa necesaria tanto para ajustar los modelos como para obtener predicciones se obtiene en el programa TSW, será necesario cargar dos archivos en formato Excel (.xlxs) con estos datos. Por defecto, y en función del último dato de la variable explicativa, la rutina detecta qué modelo emplear, lo ajusta y muestra el mensaje de que modelo se ha ajustado. Destacar que, como se verá en el capítulo siguiente, uno de los beneficios de esta clase de metodología será que se va a poder medir la sensibilidad de los modelos a distintas situaciones, reflejadas en la serie de afiliados. Es por ello que el usuario puede modificar la

serie que se toma por defecto para proyectar y obtener las proyecciones que se obtienen a partir de esa nueva serie (a modo de análisis de sensibilidad a distintos escenarios). Las pestañas son:

Descripción: Se describe el funcionamiento de esta pestaña.

Modelos ajustados: Se describen los modelos ajustados.

Variable explicativa: El usuario tiene que, mediante un panel desplegable, seleccionar los archivos en los que están las series para el ajuste y la predicción. Se muestran estas series, tanto en una tabla como en un gráfico. Además, el usuario puede modificar los datos de la proyección de la variable explicativa y después proyectar con estos valores en el modelo final.

Modelo: Se ajusta el modelo pertinente; la salida es un mensaje con el modelo seleccionado, una tabla con el ajuste y la validación. Además se muestran las predicciones para el horizonte deseable, tanto para la serie cargada por el usuario como serie proyectada como para la que el usuario ha modificado.

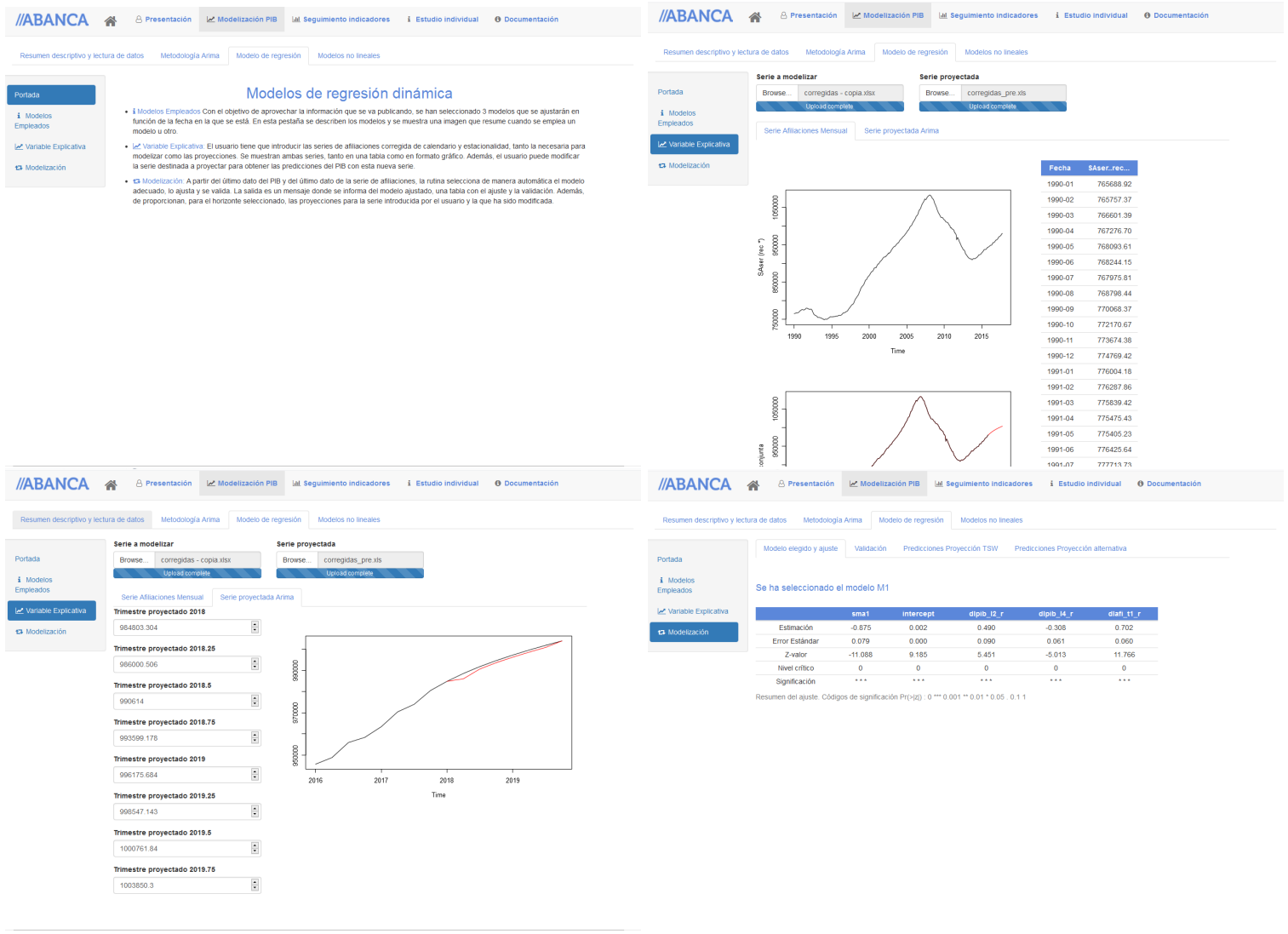

Figura 8.3: Aspecto de las pestañas de la aplicación en las que se ha implementado la metodología de modelos de regresión dinámica.

Capítulo 9

Comparativa de las distintas metodologías

Se ha trabajado con dos metodologías muy vinculadas y el objetivo de este capítulo es compararlas, así como destacar la capacidad de los modelos de regresión dinámica para adelantarse a los cambios económicos.

Por un lado, en la Sección 9.1 se desarrolla un ejercicio de "backtesting" para los modelos ajustados con el fin de comprobar su capacidad predictiva. Esta clase de ejercicios, en el contexto de las series temporales, consiste en recortar la serie, ajustar el modelo con la serie recortada, obtener predicciones y, por último, comparar dichas proyecciones con los valores reales de la serie que se dejaron fuera del estudio.

Cabe destacar que por tratarse la serie con la que se trabaja de una serie "especial", este ejercicio no es un backtesting habitual. Si se recorta la serie del PIB a fecha de 2016–$Q1$ como último valor, esta serie recortada no es la serie que se tenía en los meses de marzo o abril cuando se publicó el dato (por la corrección que se le hace). Sin embargo, puede servir para ver cómo funcionan las metodologías estudiadas a la hora de predecir.

Para el seguimiento macroeconómico de Galicia, se está interesado en las proyecciones del PIB desde dos puntos de vista:

- Interesa predecir el dato que está pendiente de publicar.
- Otro dato que también es importante es la media de las cuatro variaciones trimestrales interanuales para el año en cuestión. Este dato se denomina variación anual o cierre del año y, aunque a principios de año es una media de cuatro predicciones, a medida que avanza el año es una media de datos reales y de predicciones.

Por otro lado, en la Sección 9.2, a través de un ejemplo, se motiva el uso y la capacidad de anticiparse a shocks y a los cambios económicos de los modelos de regresión dinámica.

9.1. Capacidad predictiva

En esta sección se prueba la capacidad predictiva de los modelos Box-Jenkins y de los modelos de regresión dinámica. En los últimos, para un horizonte de predicción mayor que uno se necesitan proyecciones de la variable explicativa. A lo largo de este ejercicio se ha decidido emplear los valores reales de la misma. El motivo es que, si procediendo de esta manera se obtienen buenos resultados, significa que el modelo es bueno y todo el peso de obtener unas buenas predicciones recaerá en tener un buen modelo para la variable explicativa.

Para los modelos Box-Jenkins, una vez que se recorte la serie se elegirán los mejores modelos con la rutina detallada en el Capítulo 7 y se seleccionarán aquellos cuyo BIC diste menos de 2 unidades del

BIC del mejor modelo. Para los modelos de regresión dinámica, se ajustarán los modelos propuestos en el Capítulo 8.

Último dato real: 2015-Q4

La situación en la siguiente; nos encontramos en febrero de 2016 cuando se ha publicado el dato del PIB de 2015-Q4. En este momento, la serie de Afiliados tiene como último dato el de $2016 - M1$, por lo que el modelo a emplear sería el M1. El día 15 de marzo será el momento de ajustar el modelo M2 y el 15 de abril se ajustará el modelo M3.

Los ajustes de los modelos se han obviado y solo se muestran en el Cuadro 9.1 las proyecciones. Se observa como, en general, tanto las predicciones del dato de $2016 - Q1$ como la proyección para el cierre del año son bastante buenas.

		2016 Q1	2016 Q2	2016 Q3	2016 Q4	Cierre de año
	Real	100.68	101.42	102.29	103.00	3.08
Box-Jenkins	Modelo1	100.669	101.434	102.151	102.664	2.97
	Modelo2	100.692	101.480	102.263	102.859	3.06
	Modelo3	100.695	101.472	102.239	102.851	3.05
	Comb.mod	100.686	101.463	102.221	102.797	3.03
Regresión dinámica	M1	100.606	101.350	102.303	102.876	3.02
	M2	100.629	101.405	102.325	102.991	3.07
	M3	100.631	101.411	102.217	102.000	3.05

Cuadro 9.1: Proyecciones del backtesting cuando el último dato real es 2015-Q4.

Último dato real: 2016-Q1

Una vez que en mayo se publique el dato de $2016 - Q1$, se vuelven a ajustar los modelos y las predicciones obtenidas se muestran en el Cuadro 9.2. Se observa que, para esta ventana temporal, ambas metodologías proporcionan buenas proyecciones.

		2016 Q1	2016 Q2	2016 Q3	2016 Q4	Cierre de año
	Real	100.68	101.42	102.29	103.00	3.08
Box-Jenkins	Modelo1		101.454	102.185	102.713	2.99
	Modelo2		101.458	102.227	102.805	3.03
	Modelo3		101.445	102.195	102.789	3.01
	Comb.mod		101.452	102.204	102.772	3.01
Regresión dinámica	M1		101.428	102.421	102.998	3.12
	M2		101.458	102.405	103.073	3.14
	M3		101.462	102.291	103.078	3.11

Cuadro 9.2: Proyecciones del backtesting cuando el último dato real es 2016-Q1.

Último dato real 2015-Q2

Se propone ahora la situación en la que el último dato del PIB es el de $2015 - Q2$. En este caso, para dar el cierre del año 2015 son solamente necesarias dos proyecciones, pues el resto de datos son reales. En el Cuadro 9.3 se muestra una comparativa de las proyecciones con las distintas metodologías y se aprecia que, para este caso, los modelos de regresión dinámica generan mejores resultados.

		2015 Q1	2015 Q2	2015 Q3	2015Q4	Cierre de año
	Real	97.68	98.51	99.06	99.95	2.44
Box-Jenkins	Modelo1			99.010	99.659	2.35
	Modelo2			98.982	99.657	2.34
	Modelo3			99.047	99.747	2.38
	Comb.mod			99.010	99.682	2.36
Regresión dinámica	M1			99.098	100.064	2.48
	M2			99.082	99.977	2.45
	M3			99.025	99.992	2.44

Cuadro 9.3: Proyecciones del backtesting cuando el último dato real es 2015-Q2.

9.2. Capacidad de anticipar shocks

La decisión y motivación de ajustar modelos de regresión dinámica no era tanto el tener un modelo que diera mejores predicciones, ya que los modelos Box-Jenkins funcionan bien; sino tener la capacidad de adelantarse a cambios o shocks en el entorno macroeconómico. Con los modelos de regresión se pueden tener en cuentan estos cambios a través de los indicadores mensuales que se van publicando.

Situémonos en el año 2016, en concreto a finales de mayo cuando ya se ha publicado el dato del PIB del primer trimestre del año. A continuación se van a suponer dos situaciones en las que un shock provoca un cambio brusco en la economía:

Supuesto 1: se supone que un shock repentino en la economía en el mes de abril provoca que las afiliaciones empiecen a caer de manera brusca. Para generar este hipotético escenario adverso se han replicado las variaciones intermensuales de las afiliaciones de los años en los que comenzó la crisis, siendo estas -0.41, -0.45 y -0.66.

Supuesto 2: se supone una situación igual a la anterior, sólo que en sentido inverso. El shock provoca que afiliaciones comiencen a subir de manera brusca. Para replicar este escenario se han tomado las variaciones intermensuales de las afiliaciones en años de pleno crecimiento económico, siendo estas 0.47, 0.56 y 0.44.

Con estos escenarios para las afiliaciones fijados, se va a hacer una comparativa entre la proyección para el segundo trimestre de los modelos Box-Jenkins (nótese que, fijado el modelo, se obtiene siempre la misma predicción desde finales de mayo hasta agosto que se publica la cifra de $2017 - Q2$) con las 3 predicciones que se van obteniendo de los modelos M1, M2 y M3 a medida que trascurre el tiempo.

En el Cuadro 9.4 se muestran los distintos escenarios con los que se ha trabajado para las afiliaciones.

Fecha	Afiliaciones real	Afiliaciones adversas	Afiliaciones optimistas
2015-M3	948760.7		
2016-M4	945777.7	944798.6	953229.9
2016-M5	947232.1	940500.7	958639.9
2016-M6	948760.7	939878.5	962847.1

Cuadro 9.4: Escenarios que se plantean para las afiliaciones.

En los Cuadros 9.5 y 9.6 se encuentran las proyecciones que se obtienen para estos dos escenarios.

		2016-Q2	Variación interanual 2016-Q2
Box-Jenkins	Modelo 1	101.454	2.99
	Modelo 2	101.458	2.99
	Modelo 3	101.445	2.98
	Comb.mod	101.452	2.99
Regresión dinámica	M1	101.121	2.65
	M2	100.963	2.49
	M3	100.82	2.35

Cuadro 9.5: Capacidad de anticiparse a las situaciones adversas de la metodología de regresión dinámica.

		2016-Q2	Variación interanual 2016-Q2
Box-Jenkins	Modelo 1	101.454	2.99
	Modelo 2	101.458	2.99
	Modelo 3	101.445	2.98
	Comb.mod	101.452	2.99
Regresión dinámica	M1	101.753	3.29
	M2	101.962	3.50
	M3	102.051	3.54

Cuadro 9.6: Capacidad de anticiparse a las situaciones optimistas de la metodología de regresión dinámica.

En el Cuadro 9.5 se muestran las predicciones que se obtienen para el escenario adverso. Se observa que la metodología Box-Jenkins no da cuenta de este cambio económico. Sin embargo, en los modelos de regresión las afiliaciones se hacen eco del cambio que se ha producido en la situación económica y esto se traslada al modelo. Como consecuencia, con la metodología de modelos de regresión dinámica se obtienen cifras del PIB coherentes con el shock negativo.

Lo mismo ocurre, pero en sentido inverso, para el escenario optimista. Tal y como se puede ver en el Cuadro 9.6, con la metodología Box-Jenkins no se tiene en cuenta el shock, mientras que los modelos de regresión dinámica sí que son capaces de detectarlo y dar proyecciones coherentes con esta situación, obteniéndose valores más elevados del PIB.

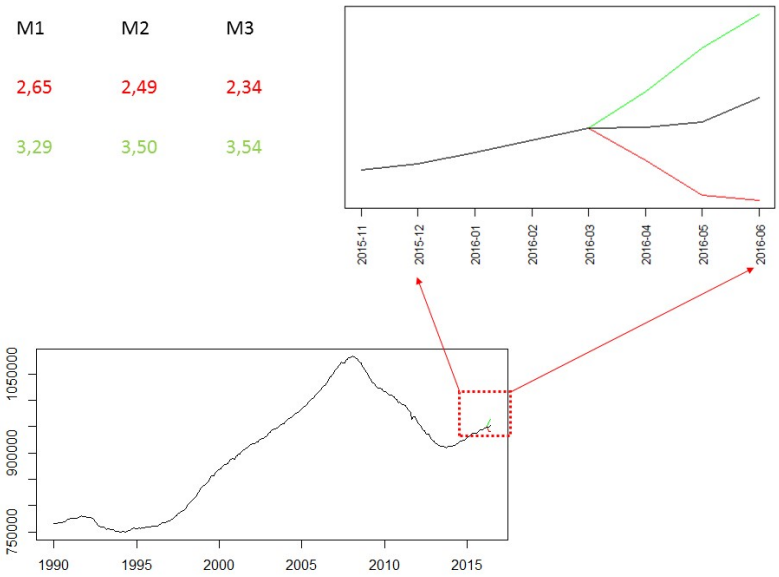

Figura 9.1: Escenario de la serie de afiliaciones y las proyecciones de los modelos de regresión dinámica. En negro se muestra la serie real. Los valores son las proyecciones del PIB en función de los distintos modelos de regresión dinámica.

En la Figura 9.6 se muestran en un formato gráfico los escenarios y las distintas proyecciones del PIB que se obtienen en función de los escenarios y los modelos. En color verde se muestran los datos para el supuesto dos (optimista), y en rojo para el supuesto uno (pesimista).

A través de este ejemplo queda claro el motivo por el que se ha decidido emplear la metodología de regresiones dinámicas y cómo es posible anticiparse a giros inesperados de la economía o a cambios de comportamiento.

Se puede concluir que si todo sigue el curso "normal" de la evolución económica (entendiendo por normal lo que ha estado ocurriendo en los instantes anteriores), entonces los modelos Box-Jenkins son una buena opción y no se requiere de un segundo modelo para proyectar la variable explicativa. Por otro lado, los modelos de regresión propuestos son una alternativa muy válida cuando se produzcan cambios o shocks en la situación económica.

Capítulo 10

Conclusiones y líneas futuras

Para cerrar esta memoria, en la Sección 10.1 se concluye con un breve resumen de lo que se ha hecho a lo largo de este trabajo. En la Sección 10.2 se describen las líneas de trabajo que han ido surgiendo y están en desarrollo o en vías de ser desarrolladas.

10.1. Conclusiones

El objetivo de este trabajo era encontrar un modelo para proporcionar proyecciones del PIB de Galicia. Para ello se han revisado dos metodologías muy vinculadas. Por un lado, se han puesto en práctica los modelos Box–Jenkins para la modelización de series temporales. Por otro lado, con el foco puesto en anticipar movimientos y cambios en la dinámica de la economía, se han ajustado modelos de regresión dinámica. Con ambas metodologías los resultados han sido bastante certeros y la acogida por parte de los expertos de la Entidad que van a hacer uso de estas proyecciones ha sido buena.

El transcurso de estas prácticas y el recorrido seguido para realizar este Trabajo Fin de Máster ha sido un proceso de continuo aprendizaje. Desde el aspecto económico, conocer qué es en realidad el PIB y el significado de muchos de los indicadores con los que se trabaja en el estudio de la situación macroeconómica, así como la importancia de la corrección de calendario y estacionalidad de las series económicas, tratamiento para mi desconocido. Desde el punto de vista estadístico, ha sido un reto enfrentarse al estudio de series temporales reales y se han revisados nuevas metodologías y herramientas para su estudio. Por último, el descubrimiento de la librería `Shiny` y el haber podido crear una herramienta a la que se le va a dar uso ha sido muy enriquecedor.

10.2. Líneas futuras

Durante la realización de este Trabajo Fin de Máster han surgido ciertas líneas de trabajo en las que el Departamento de Planificación Estratégica y PMO de ABANCA quiere poner el foco y seguir desarrollando:

- Ante el buen funcionamiento de los modelos de regresión dinámica a la hora de incorporar información de manera contemporánea para la proyección del PIB, se ha decidido probar este enfoque desde el punto de vista de la metodología no lineal. Se ha comenzado a ajustar modelos gam con errores corregidos por un proceso $ARMA(p,q) \times (P,Q)$ y, aunque no se contempla en esta memoria, sí está incorporada ya en la aplicación y en vías de desarrollo.
- El conocimiento de la API para descargar datos de manera automática del IGE ha dado pie a descubrir que el INE dispone de una herramienta similar. Del mismo modo, también se puede automatizar la descarga de los indicadores que publica el Banco de España. El objetivo es crear

una rutina para automatizar la descarga de la información que publican estos tres organismos públicos; información que se usa para el seguimiento macroeconómico (véase Capítulo 1).

- Para hacer el seguimiento macroeconómico se empleaban tanto las series corregidas como sin corregir. No obstante, algunas de las series que se estudian no se publican corregidas o se pueden consultar sin corregir en otro organismo antes de que las publiquen los Institutos. El haber conocido y entendido cuál es el procedimiento para su corrección, permite tratar aquellas series que no estén corregidas. Actualmente se está valorando automatizarlo de alguna manera.
- La librería Shiny ha sido un valioso descubrimiento y permite que los usuarios no conocedores del lenguaje de programación puedan acceder a las rutinas programadas. Está en desarrollo el implantar algunos de los procedimientos y estudios que se llevan a cabo en el Departamento por medio de otra aplicación.

Apéndice A

TSW

Se describen a continuación los parámetros principales de TRAMO-SEATS para llevar a cabo un ajuste. Para más detalle se recomienda Caporello y Maravall (2004).

Formato datos de entrada

Las series con las que se trabaja tienen que estar por columnas en un archivo `.xlsx` con el siguiente formato:

- Fila 1: El nombre de la serie.
- Fila 2: **NO AC PC F**, donde:
  - **NO:** Número de observaciones.
  - **AC:** Año de comienzo.
  - **PC:** Periodo en el que empieza.
  - **F:** Frecuencia.

Por ejemplo, para la serie de consumo de gasóleo del Capítulo 5 que comienza en 1993-M1 y finaliza en 2017-M7 el formato sería el siguiente: `295 1993 1 12`, siendo 295 el número de observaciones.

Procesamiento automático

El parámetro RSA permite un procesamiento totalmente automático. A continuación se muestran algunos de los valores y las funciones:

RSA=0: El parámetro no está activo.

RSA=4: Se emplea un modelo "Airline model" .

RSA=4: El programa hace una prueba para la especificación log/level, interpola observaciones faltantes y desarrolla una identificación automática del modelo. Además, se hace una prueba para determinar la presencia de efectos de calendario.

Órdenes del modelo Arima

P: Orden autoregresivo (p en este trabajo).

D: Orden diferencia regular (d en este trabajo).

Q: Orden medias móviles (q en este trabajo).

BP: Orden autoregresivo estacional (P en este trabajo).

BD: Orden diferencia estacional (D en este trabajo).

BQ: Orden medias móviles estacional (Q en este trabajo).

Transformación

LAM=0: Toma logaritmo.

LAM=1: No hay transformación.

LAM=-1: El programa hace una prueba y elige.

Corrección

IMEAN=0: No hay corrección de media.

IMEAN=1: Corrección de media.

Corrección de los efectos de calendario

IEAST=0: No hay efecto de Semana Santa.

IEAST=1: Se requiere la corrección del efecto de la Semana Santa.

IEAST=-1: El programa determina por defecto si es necesario la corrección del efecto de la Semana Santa.

IDUR=0: Es la duración de la Semana Santa (normalmente se elige 6).

ITRAD=0: No se corrige el efecto de Ciclo Semanal.

ITRAD=1 Se corrige el efecto de Ciclo Semanal con un parámetro (como en el Capítulo 4).

ITRAD=2 Se corrige el efecto de Ciclo Semanal con un parámetro y se corrige también el Efecto de Año bisiesto (como en el Capítulo 4).

Bibliografía

[1] Aneiros G (2016) Series de tiempo. Apuntes de la asignatura, Universidade da Coruña.

[2] Box GEP, Cox DR (1964) An Analysis of Transformations. Journal of the Royal Statistical Society. Series B (Methodological), vol. 26 (2), pp. 211-252.

[3] Box GEP, Jenkins GM (1970) Time Series Analysis, Forecasting, and Control. Holden-Day, San Francisco.

[4] Caporello G , Maravall A (2004) PROGRAM TSW. REVISED REFERENCE MANUAL. Banco de España.

[5] Carmona F, Subirana I (2015) Shiny: aplicaciones web interactivas con R. IV Jornadas de enseñanza y aprendizaje de la estadística y la investigación operativa.

[6] Chan KS, Ripley B (2015) TSA: Time Series Analysis https://cran.r-project.org/web/packages/TSA. R package version 1.01. Accedido 1 de septiembre de 2017.

[7] Chang W, Joe Cheng J, Allaire JJ, Xie Y, McPherson J (2017) shiny: Web Application Framework for R. R package version 1.0.5 https://cran.r-project.org/web/packages/shiny/index.html. Accedido 1 de septiembre de 2017.

[8] Cryer JD, Chan KS (2011) Time Series Analysis with Applications in R. Springer, New York.

[9] Eurostat (2009) ESS Guidlines on the seasonal adjustment, Methodologies and Working Papers.

[10] Faraway JJ (2014) Linear models with R. Taylor & Francis.

[11] Fox J , Weisberg S (2010) Time-Series Regression and Generalized Least Squares in R. An R Companion to Applied Regression, Second Edition.

[12] Gómez V (1998) Butterworth filters: A new perspective. Ministerio de Economía y Hacienda.

[13] Grolemund G (2015) How to start with Shiny.

[14] Instituto Nacional de Estadística (2002) Ajuste estacional y extracción de señales en la Contabilidad Trimestral Nacional Trimestral. Boletín Trimestral de Coyuntura, Num 84.

[15] Instituto Nacional de Estadística (2002) Extracción de señales y ajuste estacional en la CNTR. Estudio de un caso.Boletín Trimestral de Coyuntura, Num 85.

[16] Instituto Nacional de Estadística (2016) Contabilidad Nacional Trimestral de España. Metodología.

[17] Instituto Galego de Estatística (2017) Banco de series de conxuntura. Ciclotendencia e series corrixidas de estacionalidade e calendario. NOTAS EXPLICATIVAS.

[18] Jarque CM, Bera AK (1987) A test for normality of observations and regression residuals. International Statistical Review 55 (2), pp. 163-172.

[19] Hyndman R, O'Hara-Wild M, Bergmeir C, Razbash S, Wang E (2017) forecast: Forecasting Functions for Time Series and Linear Models. R package version 8.2 https://cran.r-project.org/web/packages/shiny. Accedido 1 de septiembre de 2017.

[20] Mochón F (2006) Principios de economía. McGraw-Hill, Madrid.

[21] Mora MA (2017) Introducción a la Inteligencia de negocios con ayuda de R. Proyecto Fin de Grado, Universidad de Sevilla.

[22] Neyman J (1952) Lectures and Conferences on Mathematical Statistics. Graduate School, US Department of Agriculture, pp. 143-154, Washington DC.

[23] de Pablo JC (2013) Por qué y cómo desestacionalizar series de datos.

[24] Peña D (2010) Análisis de series temporales. Alianza Editorial, Madrid.

[25] R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

[26] Shapiro SS, Wilk MB (1965) An Analysis of Variance Test for Normality (Complete Samples). Biometrika, Vol. 52 (34), pp. 591-611.

[27] Shumway RH, Stoffer DS (2011) Time Series Analysis and Its Applications: With R Examples. Springer, New York.

[28] Tsay RS (2005) Analysis of Financial Time Series. John Wiley and Sons, Hoboken, New Jersey.

[29] Wold HO (1938) A Study in the Analysis of Stationary Time Series. Almquist and Wicksell, Uppsala.

[30] Wei WW (2006) Time series analysis: univariate and multivariate methods. Pearson Addison Wesley, Boston.