Trabajo Fin de Máster

Modelización de la serie temporal de la inversión privada en bienes y servicios en Galicia sobre la base de variables internas de ABANCA

Alejandro Figueroa Silva

Máster en Técnicas Estadísticas

Curso 2018-2019

Propuesta de Trabajo Fin de Máster

Título en galego: Modelización da serie temporal da inversión privada en bens e servizos en Galicia sobre a base de variables internas de ABANCA

Título en español: Modelización de la serie temporal de la inversión privada en bienes y servicios en Galicia sobre la base de variables internas de ABANCA

English title: Time series modelling of private investment in Galicia bases on internal variables of ABANCA

Modalidad: Modalidad B

Autor: Alejandro Figueroa Silva, Universidad de A Coruña

Director: José Antonio Vilar Fernández, Universidad de A Coruña

Tutora: Belén María Fernández de Castro, ABANCA

Breve resumen del trabajo:

El principal objetivo es desarrollar un modelo estadístico adecuado para la serie temporal de inversión privada en bienes y servicios en Galicia basándose en la evolución de variables internas de ABANCA. La entidad dispone ya de un modelo predictivo del consumo de bienes y servicios por parte de los hogares (desarrollado en el marco de un TFM anterior), de modo que la utilización conjunta de ambos modelos permitiría a la entidad obtener un pulso de la evolución de la economía de Galicia con antelación a la publicación del PIB, toda vez que ambas componentes determinan en gran medida la evolución de este indicador macroeconómico.

En principio, se pretende aplicar una metodología que consistiría de cuatro fases: (i) selección, análisis y tratamiento de variables internas de la entidad que potencialmente podrían ser de interés para incluir en el modelo, (ii) aplicación de técnicas de reducción de la dimensión de este conjunto de variables explicativas teniendo en cuenta el carácter dinámico de las mismas, (iii) modelización de las series seleccionadas para el modelo, (iv) establecimiento y evaluación de un modelo de regresión apropiado para explicar la relación entre la serie de inversión privada en bienes y servicios y las series seleccionadas como explicativas.

Don José Antonio Vilar Fernández, Catedrático de la Universidad de A Coruña, doña Belén María Fernández de Castro, Coordinadora de ABANCA, informan que el Trabajo Fin de Máster titulado

Modelización de la serie temporal de la inversión privada en bienes y servicios en Galicia sobre la base de variables internas de ABANCA

fue realizado bajo su dirección por don Alejandro Figueroa Silva 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 5 de Septiembre de 2019.

El director:

Don José Antonio Vilar Fernández

La tutora:

Doña Belén María Fernández de Castro

El autor:

Don Alejandro Figueroa Silva

Agradecimientos

Este Trabajo de Fin de Máster no sería posible sin las recomendaciones y pautas sugeridas por José Antonio Vilar Fernández por parte de la Universidad de A Coruña, como por los consejos y asesoramiento de los compañeros de ABANCA, con especial mención a Belén Fernández de Castro y a Teresa Veiga Rodríguez, ya que con sus consejos y continua dedicación han enriquecido de forma notable el contenido de este trabajo.

Índice general

Resumen

Resumen

Desde el área de Planificación Estratégica y PMO de la entidad *ABANCA Corporación Bancaria S.A.* se realizan múltiples tareas orientadas al desarrollo de los planes estratégicos, siendo de especial interés el seguimiento de la evolución de las principales macromagnitudes e indicadores coyunturales de la economía gallega, ya que es en esta comunidad autónoma donde reside la mayor parte de su negocio.

En el presente trabajo se llevará a cabo la construcción de un indicador para la formación bruta de capital (`FBC`), variable que recoge el dinamismo inversor de la economía gallega y que representa un 33 % del producto interior bruto de la comunidad autónoma. Dado que la publicación de esta variable, de frecuencia trimestral, tiene un desfase de 53 días, resulta de importancia llevar a cabo la construcción de un modelo que nos permita conocer la evolución de la misma a través de información interna proporcionada por las variables de negocio de la entidad.

Para ello se utilizarán técnicas de análisis de series temporales, en concreto técnicas Box-Jenkins, así como diferentes modelos de regresión que nos permitan obtener un reflejo fiel del comportamiento de la inversión gallega a cierre de mes.

Abstract

In the Strategic Planning and PMO area of the entity *ABANCA Corporación Bancaria S.A.*, some multiple assignements are performed, oriented to the development of strategic plans, being of special interest the monitoring of the evolution of the main macromagnitudes and short-term indicators of the galician economy, due to that in this autonomous community most of its business resides.

In the current work will be carried out the construction of an indicator for Gross Capital Formation (`GFC`), a variable that resumes the investment dynamism of the galician economy and that represents a 33 % of the gross domestic product of the Galicia. Given that the publication of this quarterly frequency variable has a delay of 53 days, it is important to carry out the construction of a model that allows us to know its evolution through internal information provided by the entity's business variables.

In order to do so, we will use time series analysis techniques, specifically Box-Jenkins techniques, as well various regression models that allow us to obtain a faithful reflection of the behaviour of galician investment at the end of the month.

Parte I

Metodología

Capítulo 1

Motivación e introducción al problema

1.1. Motivación

En el área de Planificación Estratégica y PMO de la entidad *ABANCA Corporación Bancaria S.A.* (ABANCA de ahora en adelante) se realizan, entre muchas otras tareas, el seguimiento de la evolución macroeconómica, el desarrollo de los Planes Estratégicos, la presupuestación anual o los procedimientos necesarios para cumplir distintos requerimientos del supervisor (Stress Test, ICAAP,...). En concreto, el seguimiento de la evolución macroeconómica se realiza para distintas economías, siendo de especial interés la evolución de la economía gallega, ya que es en ésta donde reside la mayor parte del negocio de la entidad. Para llevar a cabo esta tarea, se realizan seguimientos de las principales macromagnitudes publicadas por distintos organismos oficiales e institutos de estadística (Banco de España, Instituto Nacional de Estadística, Instituto Galego de Estatística, ...), así como indicadores de coyuntura que nos ofrezcan una imagen fiel de como se está comportando la actividad económica.

Gran parte de las variables macroeconómicas e indicadores coyunturales de los que se hace seguimiento son publicados con bastante decalaje, provocando retrasos en los análisis realizados en base a los mismos. Además, algunos tienen una frecuencia de publicación trimestral, lo que impide incluir el análisis de su comportamiento en el seguimiento mensual que se hace dentro de la Entidad. El producto Interior bruto, así como sus principales componentes, son uno de estos indicadores, que se publican trimestralmente y además con un desfase de 53 días desde el final del trimestre. Es decir, para conocer cual ha sido la evolución de la economía gallega en el primer trimestre del año tendríamos que esperar teóricamente hasta marzo, pero la publicación de dicho indicador no se realiza hasta el 23 de mayo[1]. Este desfase en la fecha de publicación motiva la necesidad de construir un indicador que permita anticipar y conocer el estado de la economía gallega con una frecuencia mayor.

La modelización de las distintas componentes del `PIB` se realizará a través de las variables internas de negocio de la entidad. En concreto, en este trabajo se intentará modelizar el dinamismo inversor de la economía gallega resumido en la componente de la formación bruta de capital (`FBC` de ahora en adelante). Esta modelización le otorga a la entidad una ventaja competitiva muy importante, ya que a cierre de cada mes dispondría de un indicador que recoge el comportamiento de la `FBC` de la economía gallega, a partir del cual también se podrían realizar predicciones para anticipar el comportamiento de la inversión en los diferentes trimestres del año. Destacar que en trabajos anteriores se ha realizado la modelización de la componente del consumo del `PIB`, de esta forma la ABANCA dispondría de dos indicadores creados a partir de variables internas de negocio de la entidad que modelizan la mayor parte del `PIB` de la comunidad gallega.

Para llevar a cabo esta modelización se abarcarán múltiples técnicas estadísticas, desde el tratamiento de series temporales mediante la metodología Box-Jenkins, como el uso de distintos modelos de regre-

[1] Ver fechas de publicación: `https://www.ige.eu/web/Controlador?operacion=calendario&idioma=gl&mes=5&ano=2019`.

sión con el fin de encontrar la especificación adecuada que nos permita obtener un indicador fiel del comportamiento de la inversión en la economía gallega.

1.2. Producto Interior Bruto: Una breve introducción

Para comenzar, explicaremos brevemente en qué consiste el `PIB`, qué información resume, quiénes son las entidades encargadas de su publicación y cuáles son sus principales componentes.

El `PIB` es una de las principales macromagnitudes de síntesis, de carácter coyuntural, cuyo objetivo es proporcionar una descripción cuantitativa y coherente de la evolución reciente de la economía. Las estimaciones de la misma se ajustan a los principios de coherencia y equilibrio contable entablados en el marco del Sistema Europeo de Cuentas 2010 (SEC-2010), normativa que rige a los organismos encargados de la publicación de dicha macromagnitud (para más información ver Eurostat (2014)). En el ámbito nacional, es el Instituto Nacional de Estadística (INE)[2] el organismo encargado de su publicación, y en el caso de la comunidad autónoma de Galicia es el Instituto Galego de Estatística (IGE).[3]

De acuerdo con su definición en la contabilidad nacional, el `PIB` se define como el valor de todos los bienes y servicios **finales** (se excluyen aquellos bienes de carácter intermedio con el fin de evitar una doble contabilización) producidos en un **territorio** durante un período de tiempo determinado, generalmente un año. Cabe destacar, que el `PIB` no mide la riqueza o patrimonio de un país, sino su capacidad productiva, es decir, refleja la capacidad de una economía para producir riqueza a lo largo de un período.

La estimación del mismo puede realizarse mediante 3 vías o enfoques:

- **Enfoque de la demanda agregada o gasto:**

  Se tienen en cuenta la suma de los destinos finales de todos los bienes y servicios producidos por los factores que operan en el seno del territorio. Los componentes del `PIB` bajo el enfoque de la demanda agregada son: el consumo final efectivo (CF), la formación bruta de capital (`FBC`) y el sector exterior, en el que se tienen en cuenta las importaciones (M) y exportaciones (X) realizadas. De modo que, la expresión de su cálculo es la siguiente:

  Definición 1.2.1. *Bajo el enfoque de la demanda agregada el `PIB` se puede expresar:*

  $$PIB = CF + FBC + X - M$$

- **Enfoque de la producción u oferta:**

  Ofrece un procedimiento alternativo para estimar el `PIB`, a partir de la suma de los valores añadidos de los diferentes sectores institucionales o de las distintas ramas de actividad. En este caso, el `PIB` se corresponderá con la suma del valor añadido bruto a precios básicos de cada rama de actividad (VAB), sumando los impuestos netos (I), y restando subvenciones (S) sobre los productos.

  Definición 1.2.2. *Bajo el enfoque de la producción el `PIB` se puede expresar:*

  $$PIB = \text{VAB}(\text{Sector Primario}) + \text{VAB}(\text{Sector Secundario}) +$$
  $$\text{VAB}(\text{Sector Terciario}) + \text{I} - \text{S}$$

[2] `http://www.ine.es/welcome.shtml`
[3] `http://www.ige.eu/web/index.jsp?paxina=001&idioma=gl`

- **Enfoque de la renta:**

  El cálculo mediante esta vía se obtiene agregando el pago por los servicios de todos los factores productivos integrados en la producción. Es decir, se tendrá en cuenta la remuneración de asalariados (RA), el excedente bruto de explotación (EBE) y, los impuestos (I) y subvenciones (S) sobre la importación y la producción.

  Definición 1.2.3. *Bajo el enfoque de la renta el* `PIB` *se puede expresar:*

$$PIB = RA + EBE + I - S$$

De los tres enfoques anteriormente vistos, en el presente trabajo se estudiará el comportamiento de los agregados macroeconómicos bajo el enfoque de la demanda agregada o gasto (ver Definición 1.2.1). En concreto, nos centraremos en intentar aproximar el comportamiento de la formación bruta de capital de la economía gallega mediante una batería de variables internas de negocio de ABANCA, que presentan una frecuencia mensual. Como se ha mencionado en la introducción de este trabajo, la frecuencia de publicación de la `FBC` (trimestral) ralentiza la adopción de medidas y estrategias a corto plazo de la entidad, siendo de gran utilidad la construcción de un modelo que nos permita obtener una estimación mensual de la misma y que además permita obtener predicciones.

A continuación, se explicará brevemente en qué consiste la `FBC`, cuáles son sus componentes, y por último, cada una de las variables internas implicadas en el proceso de estimación de la `FBC`.

1.2.1. Formación bruta de capital

Toda actividad productiva requiere de una inversión previa para poder llevarse a cabo, es decir, es necesario destinar recursos de capital a la producción de nuevos bienes y servicios. La contabilidad nacional mide esa actividad de inversión mediante la formación bruta de capital. La `FBC` tiene un papel relevante en el ámbito económico por diversos motivos, entre los que destaca, su importancia sobre la producción futura, y su relación con las expectativas de los individuos sobre el comportamiento de la economía. De modo que, la inversión, entendida como la adquisición de maquinaria y edificios, más la variación de existencias, o variación de inventarios, se denomina formación bruta de capital (para más información consultar Lequiller, François and Blades, Derek (2018)).

La formación bruta de capital está compuesta por los siguientes componentes:

- **La formación bruta de capital fijo**:
  En las cuentas nacionales se define la formación bruta de capital fijo como la adquisición de activos cuyo destino es ser utilizados, durante un período superior a un año, en el proceso de producción de otros bienes y servicios. Es decir, comprende las adquisiciones (menos las cesiones) de activos fijos que llevan a cabo los productores residentes durante un período de tiempo determinado. Conviene aclarar que incluye aquella inversión destinada a cubrir la depreciación del stock de capital, de ahí el adjetivo de **bruta**, y el calificativo de **fijo** se debe a los activos materiales o inmateriales obtenidos a partir de procesos productivos que se utilizan, de forma repetida, en otros procesos de producción durante más de una año.

  El nuevo sistema SEC-2010 ha supuesto un cambio relevante en los componentes que conforman la FBCF, incluyendo el gasto en I+D así como el gasto en sistemas de armamento (ver Lequiller, François and Blades, Derek (2018, pág: 160)). Destacar que la composición de la FBCF, dependiendo del tipo de activo en cuestión, en el nuevo sistema SEC-2010 es el siguiente:

Definición 1.2.4. *Componentes de la FBCF en el sistema SEC-2010:*

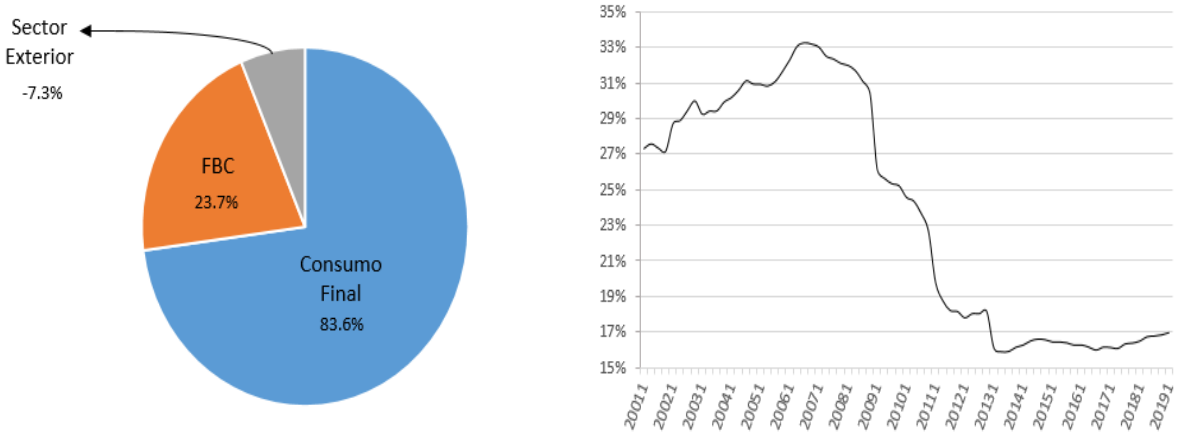
$$FBCF = Viviendas \ + \ Otros\ edificios\ y\ construcciones + \ Maquinariay\ bienes\ de\ equipo$$
$$+ \ Recursos\ biologicos\ cultivados \ + \ Propiedad\ intelectual$$

- **La variación de existencias**:
  Se mide como la diferencia entre el valor de las entradas y salidas de existencias a lo largo de un período, una vez se han descontado las pérdidas corrientes de los bienes mantenidos en existencias.[4]

En resumen, las componentes de la FBC se recogen en el siguiente esquema:

Formación Bruta Capital
- **Formación bruta capital fijo**
  - Viviendas
  - Otros edificios y construcciones
  - Maquinaria y bienes de equipo
  - Recursos biológicos cultivados
  - Propiedad intelectual
- **Variación existencias y objetos valiosos**

La FBC es una componente del PIB bajo el enfoque de la demanda agregada que representa actualmente entorno al 17 % del PIB, de ahí la necesidad de poder obtener una estimación de la misma que permita preveer cómo se está comportando la inversión en Galicia. A continuación, a través de los datos del IGE, se muestra el peso relativo histórico de la FBC en el agregado del PIB para el conjunto de la economía gallega en los últimos años, y el resto de componentes de la vía de la demanda (ver Definición 1.2.1).



(a) Composición histórica del PIB (período 2000-2019).

(b) Peso relativo de la FBC.

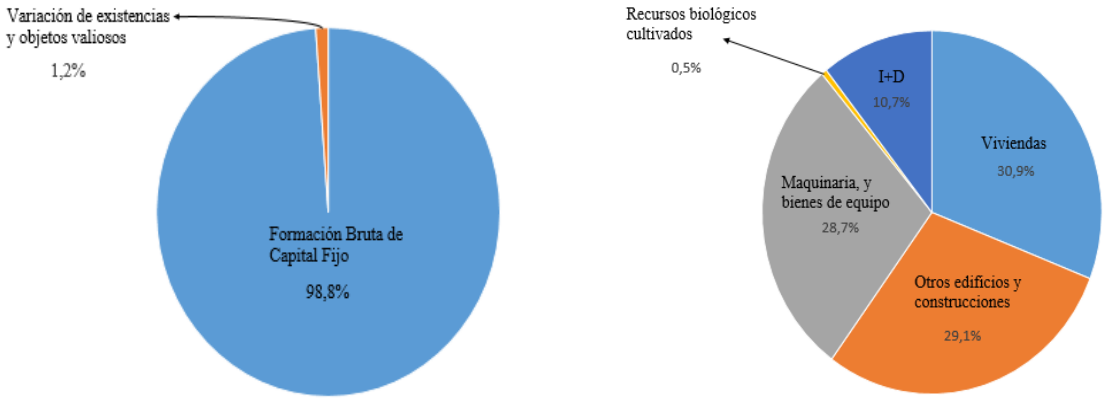
Figura 1.1: Evolución histórica de la FBC en Galicia (período 2000-2019).

En definitiva, como se aprecia en Figura 1.1, la FBC representa una parte relevante del PIB de Galicia alcanzando tasas del 33 % en los períodos de máxima expansión de la economía (año 2007). Siendo un buen indicador del grado de inversión en una economía, en este caso la gallega, y por ende, es necesario una monitorización del mismo. Sin embargo, como hemos podido ver la FBC se subdivide en dos grandes agregados, la formación bruta de capital fijo (FBCF) y la variación de existencias (VE). Debemos aclarar que en este trabajo el comportamiento de la formación bruta de capital (FBC) vendrá determinada

[4]Dentro de esta categoría se engloba también la cuenta de objetos valiosos debido a su valor residual en el agregado total.

completamente por la formación bruta de capital fijo (FBCF), ya que debido a la naturaleza de los datos no tenemos modo alguno de determinar cuál sería el comportamiento de la variación de existencias. En todo caso, el porcentaje que representa la variación de existencias y objetos valiosos es residual, un 1.2 % del total de la `FBC`, como podemos ver en la serie histórica de la `FBC` para España[5].

A continuación, se puede ver la composición histórica tanto de la `FBC` como de la FBCF para la economía española entre los años 2000 y 2019.



(a) Componentes de la `FBC`.

(b) Componentes de la FBCF en el `PIB`.

Figura 1.2: Composición histórica de la FBC y FBCF para España (período 2000-2019).

1.3. Extracción de las variables

Tal y como se ha comentado, se pretende construir un modelo que explique la FBC a partir de información interna de ABANCA. Por tanto, será necesario construir métricas a partir de la información de negocio de la Entidad y, para ello, es necesario acceder a los sistemas informacionales. ABANCA dispone de un Data Warehouse (DWH) que se alimenta diariamente de la información que se genera en el operacional. El DWH está diseñado como una base de datos relacional enriquecido, además, con una amplia colección de vistas que facilitan el acceso y la compresión de la información almacenada. La extracción de información de la base de datos se realiza mediante lenguaje SQL. A continuación, se explican brevemente algunas de las instrucciones de SQL más habituales con algunos ejemplos que ilustran su funcionamiento.

Las principales "instrucciones" de *SQL* son:

- **SELECT** : Nos permite seleccionar las columnas (variables) necesarias de la consulta realizada.
- **FROM**: Cláusula en la que debemos indicar cuál es la tabla o subconsulta donde se encuentran las columnas declaradas en el SELECT.
- **WHERE**: Nos permite filtrar los resultados de una consulta mediante operadores lógicos, siendo los más comunes los operadores **AND** y **OR**.
- **GROUP BY**: Agrupa las variables seleccionadas por los "niveles" de la variable o variables de agrupación escogidas. Existe otro conjunto de "instrucciones" que suelen ser utilizadas con **GROUP BY** permitiéndonos realizar las siguientes operaciones:
  - **COUNT**: Nos devuelve el número de filas de una o varias columnas.

[5]Se utilizan los datos del INE y no del IGE, ya que para la economía gallega no está disponible la desagregación de la `FBC` por componentes.

  - **AVG**: Calcula la media de la columna seleccionada.
  - **MAX /MIN** : Calcula el máximo o el mínimo de la columna seleccionada.
  - **SUM**: Calcula la suma de la columna seleccionada.

- **ORDER BY**: Nos permite ordenar los resultados de la consulta en función de la columna seleccionada, dispone de argumentos para indicar el orden ascendente o descendente.
- **BETWEEN**: Filtra los valores de una columna (variable) de la consulta entre el rango de los valores especificados.

Las "instrucciones" mostradas anteriormente son utilizadas para hacer múltiples consultas sobre una tabla, pero puede darse el caso que queramos consultar datos de una tabla en función a ciertas columnas de otra tabla, consultas conocidas como "multi-tabla". Estas consultas "multi-tabla" se realizan bajo los distintos tipos de **JOIN** que veremos a continuación:

- **INNER JOIN**: Devuelve únicamente aquellas filas que tienen valores idénticos en los dos campos que se comparan para unir ambas tablas. Su sintaxis se correspondería con:

  FROM Tabla A **INNER JOIN** Tabla B **ON** Condiciones / Vínculos entre tablas

  Una forma útil y didáctica de ver el funcionamiento del INNER JOIN es a través del diagrama de Venn que se muestra en la Figura 1.3a.

- **LEFT JOIN**: Mantiene todas las filas de la tabla A mostrándose únicamente las filas de la tabla B que sean coincidentes. Es decir, el resultado de esta operación siempre contiene todos los registros de la tabla A, aun cuando no exista un registro correspondiente en la tabla B. Su sintaxis se correspondería con:

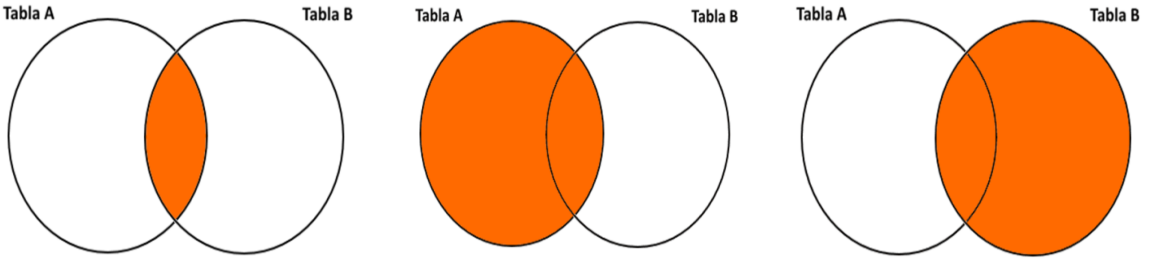
  FROM Tabla A **LEFT** Tabla B **ON** Condiciones / Vínculos entre tablas

  Siendo el diagrama de Venn el mostrado en la Figura 1.3b.

- **RIGHT JOIN**: Mantiene todas las filas de la tabla B mostrándose únicamente las filas de la tabla A que sean coincidentes. Es decir, el resultado de esta operación siempre contiene todos los registros de la tabla B, aun cuando no exista un registro correspondiente en la tabla A. Su sintaxis se correspondería con:

  FROM Tabla A **RIGHT** Tabla B **ON** Condiciones / Vínculos entre tablas

  Siendo el diagrama de Venn el correspondiente a la Figura 1.3c



(a) INNER JOIN. (b) LEFT JOIN. (c) RIGHT JOIN.

Figura 1.3: Diagrama de Venn para las distintas clases de INNER.

Por último, y para finalizar esta sección, ilustraremos con varios ejemplos de cómo utilizar algunas de las "instrucciones" explicadas.

Imagínese una empresa que dispone de un sistema de bases relacionales para explotar la información generada por el funcionamiento de su actividad económica, y que dispone de las siguientes tablas dentro de la misma:

- Tabla de clientes ($CLIENTES$), en la que se recogen los siguientes datos: [6]
  - CLI_ID: **ClienteID**.
  - NOM: Nombre del Cliente.
  - APE: Apellidos del Cliente.
  - TLF: Teléfono de contacto.
  - $CCAA$: Comunidad Autónoma.
  - $PROV$: Provincia.

CLI_ID	NOM	APE	TLF	$CCAA$	$PROV$
1	Paula	Fernández Álvarez	678904656	Galicia	A Coruña
2	Fernando	González Pita	658363749	Galicia	Pontevedra
3	Cristina	Lameiro Rodríguez	678113274	Madrid	Madrid
4	Jorge	Rodríguez Rodríguez	690284672	Cataluña	Barcelona
5	Alejandro	De la fuente Álvarez	623512403	Galicia	Lugo
6	Belén	Alonso Maquieria	646820371	Aragón	Huesca
7	Pedro	Crespo Rodríguez	639871917	Cataluña	Girona
8	Manuel	Fraga Luaces	654291201	Galicia	A Coruña
9	Cristina	Patiño Lago	679899809	Galicia	A Coruña
10	María	Naveira Ramos	645372895	Cataluña	Barcelona

Tabla 1.1: Tabla Clientes.

- Tabla de pedidos ($PEDIDOS$), en ella se recoge la información correspondiente a los pedidos efectuados:
  - PED_ID: **PedidoID**.
  - CLI_ID: ClienteID.
  - $FACT$: Factura.
  - IMP: Importe.
  - DIA: Día.

[6] Nótese que el campo destacado en negrita se corresponde con la clave primaria de la tabla.

PED_ID	*CLI_ID*	*FACT*	*IMP*	*DIA*
234	10	30	140	17/02/2018
456	4	31	360	17/02/2018
236	3	90	1000	23/05/2012
478	8	40	800	16/05/2014
293	5	23	300	15/04/2017
398	6	1	50	14/09/2019
142	1	10	900	26/10/2018
109	7	11	56	15/06/2018
180	2	20	40	07/12/2017
345	9	45	800	18/01/2017

Tabla 1.2: Tabla Pedidos.

Si queremos saber qué facturas tienen un importe menor o igual a 300 €, y en qué día se han realizado, la sentencia en SQL debería ser la siguiente:

```
SELECT FACT, IMP, DIA
FROM PEDIDOS
WHERE IMP <= 300
```

El resultado de dicha consulta se muestra en la Tabla 1.3:

FACT	*IMP*	*DIA*
30	140	17/02/2018
23	300	15/04/2017
1	50	14/09/2019
11	56	15/06/2018
20	40	07/12/2017

Tabla 1.3: Consulta Tabla Pedidos.

Ahora, si lo que queremos saber es a qué comunidad autónoma corresponden esos importes, la consulta de SQL debería ser:

```
SELECT CLIENTES.CCAA, PEDIDOS.FACT, PEDIDOS.IMP, PEDIDOS.DIA
FROM PEDIDOS
INNER JOIN CLIENTES
ON CLIENTES.CLI_ID = PEDIDOS.CLI_ID
WHERE PEDIDOS.IMP <= 300
```

siendo el resultado de la consulta el mostrado en la Tabla 1.4:

CCAA	*FACT*	*IMP*	*DIA*
Galicia	30	140	17/02/2018
Galicia	23	300	15/04/2017
Aragón	1	50	14/09/2019
Cataluña	11	56	15/06/2018
Cataluña	20	40	07/12/2017

Tabla 1.4: Consulta Tabla pedidos y Clientes.

Realizando varias consultas mediante SQL y utilizando *IBM SPSS Modeler* para fusionar, seleccionar, filtrar y detectar anomalías en las series, se han obtenido las variables internas descritas en la siguiente sección, que serán las utilizadas como variables explicativas en el modelo.

1.4. Construcción de las variables

El objetivo principal del trabajo es obtener una modelización para la `FBC`, agregado macroeconómico que recoge el esfuerzo inversor realizado en una economía (ver Subsección 1.2.1). Es por ello que la selección de variables internas con las que trabajar se ha enfocado desde un primer momento a la búsqueda de aquellos productos orientados a personas jurídicas que a juicio de expertos de la entidad, tengan relación con el dinamismo inversor de las mismas. Para la identificación de variables que puedan ser relevantes se han mantenido reuniones con las áreas de negocio que tienen un conocimiento más extenso acerca de cuáles son los productos financieros de la entidad, así como cuáles son los más utilizados por las empresas para acometer sus inversiones. En esta línea, el objetivo se centra en encontrar series que sean reflejo de:

- **Inversión productiva**: vista como inversiones a largo plazo llevadas a cabo por las empresas a través de créditos con distintos tipos de garantía. Se ha decidido optar por intentar reconstruir series relacionadas con formalizaciones ya que estas presentan un mayor dinamismo que la evolución de los saldos.
- **Inversión recurrente**: entendida como la inversión necesaria para el desarrollo de la actividad diaria de las empresas. El enfoque se ha centrado en reconstruir series relacionadas con productos de circulante.
- **Calidad del riesgo**: variables relacionadas con el riesgo crediticio y la salud financiera de las empresas. Siguiendo las indicaciones de los expertos del área de calidad del riesgo se ha optado por reconstruir series relacionadas con impagos en líneas de descuento, excedidos en líneas de crédito y descubiertos en cuentas de depósito.

Además del reto de encontrar productos que reflejen la inversión en la economía gallega, nos encontramos ante un posible problema de "contaminación" en las series, provocado por diversos motivos como pueden ser, la representatividad de la entidad ABANCA en la economía gallega, las propias dinámicas internas de la entidad, como pueden ser las diferentes estrategias llevadas a cabo a la hora de maximizar su beneficio, o la elección del perímetro de aplicación de los agentes económicos implicados en la `FBC`.

En resumen, los problemas previos a la modelización estadística que tenemos que afrontar para poder construir las series temporales candidatas a entrar como variables explicativas del modelo son:

1. Elección de métricas para la inversión (ver Subsección 1.4.1).
2. Elección del público objetivo a tener en cuenta (ver Subsección 1.4.2).
3. Representatividad de ABANCA en la economía gallega (ver Subsección 1.4.3).
4. Coherencia temporal, entendida como la estabilidad muestral a la hora de reconstruir las series (ver Subsección 1.4.4).

1.4.1. Elección de métricas para la inversión

Las variables empleadas se han construido a partir de distintas tipologías de producto (circulante y préstamos a largo plazo), movimientos en cuenta (transferencias y nóminas) e indicadores de riesgo (impagos y descubiertos en cuenta), construyendo para cada uno de ellos distintas métricas que nos aportan información acerca de la inversión.

Esquema ilustrativo de las variables utilizadas:

- Variables
  - Activo Balance
    - **Circulante**
      - Factoring directo
      - Confirming
      - Descuento
      - Cuenta de crédito
    - **Préstamos a largo plazo**
      - Garantía personal
      - Garantía real
      - Leasing
  - Movimientos
    - **Transferencias**
    - **Nóminas**
  - Riesgo
    - **Impagos**
    - **Descubiertos en cuenta**

En donde los productos de **circulante** están relacionadas con la inversión recurrente, los productos de **préstamos a largo plazo** están relacionadas con la inversión productiva, **transferencias y nóminas** están relacionadas con el pulso de la actividad económica, y por último, **impagos y riesgos** están relacionados con la calidad del riesgo.

A continuación, describiremos cada una de ellas y aportaremos información acerca de las métricas utilizadas.

Productos de Circulante

Dentro de la agrupación de productos de circulante, encontraremos aquellos productos utilizados por las empresas para el desempeño de la actividad diaria, como pueden ser:

- **Cuenta de crédito**: Líneas de crédito con garantía personal o real.
- **Líneas de descuento comercial**: Operación de financiación bancaria a corto plazo, que consiste en adelantar al cliente el importe de los efectos o títulos de crédito (pagarés, letras, pagos domiciliados, etc.).
- **Productos factoring**: Servicio que consiste en la cesión, antes de su vencimiento, de los derechos de cobro de los créditos comerciales de un cliente al factor, quien prestará servicios administrativos y financieros al cedente.

- **Confirming**: Servicio por el cual una empresa encarga a la entidad financiera todos los aspectos relativos a la gestión de pago a sus proveedores por suministro de mercancías o servicios, con la posibilidad de acceder a otras alternativas financieras.

Para este producto recogemos la siguiente información:

- Límite mensual: Saldo máximo del que pueden disponer las empresas, el cual se asigna a partir de criterios de solvencia por lo que nos dará reflejo de la vitalidad del tejido empresarial gallego.
- Saldo medio dispuesto: En qué medida las empresas necesitan utilizar crédito u otros productos de circulante para poder financiar su actividad diaria.
- Ratio de disposición medio: Su construcción deriva del cociente entre el saldo medio dispuesto y el límite mensual de cada contrato, por tanto, como máximo un cliente solo podrá disponer del límite de su contrato provocando que el valor de dicho ratio sea igual a la unidad, en el caso en el cual el cliente no haga disposición de la la línea el valor de este ratio será cero.

Préstamos a largo plazo

Dentro de esta categoría encontraremos productos destinados a financiar la actividad empresarial o profesional, orientadas a más a largo plazo en comparación con los productos de circulante. Para estos productos recogemos la siguiente información:

- Número de formalizaciones: Recoge el número de préstamos formalizados cada mes.
- Saldo Medio formalizado: Importe medio formalizado para cada mes.

Impagos

Se recogen distintas métricas sobre los efectos impagados de las líneas de descuento que nos permitan captar un indicador de deterioro o riesgo en la actividad empresarial. Para los impagos obtenemos información como:

- Importe de efectos impagados: Se corresponde con la suma de todos los importes de los efectos impagados para cada mes.
- Importe de efectos totales: Se corresponde con la suma de todos los importes de los efectos de descuento emitidos cada mes.
- Proporción importe impagados: Se corresponde con la proporción de importes impagados sobre el importe total de efectos emitidos cada mes.
- Número de efectos impagados: Recoge el número de efectos impagados cada mes.
- Número de efectos totales: Suma de todos los efectos de descuento emitidos cada mes.
- Proporción de efectos impagados: Ratio que recoge el porcentaje de efectos impagados sobre el total de efectos emitidos cada mes.

Descubiertos

Recogen información acerca de cuentas a la vista con descubiertos más del 50 % de los días en los últimos 3 meses. De esta manera, obtenemos las siguientes métricas:

- Número de días medio de descubierto: Para cada mes tenemos información sobre el número de días en descubierto en las cuentas a la vista.
- Importe máximo descubierto: Nos aporta información sobre la media del los importes máximos descubiertos por mes.
- Importe descubierto medio: Representa el importe medio de los descubiertos para cada mes.

Transferencias y nóminas

En referencia a las nóminas, se tendrán en cuenta nóminas de los clientes de ABANCA que tienen contratado el servicio de nóminas, recogiendo en cierta medida el dinamismo del empleo a través del número de nóminas por mes, o el importe de las mismas.

En cuanto a las transferencias, se tienen en cuenta las realizadas por los clientes de ABANCA y las que reciben clientes de la entidad desde cuentas de la propia entidad y desde otras entidades. Con estas variables intentaremos recoger el dinamismo generado entre empresas, captando de alguna manera los pagos realizados derivados de las relaciones bilaterales entre las mismas.

De esta manera, obtenemos las siguientes métricas:

- Importe medio de las nóminas.
- Número de nóminas.
- Importe de transferencias recibidas y emitidas por clientes de ABANCA.
- Número de transferencias recibidas y emitidas por clientes de ABANCA.

1.4.2. Elección del perímetro de aplicación

Una vez elegidas las métricas a utilizar nos encontramos con la problemática de encontrar cuál es el público objetivo a tener en cuenta en la muestra. Dado que la formación bruta de capital hace referencia a las inversiones que se cometen por los agentes privados de la economía, nos centramos en aquellos productos relacionados con personas jurídicas (empresas). Dentro de la categoría de empresa, debemos decidir en función del tamaño de las mismas, cuáles son las que debíamos introducir en la muestra. Destacar, que la clasificación por tipología de empresa se rige por la normativa de la UE, [7] recogida en el manual Comisión Europea (2015). En este documento se establece la categorización de la tipología de empresa en función a determinados parámetros:

- **Micro empresa**: se corresponderá con aquella entidad con un número de empleados en plantilla inferior a 10 personas, un volumen de negocios inferior o igual a 2 millones de euros, o un balance inferior o igual a 2 millones de euros.
- **Pequeña empresa**: se denomina pequeña empresa a aquella entidad con un número de empleados en plantilla inferior a 50 personas, un volumen de negocios inferior o igual a 10 millones de euros, o un balance inferior a 10 millones de euros.
- **Mediana empresa**: será aquella entidad que tenga un número de empleados en plantilla inferior a 250 personas, con un volumen de negocio inferior o igual a 50 millones de euros, o un balance inferior a 43 millones de euros.
- **Gran empresa**: se corresponderán con aquellas empresas que sobrepasen los umbrales anteriormente mencionados.

Esta categorización es la utilizada por la entidad ABANCA, y por tanto, la incluida en este trabajo.

Además de las personas jurídicas, debemos de incluir un agente, que aun teniendo la denominación de persona física, es partícipe de la actividad empresarial de cualquier economía. Estamos hablando de los **Autónomos**, en este trabajo se ha aplicado la definición de autónomos utilizada por el área de Márketing de la entidad. En esta definición se establece como autónomo, a toda persona física que posea productos financieros clasificados como productos propios de empresa y, además, se categorizarán como autónomos a aquellos clientes que según bases de datos externas compradas por la entidad, estén reflejados como tal,

[7] Reglamento (UE) nº 651/2014 de la comisión.

y dichas personas hayan dado su consentimiento según el Reglamento General de Protección de Datos[8].

Reglamento General de Protección de Datos implica que es necesaria la obtención y gestión de nuevos consentimientos que hasta el momento no se registraban. En concreto, para el perfilado en base a bases de datos externas, si el cliente acepta, ABANCA podrá acceder a fuentes externas para enriquecer o completar la información que ya tiene del cliente y así poder ofrecerle productos más ajustados a sus gustos, preferencias y necesidades, según su concreto perfil.

1.4.3. Representatividad de ABANCA

Una cuestión a tener en cuenta en este estudio es la representatividad de los datos obtenidos a través de las variables internas de la entidad. Es decir, si podemos asumir que la muestra obtenida mediante las variables internas de la entidad ABANCA es representativa de la economía gallega en su conjunto. Según datos de evolución de negocio reportados por el Banco de España, ABANCA gestiona más de un tercio del negocio financiero de la comunidad gallega, tanto en créditos como en depósitos. Por tanto, este hecho justifica que los datos de ABANCA son una muestra representativa del comportamiento de la economía gallega.

1.4.4. Coherencia temporal

Por último, tenemos que tener en cuenta una problemática que afecta a las unidades de medición de todas las series. Dicha problemática se corresponde con la evolución del número de clientes o el número de contratos formalizados por la entidad a lo largo del período que es objeto de estudio.

En este trabajo se busca obtener un conjunto de series relacionadas con la `FBC` que reflejen el comportamiento de la economía gallega en términos de inversión. Pero desde el momento en el que la fuente de obtención de datos es una entidad bancaria que se ve afectada en parte por los ciclos de la economía y también por su propia estrategia, podría ser que estuviésemos encontrando patrones de comportamiento que no fuesen un fiel reflejo de la evolución de la inversión de la economía gallega, si no del buen o mal hacer de la entidad ABANCA en el desempeño de su ejercicio económico como otro agente más del conjunto de la economía. Por tanto, para eliminar esa componente de crecimiento o estrategia de la entidad, se ha decidido extraer todas las variables en términos relativos en función del número de contratos por mes.

1.4.5. Depuración de datos

A continuación, se ilustrarán algunas de las estrategias que se han llevado a cabo para solucionar las cuestiones que se han expuesto anteriormente[9].

1. Durante la limpieza y depuración de las series, se ha optado por la eliminación de las grandes empresas en la muestra por varios motivos. En primer lugar, la poca adecuación de las mismas a la coyuntura del ciclo económico (nótese que las grandes empresas pueden ser entendidas como grandes barcos que tardan en reaccionar ante cambios en el mercado y con suficiente poder y margen de maniobra para afrontar situaciones económicamente desfavorables). En segundo lugar, los importes de las operaciones que realizan introducen "ruido" en las series debido a sus elevadas cifras. Por último, en base a criterio de los expertos de las distintas áreas de negocio, la cuota de mercado de la entidad es menor, tanto en cuanto mayor es el tamaño de la empresa, ya que es normal que una gran empresa recurra a varias entidades bancarias para obtener distintos tipos de financiación.

A continuación, en la Figura 1.4, se muestran las tasas de variación interanual del saldo del producto de circulante. Se puede ver cómo las tasas de variación de las grandes empresas, representadas por una línea discontinua de color rojo, siguen un comportamiento diferente al resto de tipologías de empresas

[8]El Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo, de 27 de abril de 2016, relativo a la protección de las personas físicas en lo que respecta al tratamiento de sus datos personales y a la libre circulación de estos datos y por el que se deroga la Directiva 95/46/CE (Reglamento General de Protección de Datos), es plenamente aplicable en España desde el pasado 25 de mayo.

[9]Los ejes de todas las figuras se han eliminado por cuestiones de confidencialidad.

consideradas en este trabajo (autónomos, micro, pequeña y mediana empresa), representadas por una línea sólida de color azul.

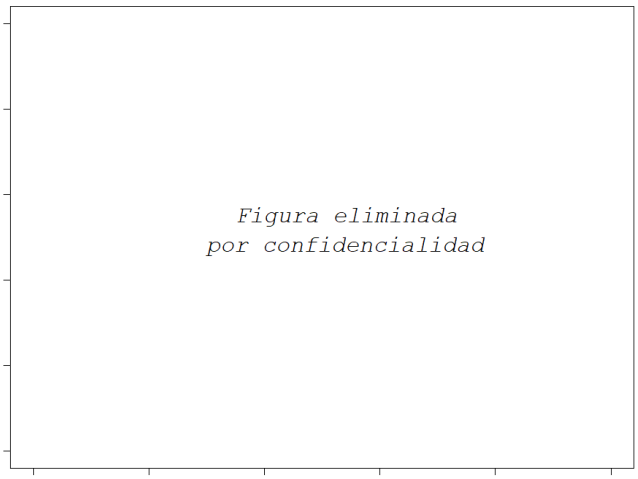


Figura 1.4: Tasas de variación del producto del saldo del circulante por tipología de empresa.

2. A lo largo del período temporal contemplado en este trabajo, el catálogo de productos financieros ofrecidos a las empresas no es estable debido al dinamismo comercial. Ya que las preferencias de los consumidores no se mantienen constantes a lo largo del tiempo, provocando que el catálogo de productos ofrecido sufra modificaciones, las cuales vendrán motivadas principalmente por:

- Cambios en la categorización de los productos debido a la adecuación de los mismos en base a las nuevas preferencias de los consumidores.
- Aparición de nuevos productos en el catálogo con el fin de satisfacer nuevas necesidades.

Para solucionar el problema relacionado con la estabilidad de los productos financieros existentes en el período de análisis, se ha decidido tomar como producto del contrato el último producto que tiene informado[10].

En la Figura 1.5 se muestra ejemplo de cómo la aparición de nuevos productos puede provocar cambios en las series, es el caso particular de un producto vinculado con el pago de impuestos.

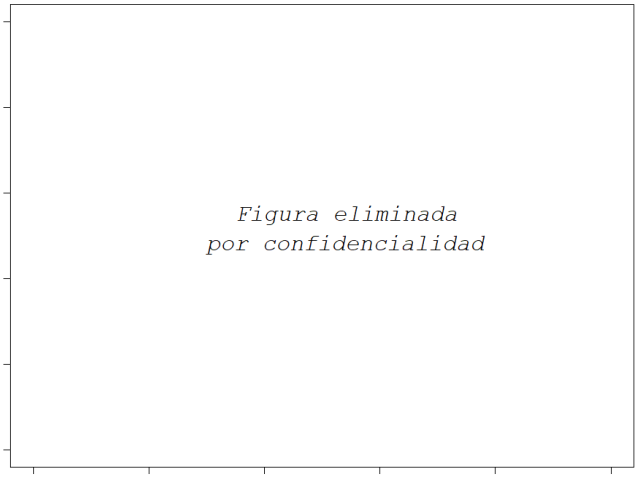
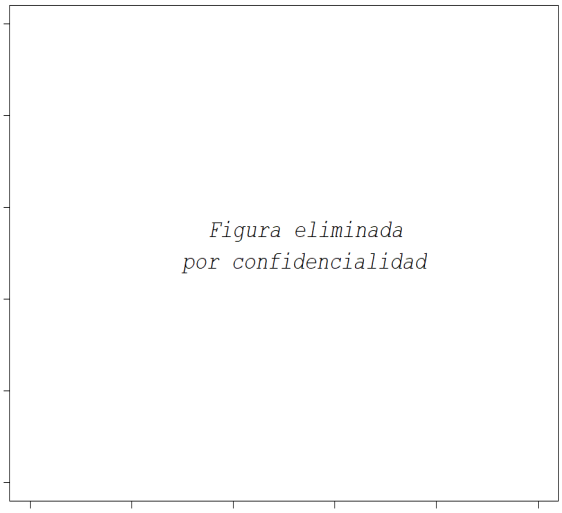


Figura 1.5: Número de formalizaciones por mes.

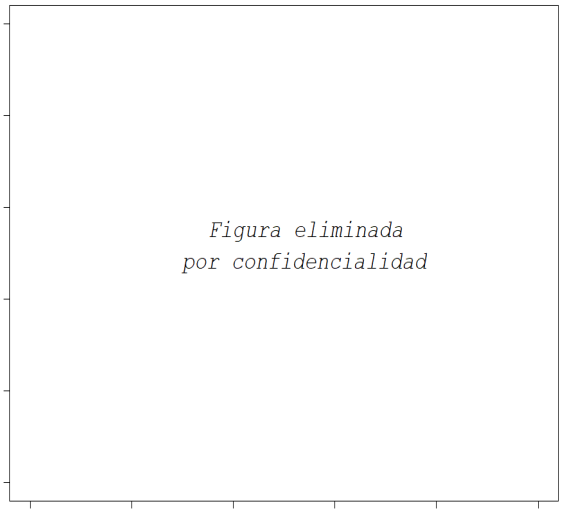
[10] Destacar que por una parte, comprender la operativa de cada uno de los productos financieros de la entidad requiere bastante tiempo, y por otra, conocer las distintas estrategias de la misma a lo largo del tiempo, que pueden afectar a la reconstrucción de las series, resulta a veces inabarcable.

Se ha detectado que a partir de una fecha concreta, se pone en marcha un producto especifico para el pago de impuestos dentro de la categoría de productos de largo plazo. Esto provoca una distorsión en la serie de formalizaciones a partir de la fecha de puesta en marcha de este producto, donde el número de formalizaciones por mes de productos de préstamos a largo plazo comienza a presentar un claro patrón estacional señalado a la derecha de la línea roja discontinua de la Figura 1.5.

3. Haciendo referencia a la problemática de la coherencia temporal, se puede apreciar que en la Figura 1.6a, el número de contratos ABANCA para el producto de circulante, y el número de nóminas pagadas por empresas con el servicio de nóminas de ABANCA aumenta a lo largo del período. Tal comportamiento puede derivarse en parte a los ciclos económicos, pero también a la evolución de las estrategias de la compañía. Dado que estos dos efectos están mezclados y son difícilmente separables, para evitar esta problemática se ha decidido extraer todas las variables en función del número de contratos por mes, consiguiendo de esta forma medidas relativas que amortigüen en cierta media este efecto.



(a) Número de contratos de circulante.



(b) Número de nóminas por mes.

Figura 1.6: Evolución del número de contratos para productos de circulante y el servicio de pago de nóminas.

En resumen, tras un estudio exhaustivo de cada uno de los productos financieros relacionados con la inversión privada gallega, reuniones con distintas áreas de negocio de la entidad para conocer el funcionamiento de los mismos, y la corrección y depuración mediante *IBM SPSS Modeler* de cada una de las series, se ha conseguido construir una batería de 23 métricas distintas reflejadas en la Tabla 1.5. Estas métricas nos aportan información, a cierre de mes, acerca del comportamiento de las empresas en el corto plazo, las inversiones que éstas realizan en el medio y largo plazo, así como el riesgo o deterioro de la actividad empresarial de la comunidad autónoma de Galicia que nos permitirán modelizar la `FBC`.

Métrica	Variables	Descripción
Circulante	CIRC_RATIO_Mean	Ratio de la disposición media de las líneas de circulante
Circulante	CIRC_SALDO_Mean	Saldo dispuesto medio en las líneas de circulante
Circulante	CIRC_LIMITE_Mean	Límite medio de las lineas de circulante
Circulante	CIRC_NUMCONT	Número de contratos de circulante
Circulante	CIRC_SALDO_EXC_Mean	Saldo excedido medio en las líneas de circulante
Circulante	CIRC_RATIO_EXC_Mean	Ratio de saldo excedido en líneas de circulante
Préstamos a largo plazo	LARG_NUMFORMA_Sum	Número de formalizaciones en líneas de largo
Préstamos a largo plazo	LARG_SALDOFORM_Mean	Saldo medio de las formalizaciones en líneas de largo
Nóminas	NOM_IMPT_Mean	Importe medio de las nóminas
Nóminas	NOM_NUM_Sum	Número de nóminas
Transferencias	TRANSF_E_IMPT_Mean	Importe de transferencias emitidas por clientes de ABANCA
Transferencias	TRANSF_E_NUM_Sum	Número de transferencias emitidas por clientes de ABANCA
Transferencias	TRANSF_R_IMPT_Mean	Importe de transferencias recibidas por clientes de ABANCA
Transferencias	TRANSF_R_NUM_Sum	Número de transferencias recibidas por clientes de ABANCA
Impagos	IMP_IMPAGADO	Saldo impagado medio en líneas de descuento
Impagos	IMP_TOTALES	Saldo medio en líneas de descuento
Impagos	IMP_PROP	Proporción de impagados en líneas de descuento
Impagos	IMP_N_EFE_IMP	Número de efectos impagados en líneas de descuento
Impagos	IMP_N_EFE_TOT	Número de efectos emitidos en líneas de descuento
Impagos	IMP_PROP_EFECT	Proporción de impagados en líneas de descuento
Descubiertos	DESC_N_DIAS	Número de días en descubierto para cuentas a la vista
Descubiertos	DESC_MAX	Importe descubierto máximo para cuentas a la vista
Descubiertos	DESC_MEDIO	Importe descubierto medio para cuentas a la vista

Tabla 1.5: Variables construidas a partir de variables de negocio de ABANCA.

Capítulo 2

Modelos Box-Jenkins

En este capítulo se hará una breve introducción sobre los conceptos más relevantes de la metodología Box-Jenkins, en la que se engloba una clase de procesos a partir de los cuales se puede modelizar la evolución de múltiples series de tiempo, entendidas éstas como realizaciones de un proceso estocástico.

A continuación, en la Sección 2.1 y Sección 2.2 se muestran diferentes procesos que permiten la modelización de múltiples series de tiempo en función de las características de las mismas, en concreto los modelos $ARMA$ y $ARIMA$. En la Sección 2.3 se muestran las herramientas necesarias para aproximar los órdenes de los parámetros generadores de las series. En la Sección 2.4 se analizan los distintos métodos de estimación de los parámetros. Una vez estimados los parámetros del modelo es necesario comprobar si se cumplen las hipótesis básicas del mismo como veremos en la Sección 2.5, y entre la duda de varios modelos usaremos los criterios de la Sección 2.6 para decidir cual es el mejor. Por último, si el modelo es válido en la Sección 2.7 se mostrará cómo realizar predicciones con el mismo.

En resumen, la estructura del capítulo girará en torno al proceso de construcción de un modelo válido bajo la metodología Box-Jenkis, la cual se asienta sobre el cuadro conceptual que se muestra en la Figura 2.1:

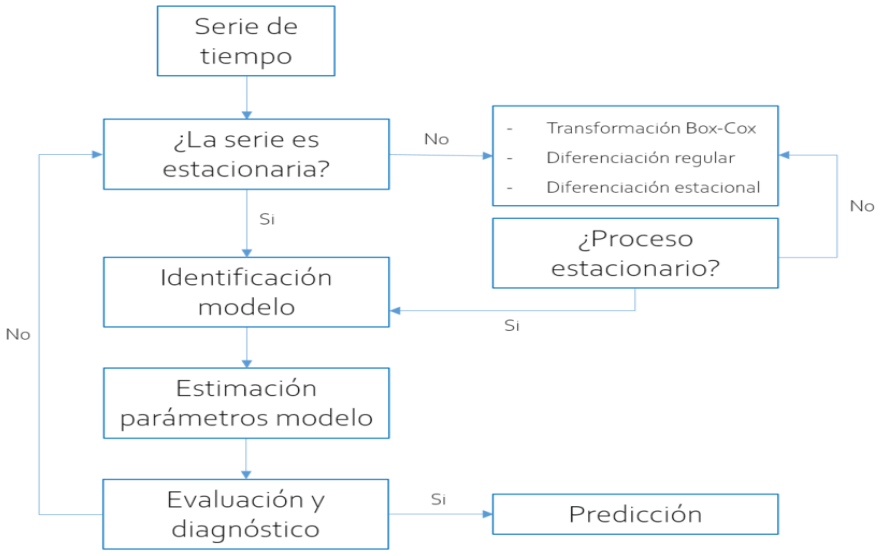


Figura 2.1: Metodología Box-Jenkins.

2.1. Modelos para series estacionarias

2.1.1. Proceso autorregresivo AR(p)

Un proceso autorregresivo de orden p es un proceso estacionario $\{X_t\}_{t\in\mathbb{R}}$ que admite la representación:

Definición 2.1.1. *Proceso $AR(p)$*

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + a_t$$

donde $c,\ \phi_1, \ldots, \phi_p$ son constantes y $\{a_t\}_{t\in\mathbb{R}}$ un proceso de ruido blanco conocido como innovaciones.

Es decir, un proceso autorregresivo de orden p, es aquel proceso cuyo valor actual depende linealmente de los p instantes temporales anteriores.

La ecuación anterior puede reescribirse en función del operador retardo, tal que:

$$(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p) X_t - c = a_t$$

donde B se corresponde con el operador retardo $BX_t = X_{t-1}$.

Para estos procesos se verifica que:

- El proceso será estacionario si la ecuación característica del proceso no tiene raíces unitarias. En otros términos, cuando se satisface:

$$1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p \neq 0 \qquad \text{para todo } z \text{ tal que } \mid z \mid = 1$$

- El proceso será causal si se puede expresar como combinación lineal de un ruido blanco y las raíces de la ecuación característica del proceso están fuera del círculo unidad. Es decir, se satisface:

$$1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p \neq 0 \qquad \text{para todo } z \text{ tal que } \mid z \mid \leq 1$$

- Siempre es un proceso invertible.

2.1.2. Proceso de medias móviles MA(q)

Se conoce como un proceso de medias móviles de orden q a un proceso estacionario $\{X_t\}_{t\in\mathbb{R}}$ que admite la representación:

Definición 2.1.2. *Proceso $MA(q)$*

$$X_t = c + a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \cdots + \theta_q a_{t-q}$$

donde $c,\ \theta_1, \ldots, \theta_p$ son constantes y $\{a_t\}_{t\in\mathbb{R}}$ un proceso de ruido blanco conocido como innovaciones.

Es decir, un proceso de medias móviles de orden q, es aquel proceso cuyo valor depende únicamente de los q retardos de las innovaciones del proceso.

La ecuación anterior puede reescribirse en función del operador retardo tal que:

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

siendo B el operador retardo $Ba_t = a_{t-1}$.

Para estos procesos se verifica que:

- El proceso es estacionario y causal.
- Será invertible si se satisface:

$$1 + \theta_1 z + \theta_2 z^2 + \cdots + \theta_q z^q \neq 0 \qquad \text{para todo } z \text{ tal que } \mid z \mid \leq 1$$

Destacar que, los procesos $MA(q)$ siempre serán estacionarios, dado que son una combinación lineal de ruidos blancos, y un ruido blanco siempre es estacionario.

2.1.3. Proceso ARMA(p,q)

Se conoce como un proceso $ARMA(p,q)$ a un proceso estacionario $\{X_t\}_{t\in\mathbb{R}}$ que admite la representación:

Definición 2.1.3. *Proceso* $ARMA(p,q)$

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \cdots + \theta_q a_{t-q}$$

donde $c, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q$ *son constantes.*

Del mismo modo que los procesos anteriores, un proceso $ARMA(p,q)$ puede reescribirse de forma compacta tal que:

$$\phi(B)X_t = c + \theta(B)a_t$$

en donde

$$\phi(B) = (1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p)$$

$$\theta(B) = (1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q)$$

con B definido como operador retardo.

Para estos procesos se verifica que:

- La constante será el producto de la media del proceso por los coeficientes autorregresivos del mismo $c = \mu(1 - \phi_1 - \cdots - \phi_p)$.
- El proceso será estacionario si la ecuación característica del proceso no tiene raíces unitaria, i.e. si se satisface:
$$1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p \neq 0 \qquad \text{para todo } z \text{ tal que } \mid z \mid = 1$$
.
- El proceso será causal si se puede expresar como combinación lineal de un ruido blanco tal que:
$$1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p \neq 0 \qquad \text{para todo } z \text{ tal que } \mid z \mid \leq 1.$$
- Será invertible si se satisface:
$$1 + \theta_1 z + \theta_2 z^2 + \cdots + \theta_q z^q \neq 0 \qquad \text{para todo } z \text{ tal que } \mid z \mid \leq 1.$$

Conviene destacar que los modelos ARMA(p,q) son muy flexibles y permiten modelizar gran cantidad de series generadas por procesos estacionarios. En efecto, si $\{Y_t\}_{t\in\mathbb{R}}$ es un proceso estacionario con autocovarianzas de retardo h tendiendo a cero con h, i.e. $\gamma_{Y,h} \to 0$ cuando $h \to \infty$, entonces para un entero arbitrario $k > 0$ existe un proceso $\{X_t\}_{t\in\mathbb{R}}$ siguiendo un modelo ARMA tal que $\gamma_{X,h} = \gamma_{Y,h}$ para todo $h = 0, 1, \ldots, k$.

2.2. Modelos para series no estacionarias

Como hemos visto en la sección anterior, los procesos $ARMA$ permiten modelizar gran variedad de procesos estacionarios, aunque teóricamente es una cualidad interesante y a destacar de estos modelos, en la realidad no abundan series reales generadas por procesos estacionarios si no que suelen estar afectadas por:

1. **Heterocedastidad**: La variabilidad de la serie no es constante en el tiempo.
2. **Tendencia**: El nivel de la serie no es estable en el tiempo.
3. **Componente estacional**: La serie suele presentar patrones repetitivos.

A continuación, se explicarán cuales son los procedimientos que nos permiten solventar cada una de estas problemáticas.

2.2.1. Heterocedasticidad

La problemática de la presencia de heterocedasticidad puede ser resuelto mediante la familia de transformaciones Box-Cox. Sea $\{X_t\}_{t\in\mathbb{R}}$ un proceso no estacionario, la familia de transformaciones Box-Cox depende de un parámetro λ a determinar previamente y viene definida por:

$$Y_t^\lambda = \begin{cases} \frac{x_t^\lambda - 1}{\lambda} \ si \ \lambda \neq 0 \\ log(x_t) \ si \ \lambda \ = \ 0 \end{cases}$$

El valor apropiado de λ se obtiene por el método de máxima verosimilitud. La serie transformada Y_t^λ será estacionaria y susceptible de ser modelada con cualquiera de los modelos expuestos en la sección anterior.

2.2.2. Tendencia

Sea el proceso $X_t = \beta_0 + \beta_1 t + V_t$, un proceso no estacionario debido a que tiene una tendencia determinista, y $\{V_t\}_{t\in\mathbb{R}}$ un proceso estacionario. El proceso diferenciando $X_t - X_{t-1} = \beta_1 + (V_t - V_{t-1})$ es estacionario. Por tanto, si una serie presenta una componente de tendencia, esta podrá ser eliminada aplicando sucesivamente d diferencias regulares.

Un proceso $ARIMA(p,d,q)$ es aquél proceso que, después de aplicarle d diferencias regulares, se convierte en un proceso $ARMA(p,q)$. Es decir:

$$\{X_t\}_{t\in\mathbb{R}} \ es \ ARIMA(p,d,q) \Leftrightarrow (1-B)^d X_t \ es \ ARMA(p,q)$$

en donde B se corresponde con el operador retardo definido por $BX_t = X_{t-1}$.

Una vez diferenciada regularmente la serie, $\{X_t\}_{t\in\mathbb{R}}$, es estacionaria, por lo que podremos ajustar cualquiera de los modelos expuestos en la sección anterior para modelar su comportamiento.

2.2.3. Componente estacional

La clase de procesos $ARIMA$ expuestos permite capturar la no estacionariedad provocada por la presencia de tendencia pero no capturan la estacionariedad provocada por una componente estacional, entendida ésta como la dependencia entre observaciones ocurridas en instantes temporales separados por múltiplos del período estacional s.

Sea el proceso $X_t = S_t + V_t$, donde $\{S_t\}_{t\in\mathbb{R}}$ no es estacionario pero $\{V_t\}_{t\in\mathbb{R}}$ sí, y $S_t = S_{t-s}$ se corresponde con una tendencia estacional determinista. Entonces el proceso diferenciado $X_t - X_{t-s} = V_t - V_{t-s}$ es estacionario. Por tanto, si una serie presenta dependencia estacional, al aplicarle D diferencias estacionales se convierte en un proceso $ARMA(p,q)$. Es decir:

$$\{X_t\}_{t\in\mathbb{R}} \ es \ ARIMA(P,D,Q) \Leftrightarrow (1-B^s)^D X_t \ es \ ARMA(P,Q)_s$$

con operador retardo estacional $B^s X_t = X_{t-s}$.

Se conoce como un proceso estacional $ARMA(P,Q)_s$ a un proceso estacionario $\{X_t\}_{t\in\mathbb{R}}$ que admite la representación:

Definición 2.2.1. *Proceso* $ARMA(P,Q)_s$

$$X_t = c + \Phi_1 X_{t-s} + \Phi_2 X_{t-2s} + \cdots + \Phi_P X_{t-Ps} + a_t + \Theta_1 a_{t-s} + \Theta_2 a_{t-2s} + \cdots + \Theta_Q a_{t-Qs}$$

donde $c, \Phi_1, \ldots, \Phi_P, \Theta_1, \ldots, \Theta_Q$ *son constantes.*

2.2.4. Tendencia y componente estacional

Si las series se viesen afectadas tanto por dependencia regular (tendencia), como por dependencia estacional, el proceso que nos permite modelar tal situación se correspondería con un proceso ARIMA estacional multiplicativo $ARIMA(p,d,q)\times(P,D,Q)_s$.

Un proceso $ARIMA(p,d,q)\times(P,D,Q)_s$ es aquel proceso que después de aplicarle d diferencias regulares y D diferencias estacionales de período s se convierte en un $ARMA(p,q)\times(P,Q)_s$.

Del mismo modo, diremos que un proceso $\{X_t\}_{t\in\mathbb{R}}$ es un *ARIMA estacional multiplicativo* si admite una representación tal que:

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

donde:

$$\begin{aligned}
\phi(B) &= (1-\phi_1 B-\phi_2 B^2-\cdots-\phi_p B^p)\\
\theta(B) &= (1+\theta_1 B+\theta_2 B^2+\cdots+\theta_q B^q)\\
\Phi(B^s) &= (1-\Phi_1 B^s-\Phi_2 B^{2s}-\cdots-\Phi_p B^{sP})\\
\Theta(B^s) &= (1+\Theta_1 B^s+\Theta_2 B^{2s}+\cdots+\Theta_q B^{sQ})
\end{aligned} \tag{2.1}$$

correspondiendose B con el operador retardo y B^s con el operador retardo estacional.

El proceso *ARIMA estacional multiplicativo* es posiblemente el más utilizado en la modelización de series de tiempo univariantes debido a su flexibilidad para ajustar múltiples series temporales.

2.3. Identificación

Para identificar una modelo ARMA como posible generador de una serie de tiempo se necesita determinar sus órdenes apropiados. Para ello se siguen una serie de pasos que se muestran a continuación de manera esquemática, tras introducir algunos conceptos previos de interés.

Conceptos previos:

Definición 2.3.1. *Sea $\{X_t\}=(x_1,x_2,\ldots,x_T)$ una serie temporal, con $t=1,\ldots,T$. Se define la* ***media muestral*** *como:*

$$\bar{x}=\frac{1}{T}\sum_{t=1}^{T}x_t$$

Definición 2.3.2. *Sea $\{X_t\}=(x_1,x_2,\ldots,x_T)$ una serie temporal, con $t=1,\ldots,T$. Se define la* ***función de autocovarianzas muestrales*** *entre observaciones separadas k instantes temporales como:*

$$\hat{\gamma}_k=\frac{1}{T}\sum_{t=1}^{T-k}(x_t-\bar{x})(x_{t-k}-\bar{x})$$

con $\hat{\gamma}_k=\hat{\gamma}_{-k},\ para\ k=0,1,\ldots,T-1$.

Definición 2.3.3. *Sea $\{X_t\}=(x_1,x_2,\ldots,x_T)$ una serie temporal, con $t=1,\ldots,T$. La* ***función de autocorrelaciones simples muestrales (FAS)*** *asigna a cada retardo k el valor:*

$$\hat{\rho}_k=\frac{\hat{\gamma}_k}{\hat{\gamma}_0}$$

Definición 2.3.4. *Sea* $\{X_t\} = (x_1, x_2, \ldots, x_T)$ *una serie temporal, con* $t = 1, \ldots, T$. *La* ***función de autocorrelaciones parciales muestrales (FAP)*** *asigna a cada retardo* k *el valor:*

$$\hat{\sigma}_k = \hat{\sigma}_{kk}$$

donde $\hat{\sigma}_{kk}$ *se corresponde con el estimador mínimo cuadrático de* σ_{kk} *en la regresión* $x_t = \sigma_{k0} + \sigma_{k1}x_{t-1} + \cdots + \sigma_{kk}x_{t-k} + \varepsilon$.

La estructura de las funciones de autocorrelación simples y parciales de los modelos expuestos hasta ahora se corresponden con la Tabla 2.1:

Proceso	**FAS** $\rho(k)$	**FAP** $\sigma(k)$
$AR(p)$	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*

Tabla 2.1: Comportamientos de las FAS y las FAP para la identificación de procesos. Con * se indica que los coeficientes no múltiplos de P (modelo AR), de Q (modelo MA) o de P(parte AR del modelo ARMA) o Q (parte MA del modelo ARMA), son nulos.

Una vez presentados los conceptos previos los pasos a seguir para la identificación de los órdenes del proceso generador de la serie se corresponden con:

1. Si la serie presenta heterocedasticidad, eliminarla a través de una transformación Box-Cox.
2. Si la serie (puede que transformada en el Paso 1) presenta tendencia, eliminarla a través de la diferenciación regular.
3. Si la serie (puede que transformada en los Pasos 1 y 2) presenta estacionalidad, eliminarla mediante la diferenciación estacional.
4. Identificar un modelo ARMA para la serie a través de las funciones de autocorrelación simple y parcial.

Para saber cuales son los órdenes del proceso generador de la serie temporal nos basaremos en la información suministrada por la **FAP** y la **FAS** muestrales ($\hat{\rho}_k$ y $\hat{\sigma}_k$ respectivamente). Dado que $\hat{\rho}_k$ y $\hat{\sigma}_k$ dependen de la serie de tiempo observada, para poder sacar conclusiones deberemos de conocer su distribución muestral.

Proposición 2.3.1. *Bajo condiciones generales y un tamaño muestral suficientemente grande se verifica que la distribución muestral de* $\hat{\sigma}_k$*:*

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

Proposición 2.3.2. *Bajo condiciones generales y un tamaño muestral suficientemente grande se verifica que la distribución muestral de* $\hat{\rho}_k$*:*

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

A través de las proposiciones anteriores podemos concluir que si la serie ha sido generada por un proceso AR(p) debería cumplirse que para cada $k > p$ (con una significación al 5 %):

$$\hat{\sigma}_k \in \left(-\frac{1,96}{\sqrt{T}}, \frac{1,96}{\sqrt{T}} \right),$$

y de la misma forma si la serie ha sido generada por un proceso MA(q) debería cumplirse que, para cada $k > q$ (con una significación al 5 %):

$$\mid \hat{\rho}_k \mid \leq 1,96 \sqrt{\frac{1 + 2(\hat{\rho}_1^2 + \cdots + \hat{\rho}_q^2)}{T}}$$

Por tanto, una vez conocida la distribución muestral de los coeficientes de autocorrelación muestral, podremos detectar gráficamente los coeficientes no nulos, los cuales se corresponderán con aquellos que sobrepasen las bandas de aceptación rechazo.

2.4. Estimación

Una vez elegido el posible modelo generador de la serie $\{X_t\} = (X_1, X_2, \ldots, X_T)$ con $t = 1, \ldots, T$, es necesario estimar los parámetros del mismo[1].

Supongamos que la serie de tiempo $\{X_t\} = (X_1, X_2, \ldots, X_T)$ ha sido generada por un proceso ARMA(p,q) cuya representación se corresponde con:

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \cdots + \theta_q a_{t-q}$$

Los parámetros susceptibles de ser estimados se corresponden con $c, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q$ y σ_a. En esta sección analizaremos la estimación mediante mínimos cuadrados y máxima verosimilitud.

2.4.1. Estimación mediante mínimos cuadrados

Los residuos asociados a las estimaciones $\widetilde{c}, \widetilde{\phi}_1, \ldots, \widetilde{\phi}_p, \widetilde{\theta}_1, \ldots, \widetilde{\theta}_q$, se definen, para $t = 1, \ldots, T$, como la diferencia entre los valores observados y los correspondientes valores ajustados, esto es,

$$\hat{a}_t = X_t - (\widetilde{c} + \widetilde{\phi}_1 X_{t-1} + \cdots + \widetilde{\phi}_p X_{t-p} + \widetilde{\theta}_1 \hat{a}_{t-1} + \cdots + \widetilde{\theta}_q \hat{a}_{t-q}).$$

La estimación de los parámetros $(c, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q)$ por medio del método de mínimos cuadrados conduce a los valores $(\hat{c}, \hat{\phi}_1, \ldots, \hat{\phi}_p, \hat{\theta}_1, \ldots, \hat{\theta}_q)$ que minimizan la función:

$$S(\widetilde{c}, \widetilde{\phi}_1, \ldots, \widetilde{\phi}_p, \widetilde{\theta}_1, \ldots, \widetilde{\theta}_q) = \sum_{t=1}^{T} \hat{a}_t^2.$$

2.4.2. Estimación mediante mínimos cuadrados condicionados

Existe una dificultad asociada a la estimación por mínimos cuadrados si $p > 0$ para obtener los residuos $\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_p$ debido a que dependen de los valores no observados de $X_0, X_{-1}, \ldots, X_{1-p}$.

Esta dificultad se puede solventar minimizando la función:

$$S_C(\widetilde{c}, \widetilde{\phi}_1, \ldots, \widetilde{\phi}_p, \widetilde{\theta}_1, \ldots, \widetilde{\theta}_q) = \sum_{t=p+1}^{T} \hat{a}_t^2$$

en lugar de la función $S(\cdot)$.

[1] A efectos de simplicidad en la presente sección se basará en un proceso ARMA(p,q).

A su vez, $\hat{a}_{p+1}$ depende de los valores de $\hat{a}_p, \hat{a}_{p-1}, \ldots, \hat{a}_{p+1-q}$, los cuales dependen de valores no observados de X_t. Nótese que a partir de la serie observada y fijados los valores de $\hat{a}_p, \hat{a}_{p-1}, \ldots, \hat{a}_{p+1-q}$ es viable obtener iterativamente los valores de $\hat{a}_{p+1}, \hat{a}_{p+2}, \ldots, \hat{a}_T$.

En definitiva, la estimación de los parámetros $(c, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q)$ por medio del método de mínimos cuadrados condicionados conduce a los valores $(\hat{c}, \hat{\phi}_1, \ldots, \hat{\phi}_p, \hat{\theta}_1, \ldots, \hat{\theta}_q)$ que minimizan:

$$S_C(\widetilde{c}, \widetilde{\phi}_1, \ldots, \widetilde{\theta}_q) = \sum_{t=p+1}^{T} \hat{a}_t^2, \qquad \text{condicionado a} \quad \hat{a}_p = \hat{a}_{p-1} = \cdots = \hat{a}_{p+1-q} = 0.$$

2.4.3. Estimación mediante máxima verosimilitud

La estimación por máxima verosimilitud de los parámetros $c, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q$ y σ_a^2 se obtiene a través de los valores que dan mayor credibilidad a la serie observada $x_1, \ldots, x_T$, i.e. los valores $\widetilde{c}, \widetilde{\phi}_1, \ldots, \widetilde{\theta}_q$ y $\widetilde{\sigma}_a^2$ maximizando:

$$L_{x_1,\ldots,x_T}(\widetilde{c}, \widetilde{\phi}_1, \ldots, \widetilde{\theta}_q, \widetilde{\sigma}_a^2) = f_{\widetilde{c},\widetilde{\phi}_1,\ldots,\widetilde{\theta}_q,\widetilde{\sigma}_a^2}(x_1, \ldots, x_T),$$

donde $f_{\widetilde{c},\widetilde{\phi}_1,\ldots,\widetilde{\theta}_q,\widetilde{\sigma}_a^2}$ denota la función de densidad conjunta de un vector aleatorio $(\widetilde{X}_1, \ldots, \widetilde{X}_T)$' procedente de un proceso ARMA con parámetros $\widetilde{c}, \widetilde{\phi}_1, \ldots, \widetilde{\theta}_q$ y $\widetilde{\sigma}_a^2$.

Bajo condiciones adecuadas, se tiene que:

- Los estimadores por máxima verosimilitud de los parámetros $\mu, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q$ definiendo un modelo ARMA(p,q) gaussiano son asintóticamente óptimos; es decir, si el tamaño T de la serie es "grande", se puede considerar que:

  1. Son centrados (o insesgados).
  2. Son eficientes.
  3. Su distribución es normal.

- El estimador de máxima verosimilitud de σ_a^2 es consistente.

Debemos notar la importancia de la propiedad 3 que nos permite construir intervalos/regiones de confianza y contrastes de hipótesis referentes a los parámetros. Además, las propiedades 1 y 3 se mantienen para estimadores basados en la verosimilitud gaussiana, aunque el proceso no sea gaussiano.

2.5. Diagnosis

Tras ajustar el modelo ARMA se deben comprobar que las hipótesis básicas realizadas sobre él se verifican. Esto se conoce como la diagnosis o chequeo del modelo ajustado.

Estas hipótesis son las siguientes:

- La hipótesis más importante es la que exige que las innovaciones $\{a_t\}_{t\in\mathbb{R}}$ sean ruido blanco:

  - Tengan media cero.
  - Tengan varianza constante.
  - Estén incorreladas.

  Si estas hipótesis no se verifican invalida al modelo ajustado como posible generador de la serie de tiempo en estudio.

- La hipótesis de normalidad es conveniente por tres motivos:

  1. Bajo normalidad, el ruido blanco equivale a la independencia. Esto garantiza que no estamos dejando información por modelizar.

2. Bajo normalidad, los estimadores que utilizamos (máxima verosimilitud gaussiana) son asintóticamente eficientes.
3. Chequeado el modelo se realizarán predicciones de valores futuros del proceso, resultando conveniente que vayan acompañadas de intervalos de predicción. Si no tenemos normalidad, no podremos garantizar su nivel de confianza.

A continuación, se presentan algunos gráficos que pueden asesorar acerca de si una muestra $y_1, \ldots, y_T$ es o no una realización de un conjunto de variables aleatorias procedentes de un proceso de ruido blanco gaussiano.

- **El gráfico de la muestra frente al tiempo**
  La representación gráfica de la muestra frente al tiempo puede ayudar a detectar de manera visual y rápida la presencia de:
  - Tendencia.
  - Componente estacional.
  - Variabilidad no constante.
  - Dependencia lineal, que puede ser de dos formas:
    - Positiva: tendencias que desaparecen a corto plazo (tendencias locales).
    - Negativa: valores altos seguidos con frecuencia por valores bajos y viceversa (zig-zag).

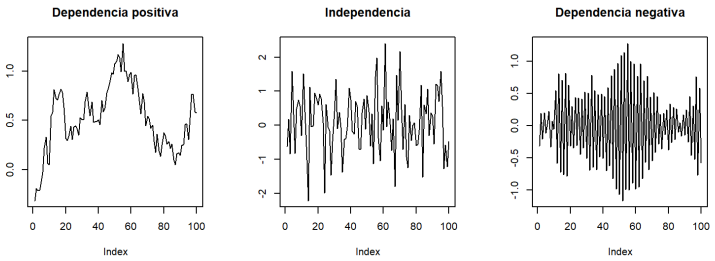


Figura 2.2: Dependencia lineal.

- **El gráfico Q-Q normal**
  El gráfico Q-Q (Cuantil-Cuantil) normal representa a los cuantiles muestrales frente a los cuantiles de una distribución $N(0, 1)$.

  Si la muestra $y_1, \ldots, y_T$ es i.i.d. con distribución normal, su gráfico Q-Q normal debería ser aproximadamente lineal.

  Esto nos lleva a que la no linealidad del gráfico Q-Q normal sugiere ausencia de normalidad.

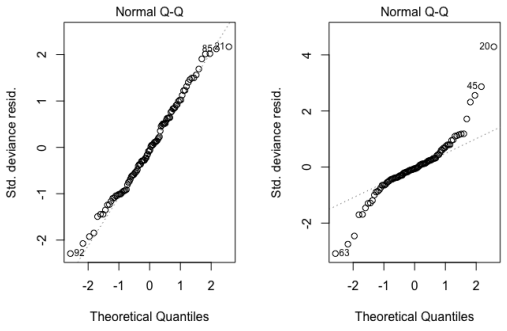


Figura 2.3: Normalidad y falta de normalidad.

También se utilizan una serie de pruebas de hipótesis diseñadas para chequear si una muestra $y_1, \ldots, y_T$ está formada por realizaciones:

- Independientes.
- Con media cero.
- Con distribución común gaussiana.

Estos contrastes son los siguientes:

- **Contraste de independencia**

  Denotemos por $\hat{\rho}_k$ $(k = 1, 2, \ldots)$ a la FAS de la muestra en estudio $y_1, \ldots, y_T$.

  Bajo la hipótesis nula de que la muestra proviene de variables aleatorias i.i.d. con varianza finita, y asumiendo que el tamaño muestral T es "grande", se tiene que:

$$\hat{\rho}_k \approx N\left(0, \frac{1}{\sqrt{T}}\right).$$

  Por tanto, la independencia (al 5 %) se rechazará si $|\hat{\rho}_k| \geq \frac{1{,}96}{\sqrt{T}}$.

  También se puede utilizar el estadístico de contraste:

$$Q_H = T(T+2)\sum_{k=1}^{H} \frac{\hat{\rho}_k^2}{T-k},$$

  para un número H de autocovarianzas.

  Bajo la hipótesis nula de que la muestra proviene de variables aleatorias i.i.d. con varianza finita, y asumiendo que el tamaño muestral T es "grande", se tiene que:

$$Q_H \approx \chi_H^2$$

  Por tanto, rechazaremos la independencia (al 5 %) si el valor de Q_H es mayor o igual que el percentil $0,95$ de la distribución χ_H^2. Este contraste se conoce como contraste de Ljung-Box.

- **Contraste de media cero**

  Utilicemos ahora los símbolos $\overline{y}$ y s_y^2 para denotar a la media y a la varianza muestrales, respectivamente.

  Bajo la hipótesis nula de que la muestra $y_1, \ldots, y_T$ proviene de variables aleatorias i.i.d. con media cero y varianza finita, y asumiendo que el tamaño muestral T es "grande", se tiene que:

$$\frac{\overline{y}}{s_y/\sqrt{T}} \approx t_{T-1} \approx N(0,1)$$

  Por tanto, rechazaremos que la media μ_y es cero (al 5 %) si $|\overline{y}| \geq 1{,}96\frac{s_y}{\sqrt{T}}$.

- **Contraste de normalidad**

  Utilizando una notación diferente,

$$G_1 = \frac{\sum_{t=1}^{T}(y_t - \overline{y})^3}{T s_y^3} \quad \text{y} \quad G_2 = \frac{\sum_{t=1}^{T}(y_t - \overline{y})^4}{T s_y^4} - 3.$$

Bajo la hipótesis nula de que la muestra $y_1, \ldots, y_T$ proviene de variables aleatorias i.i.d. con distribución gaussiana, y asumiendo que el tamaño muestral T es "grande", se tiene que:

$$T\left(\frac{G_1^2}{6} + \frac{G_2^2}{24}\right) \approx \chi_2^2.$$

Rechazaremos la normalidad (al 5 %) si el valor del estadístico es mayor o igual que el percentil 0, 95 de la distribución χ_2^2. Este contraste se conoce como contraste de Jarque-Bera.

Si por otra parte denotamos,

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

donde $y_{(t)}$ denota al estadístico ordenado de orden t y las constantes $b_{t,T}$ vienen dadas a partir de la inversa de la distribución normal estándar.

El estadístico ω puede interpretarse como el cuadrado del coeficiente de correlación lineal de los puntos muestrales dibujados sobre papel probabilístico normal. Puesto que bajo la hipótesis nula de que $y_1, \ldots, y_T$ son i.i.d. con distribución gaussiana dicho grafico debería ser aproximadamente lineal, se rechaza la normalidad para valores pequeños de ω. Shapiro y Wilk tabularon los valores de $b_{t,T}$, y la distribución de ω bajo la hipótesis nula.

Los gráficos y contrastes de hipótesis ayudan en la verificación de si el modelo ARMA propuesto es o no adecuado como generador de la serie de tiempo (etapa de chequeo o diagnosis).

Concretamente, asesoran en la toma de la decisión referente a si las innovaciones a_t del modelo ARMA son o no ruido blanco con distribución gaussiana.

Puesto que las innovaciones $\{a_t\}_{t\in\mathbb{R}}$ no son observables, lo que se hace es estimarlas y realizar el chequeo sobre dichas estimaciones (esto es, sobre los residuos $\hat{a}_t$ del modelo estimado o ajustado).

Se puede destacar que los contrastes de independencia aplicados a los residuos $\hat{a}_t$ sufren las siguientes modificaciones (con respecto a su aplicación a las innovaciones a_t):

- Contraste basado en la distribución de cada autocorrelación muestral $\hat{\rho}_k$:
  La varianza asintótica de $\hat{\rho}_k$ para retardos k "pequeños", deja de ser $\frac{1}{T}$, tiene un valor menor.

- Contraste de Ljung-Box:
  Los grados de libertad de la distribución asintótica de Q_H definido anteriormente pasan a ser $H-p-q-1$ o $H-p-q$, en función de que el ARMA tenga o no constante respectivamente. Esto quiere decir que se necesita que $H > p+q+1$ o $H > p+q$, respectivamente.

Para las dos modificaciones anteriores, la región de rechazo resulta modificada.

Notar que debido a la relación existente entre los procesos ARIMA y los procesos ARMA:

$$\{X_t\}_{t\in\mathbb{R}} \quad ARIMA(p,d,q) \Leftrightarrow (1-B)^d X_t \;\; es \;\; ARMA(p,q)$$

se tiene que para estimar y realizar diagnosis de un modelo ARIMA es suficiente:

- Identificar el orden de diferenciación regular, d.

- Saber estimar y realizar la diagnosis de un modelo ARMA.

2.6. Selección del modelo

En las secciones anteriores, se han presentado las etapas necesarias para proponer un modelo ARMA como posible generador de una serie de tiempo, que de forma resumida serían:

1. Asesorarnos (gráficamente) acerca de la "estacionariedad de la serie". Si es estacionaria, pasar a la etapa 2.
2. Identificar los órdenes p y q del ARMA: estudio de sus fas, fap y fase muestrales.
3. Estimar el modelo cuyos órdenes se identificaron en la etapa 2 mediante mínimos cuadrados (condicionados o no) o máxima verosimilitud.
4. Chequear el modelo estimado: análisis de residuos.

Si los residuos pueden ser considerados como procedentes de un proceso de ruido blanco (preferiblemente gaussiano), el modelo estimado es propuesto como posible generador de la serie de tiempo.

Como se acaba de recordar, la etapa correspondiente a la identificación de los órdenes p y q del ARMA, se basa en el estudio de las fas, fap y fase muestrales. Sin embargo:

- Las fas, fap y fase muestrales pueden identificar varios procesos como posibles generadores de la serie de tiempo; por eso, es necesario establecer criterios que nos permitan determinar, entre varios modelos, cuál es preferible.
- Es posible que haya algún proceso distinto de los identificados que sea preferible a ellos; también convendría disponer de algún método que sugiera modelos de forma automática.

Esta sección se centrará principalmente en el estudio de los métodos necesarios para poder actuar con criterios a la hora de elegir cúal es el mejor modelo.

Sean k la cantidad de coeficientes de un modelo ARMA(p,q) (esto es, $k = p + q + 1$ o $k = p + q$ si el ARMA tiene o no constante, respectivamente) y $\hat{\varphi}_{(p,q)}$ el vector formado por las estimaciones máximo verosímiles de dichos k coeficientes y de σ_a^2.

Se propone seleccionar aquel modelo ARMA que minimice el valor de una de las funciones siguientes:

- Criterio de Información de Akaike: $AIC = -2log\left(L(\hat{\varphi}_{(p,q)})\right) + 2k$.
- Criterio de Información de Akaike corregido: $AICC = -2log\left(L(\hat{\varphi}_{(p,q)})\right) + 2(kT+k+2)/(T-k-2)$.
- Criterio de Información Bayesiano: $BIC = -2log\left(L(\hat{\varphi}_{(p,q)})\right) + klog(T)$.

La estructura de los tres criterios anteriores, AIC, AICC y BIC, es similar:

- Primer sumando:
  Mide tanto la calidad del ajuste como la credibilidad que le da a la serie de tiempo. Cuanto menor es su valor mejor es el ajuste y mayor la credibilidad que le da a la serie. Su valor disminuye al aumentar el valor de p y/o q (que a su vez implica aumentar k).
- Segundo sumando:
  Penaliza el aumento en la cantidad de coeficientes del ARMA. Su valor disminuye al disminuir k.

El modelo que minimiza a una de estas 3 funciones consigue un equilibrio entre ambos sumandos (ambos serán "pequeños"); esto es, un buen ajuste sin demasiados parámetros (que darían problemas tanto a la hora de estimar como de predecir).

La comparación entre los criterios se centra en:

- El criterio BIC es consistente si realmente la serie ha sido generada por un ARMA. El BIC selecciona los órdenes correctos con probabilidad 1 (esto no ocurre con los criterios AIC y AICC).
- Los criterios AIC y AICC son asintóticamente eficientes si realmente la serie ha sido generada por un AR (posiblemente de orden ∞). El AIC y el AICC seleccionan el modelo que da lugar al menor error de predicción esperado (esto no ocurre con el criterio BIC).

Como consecuencia de la relación entre los procesos ARIMA y los procesos ARMA:

$$\{X_t\}_t \quad ARIMA(p,d,q) \Leftrightarrow (1-B)^d X_t \;\; es \;\; ARMA(p,q)$$

se tiene que para seleccionar un modelo ARIMA es suficiente:

- Identificar el orden de diferenciación regular, d.
- Saber seleccionar un modelo ARMA.

2.7. Predicción

Supongamos que la serie de tiempo $x_1, \ldots, x_T$ ha sido generada por un proceso ARMA $\{X_t\}$ cuyos parámetros son conocidos.

El objetivo de esta sección es predecir, a partir de la serie de tiempo observada, el valor futuro del proceso dentro de k instantes de tiempo. Por tanto, predecir el valor de X_{T+k} utilizando los parámetros estimados como si fueran verdaderos. Si el modelo es el correcto, estos parámetros minimizan el error de predicción a cualquier horizonte, de lo contrario no sería necesariamente real.

Dicha predicción se denomina predicción con origen en T y horizonte k, y la denotaremos por $\hat{x}_T(k)$.

- Si suponemos que la serie temporal $x_1, \ldots, x_T$ ha sido generada mediante un proceso AR(1), entonces:

  $$X_t = c + \phi_1 X_{t-1} + a_t$$

  y la predicción en origen T a horizonte 1 toma la forma

  $$X_{T+1} = c + \phi_1 X_T + a_{T+1}$$

  donde el valor de X_T es conocido y los valores de c, ϕ_1 y a_{T+1} son desconocidos y serán sustituidos por sus estimaciones.

  Destacar que la predicción de a_{T+1} a partir de la serie temporal es su media, $\mathbb{E}(a_{T+1}) = 0$, ya que no se contiene información sobre a_{T+1}. Por tanto,

  $$\hat{x}_T(1) = \hat{c} + \hat{\phi}_1 x_T$$

  Si queremos predecir en origen T a horizonte 2, se tiene:

  $$X_{T+2} = c + \phi_1 X_{T+1} + a_{T+2}$$

  donde los valores de c, ϕ_1 y a_{T+2} y X_{T+1} son desconocidos. Los tres primeros serán sustituidos por sus estimaciones, y el valor de X_{T+1} será sustituido por su predicción $\hat{x}_T(1)$.

  De nuevo la predicción de a_{T+2} a partir de la serie temporal es su media, $\mathbb{E}(a_{T+2}) = 0$, ya que no se contiene información sobre a_{T+2}. Entonces, de la misma forma,

  $$\hat{x}_T(2) = \hat{c} + \hat{\phi}_1 \hat{x}_T(1) = \hat{c} + \hat{\phi}_1(\hat{c} + \hat{\phi}_1 x_T) = \hat{c}(1 + \hat{\phi}_1) + \hat{\phi}_1^2 x_T$$

- Si suponemos que la serie temporal $x_1, \ldots, x_T$ ha sido generada mediante un proceso MA(1), entonces:

$$X_t = c + a_t + \theta_1 a_{t-1}$$

y queremos predecir en origen T a horizonte 1. Luego,

$$X_{T+1} = c + a_{T+1} + \theta_1 a_T$$

donde los valores de c, θ_1, a_{T+1} y a_T son desconocidos y serán sustituidos por sus estimaciones.

Destacar que la predicción de a_{T+1} a partir de la serie temporal es su media, $\mathbb{E}(a_{T+1}) = 0$, ya que no se contiene información sobre a_{T+1} y la predicción de a_T no es inmediata, pues sí se contiene información, la denotaremos por $\hat{a}_T(0)$.

Por tanto,

$$\hat{x}_T(1) = \hat{c} + \hat{\theta}_1 \hat{a}_T(0)$$

Si queremos predecir en origen T a horizonte 2, se tiene:

$$X_{T+2} = c + a_{T+2} + \theta_1 a_{T+1}$$

Los valores de c, θ_1, a_{T+2} y a_{T+1} son desconocidos y serán sustituidos por sus estimaciones. La predicción de a_{T+2} y a_{T+1} a partir de la serie temporal son sus medias respectivamente, $\mathbb{E}(a_{T+2}) = 0$ y $\mathbb{E}(a_{T+1}) = 0$ ya que no se contiene información sobre a_{T+2} y a_{T+1}.

Luego,

$$\hat{x}_T(2) = \hat{c}$$

El procedimiento diseñado para predecir valores de un AR(1) o un MA(1) a horizonte k=1 y k=2, se puede generalizar fácilmente para cualquier horizonte $k > 0$ y cualquier proceso de medias móviles de orden finito $p > 0$ o $q > 0$ respectivamente.

Además, puede demostrarse que la predicción a largo plazo de futuros valores de un proceso AR(p) es la media del proceso, es decir:

$$\hat{x}_T(k) \to \mu \;\; si \;\; k \to \infty.$$

De la misma forma, la predicción a horizontes mayores que q de futuros valores de un proceso MA(q) es la media del proceso:

$$\hat{x}_T(k) = \mu \;\; si \;\; k > q.$$

En resumen, los procedimientos utilizados para predecir valores futuros de procesos AR(p) y MA(q) pueden ser combinados fácilmente para predecir valores futuros de procesos ARMA(p,q).

Puede demostrarse que la predicción a largo plazo de futuros valores de un proceso ARMA(p,q) es la media del proceso; esto es,

$$\hat{x}_T(k) \to \mu \;\; si \;\; k \to \infty.$$

Ahora bien, también se puede suponer que la serie de tiempo ha sido generada por un proceso ARIMA, en este caso la predicción de valores futuros de procesos ARIMA se basa en la predicción de los valores futuros de los procesos ARMA asociados.

Los pasos a seguir en este caso serían:

1. Diferenciar la serie procedente del ARIMA hasta obtener una serie procedente de un ARMA.
2. Predecir los valores futuros del proceso ARMA.
3. Deshacer la diferenciación en las predicciones del ARMA, obteniendo entonces las predicciones del proceso original ARIMA.

2.7.1. Intervalos de predicción

Cuando los residuos pasen el contraste de normalidad se puede construir un intervalo de predicción.

A continuación, generaremos un intervalo de predicción utilizando la distribución muestral del error de predicción:

$$e_T(k) = X_{T+k} - \hat{X}_T(k).$$

Se tiene que, si T es "grande" y $\{a_t\}_t$ es gaussiano, entonces

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

donde ψ_i son los coeficientes de la representación

$$X_t = c + a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \ldots$$

Un intervalo de predicción (al 95 %) para el valor de X_{T+k} será:

$$\left(\hat{X}_T(k) \pm 1,96\sqrt{\sigma_a^2(1 + \psi_1^2 + \cdots + \psi_{k-1}^2)}\right)$$

Es importante notar que en el cálculo de la varianza asintótica del error de predicción,

$$Var\left(X_{T+k} - \hat{X}_T(k)\right) \approx \sigma_a^2\left(1 + \psi_1^2 + \cdots + \psi_{k-1}^2\right),$$

las estimaciones de los parámetros incluidos en la predicción $\hat{X}_T(k)$ fueron tratadas como fijas (esto es, como si no dependiesen de los valores del proceso estocástico $\{X_t\}_{t\in\mathbb{R}}$).

Sin embargo, puesto que dichas estimaciones dependen de la serie $x_1, \ldots, x_T$, también dependen de los valores de $\{X_t\}_{t\in\mathbb{R}}$, provocando un cambio en la varianza del error de predicción, y por tanto en los intervalos de predicción. Este cambio se puede considerar despreciable si el tamaño de la serie es "grande".

Capítulo 3

Modelos de regresión

La línea de la exposición intenta centrarse en enfatizar las principales ventajas y desventajas de cada uno de los modelos propuestos. Se pretende mostrar cómo la rigidez de las hipótesis de los modelos más sencillos puede aliviarse mediante formulaciones más complejas que otorgan mayor flexibilidad para explicar adecuadamente la conducta de la formación bruta de capital, variable respuesta objeto de estudio en este trabajo.

3.1. Modelo de regresión lineal

Siguiendo las palabras de Hastie, Trevor and Tibshirani, Robert and Friedman, Jerome (2009, Cap.3):

"*partiremos de los modelos más simples para poder llegar a comprender el funcionamiento de modelos más complejos*".

En nuestro caso empezaremos explicando los conceptos básicos de los modelos lineales para luego ir avanzando en modelos más flexibles y a su vez más complejos.

Los modelos de regresión lineal asumen que la función de regresión $\mathbb{E}(Y \mid X)$ es lineal para los valores $(X_1, \ldots, X_p)$, de modo que son fácilmente interprables y, pese a su sencillez, resultan útiles a la hora de analizar cómo las variables explicativas $X' = (X_1, \ldots, X_p)$ afectan a la variable respuesta o explicada Y.

3.1.1. Modelo de regresión lineal múltiple

Como es bien conocido la función de regresión modeliza el comportamiento del valor esperado de una variable respuesta Y condicionado a los valores de una variable explicativa X. Si X consiste de un vector de p variables explicativas $X' = (X_1, \ldots, X_p)$ y se asume linealidad se habla entonces de un modelo de regresión lineal múltiple que se formula como

$$m(X') = m(X_1, \ldots, X_p) = \mathbb{E}(Y \mid X_1, \ldots, X_p) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j, \tag{3.1}$$

siendo $\beta' = (\beta_0, \beta_1, \ldots, \beta_p)$ el conjunto de parámetros desconocidos.

De este modo, la variable Y se puede explicar mediante la Ecuación (3.1) más un término de error aleatorio ε, y equivalentemente:

$$Y = \mathbb{E}(Y \mid X_1, \ldots, X_p) + \varepsilon = \beta_0 + \sum_{j=1}^{p} X_j \beta_j + \varepsilon, \tag{3.2}$$

donde habitualmente se asume que ε es un variable aleatoria gaussiana con media 0 y varianza σ^2 constante, $\varepsilon \sim N(0, \sigma^2)$.

Estimación de los parámetros del modelo

La estimación del vector de parámetros β se realiza sobre la base de un conjunto de datos observados $(X_{1,1},\ldots,X_{1,p},Y_1),\ldots,(X_{N,1},\ldots,X_{N,p},Y_N)$, al que denominaremos *"training data"*[1]. El método de estimación más común para estimar $\beta=(\beta_0,\beta_1,\ldots,\beta_p)'$ es la estimación por **mínimos cuadrados**, consistente en minimizar la suma de los cuadrados de los residuos (RSS, por sus siglas en ingles *Residuals Sum Squares*). La función objetivo a minimizar es:

$$RSS(\beta)=\sum_{i=1}^{N}(y_i-m(x_i))^2=\sum_{i=1}^{N}\left(Y_i-\beta_0-\sum_{j=1}^{p}X_{ij}\beta_j\right)^2 \tag{3.3}$$

La Ecuación (3.2) puede reescribirse en formato matricial. Sea $\mathbf{X}$ la matriz de dimensión $N\times(p+1)$ cuya i-ésima fila, $i=1,\ldots,N$, se corresponde con i-ésima observación del *"training data"*, la primera columna está formada por 1s, y la j-ésima columna, $j=2,\ldots,p+1$, recoge los valores la variable explicativa $j-1$. Denótese por $\mathbf{Y}$ al vector de observaciones de la respuesta, por β al vector de parámetros y por **epsilon** al vector de errores aleatorios. Entonces, la Ecuación (3.2) puede escribirse como $\mathbf{Y}=\mathbf{X}\beta+\mathbf{epsilon}$, y equivalentemente

$$\begin{pmatrix}Y_1\\Y_2\\\vdots\\Y_N\end{pmatrix}=\begin{pmatrix}1&X_{1,1}&X_{1,2}&\ldots&X_{1,p}\\1&X_{2,1}&X_{2,2}&\ldots&X_{2,p}\\\vdots&\vdots&\vdots&\ddots&\vdots\\1&X_{N,1}&X_{N,2}&\ldots&X_{N,p}\end{pmatrix}\begin{pmatrix}\beta_0\\\beta_1\\\vdots\\\beta_p\end{pmatrix}+\begin{pmatrix}\varepsilon_1\\\varepsilon_2\\\vdots\\\varepsilon_N\end{pmatrix} \tag{3.4}$$

Del mismo modo, reescribiremos la Ecuación (3.3) de manera matricial y su expresión se corresponde con:

$$RSS(\beta)=(\mathbf{Y}-\mathbf{X}\beta)'(\mathbf{Y}-\mathbf{X}\beta) \tag{3.5}$$

Diferenciando la Ecuación (3.5) respecto a β, e igualando a cero su primera derivada obtendremos la expresión del estimador mínimo cuadrático de los parámetros del modelo:

$$\widehat{\beta}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{3.6}$$

Asumiendo que las observaciones de Y_i, con $i=1,\ldots,N$, no están correlacionadas y que tienen una varianza constante σ^2, la distribución en el muestreo de $\widehat{\beta}$ presenta una varianza tal que:

$$var(\widehat{\beta})=(\mathbf{X}'\mathbf{X})^{-1}\sigma^2$$

Hipótesis del modelo y diagnosis

Para poder realizar inferencia sobre los parámetros de la Ecuación (3.2), se deben cumplir las siguientes hipótesis:

- **Linealidad**: Hipótesis estructural toda vez que se ha asumido que $\mathbb{E}(\mathbf{Y}\mid\mathbf{X})$ es lineal.
- **Homocedasticidad**: La varianza de los residuos del modelo es constante $var(\widehat{\varepsilon}\sim(X_1,\ldots,X_p))=\sigma^2$.
- **Normalidad**: Los residuos siguen una distribución normal de modo que $\widehat{\varepsilon}\sim N(0,\sigma^2)$.
- **Independencia**: Los residuos del modelo $\widehat{\varepsilon}_1,\ldots,\widehat{\varepsilon}_N$ deben ser independientes.

Las hipótesis enumeradas anteriormente se chequearán de manera gráfica y con pruebas de hipótesis específicas ver Ruppert, David and Wand, M.P. and Carroll, Raymond J. (2009, pág: 21).

[1] En nuestro caso coincidirá con el período temporal en el que se encuentran las series de variables internas de ABANCA.

Bondad de ajuste

Si tenemos varios modelos candidatos, podremos recurrir al coeficiente de determinación ajustado para saber que modelo es mejor. La expresión del coeficiente de determinación ajustado se corresponde con:

$$R^2_{ajust} = 1 - \frac{RSS/(N-p)}{TSS/(N-1)} \tag{3.7}$$

donde RSS se corresponde con la Ecuación (3.5), $TSS = \sum_{i=1}^{N}(y_i - \bar{y}_i)^2$, N se corresponde con el tamaño muestral, y p con el número de parámetros del modelo.

El coeficiente de determinación ajustado es una medida de bondad de ajuste que toma valores en el intervalo $[0, 1]$. De manera que cuanto más próximo este a 1 mejor será la calidad del ajuste, y por tanto, el mejor modelo será aquel que presente un R^2_{ajust} mayor.

Esta sección se ha construido siguiendo el Capítulo 3 del libro "*The elements of statistical learning*" de Ruppert, David (2004), y el capítulo 2 del libro "*Semiparametric regression*" de Ruppert, David and Wand, M.P. and Carroll, Raymond J. (2009).

3.2. Modelo de regresión lineal dinámica

Un modelo de regresión lineal dinámica (DLM, por sus siglas en inglés *Dynamic linear models*) es una representación de la relación existente entre dos o más series temporales que permite introducir dependencia temporal entre las series ajustando un modelo Box-Jenkins a los errores del modelo (como los vistos en el Capítulo 2). En esta sección se analizarán modelos para relacionar linealmente series temporales estacionarias mediante modelos de regresión dinámica. Estos modelos, además de tener en cuenta la relación dinámica entre las variables del modelo, resultan de interés ya que si las variables explicativas son controlables, permiten simular y evaluar políticas alternativas, siendo de gran utilidad en las series macroeconómicas.

3.2.1. Comparación con los modelos de regresión lineal estándar

Para un modelo de regresión lineal simple, caso particular del modelo de regresión lineal múltiple visto en la Sección 3.1, se supone que las variables endógena $\{Y_t\}_{t\in\mathbb{R}}$ y exógena $\{X_t\}_{t\in\mathbb{R}}$ del modelo siguen ambas un proceso de ruido blanco y se relacionan conforme al siguiente modelo:

$$Y_t = \beta_0 + \beta_1 X_t + u_t \tag{3.8}$$

donde u_t son los errores o innovaciones del modelo, y siguen un proceso de ruido blanco.

Si intentamos aplicar este modelo a unas variables que no siguen un proceso de ruido blanco, como es el caso de la mayoría de series temporales, con alta probabilidad estaremos asumiendo erróneamente algunos hipótesis de trabajo. Específicamente:

1. Asumir que la relación entre las variables es instantánea, cuando entre variables dinámicas la relación puede ser más compleja y transmitirse con ciertos retardos, siendo el efecto dinámico más complejo cuantos más intervalos de tiempo sean necesarios para la transmisión del efecto.

   Por ejemplo, si trabajamos con datos trimestrales, una subida en el precio de cierto producto en un mes del trimestre aparecerá enmascarado como un efecto instantáneo, ya que todos los efectos quedan incluidos en el período de observación de la serie.

2. Asumir que la dirección de la relación de causalidad se produce de la variable $\{X_t\}_{t\in\mathbb{R}}$ hacia la $\{Y_t\}_{t\in\mathbb{R}}$, es decir, no existe causalidad bidireccional[2].

[2]Los modelos de regresión dinámica más complejos permiten contemplar una causalidad bidireccional entre las variables endógenas y exógenas del modelo, pero nosotros supondremos que esta causalidad es conocida y va solamente en una dirección. Para más información ver Peña, Daniel (2010, Cap. 19).

3. Por último, pero no menos importante, los modelos de regresión lineal estándar (ver Sección 3.1) parten de la hipótesis de que los errores del modelo son independientes, con media cero y varianza constante. Es decir, presuponen que los errores del modelo siguen un proceso de ruido blanco.

Estas limitaciones presentes en los modelos de regresión estándar, son superadas por los modelos de regresión lineal dinámica como veremos a continuación.

3.2.2. Modelo de regresión dinámica para series estacionarias

Del mismo modo que para determinar el orden del proceso Box-Jenkins utilizábamos los coeficientes de correlación simple y parcial (ver Sección 2.3). Para determinar la relación de *dependencia lineal* entre dos series estacionarias utilizaremos las funciones de covarianzas y correlaciones cruzadas. Previamente es preciso introducir los conceptos de covarianza y correlación cruzadas.

Definición 3.2.1. *Dos procesos $\{Y_t\}_{t\in\mathbb{R}}$ y $\{X_t\}_{t\in\mathbb{R}}$ se dicen conjuntamente estacionarios si cada uno de ellos es estacionario y las covarianzas cruzadas solamente dependen del retardo k entre las mismas, con independencia del instante inicial considerado.*

Función de covarianzas cruzadas

Definición 3.2.2. *Dados dos procesos $\{Y_t\}_{t\in\mathbb{R}}$ y $\{X_t\}_{t\in\mathbb{R}}$ conjuntamente estacionarios, la función de covarianzas cruzadas asigna a cada retardo k el valor $\gamma_{X,Y}(k)$ dado por:*

$$\gamma_{X,Y}(k) = \gamma_{X,Y}(t, t+k) = \mathbb{E}\big[(X_t - \mu_x)(Y_{t+k} - \mu_y)\big].$$

La estimación de $\gamma_{X,Y}(k)$ para dos series temporales estacionarias $\{Y_t\}_{t\in\mathbb{R}}$ y $\{X_t\}_{t\in\mathbb{R}}$ es:

$$\widehat{\gamma}_{xy}(k) = \frac{1}{T}\sum_{t=1}^{T-k}(x_t - \bar{x})(y_{t+k} - \bar{y})$$

En cuanto a las principales características de la función de covarianzas cruzadas destacar:

- La función $\widehat{\gamma}_{xy}(k)$ no es simétrica respecto al retardo k de modo que:
  - Para $k > 0$ los coeficientes $\widehat{\gamma}_{xy}(k)$ representan cómo los valores de x_t influyen en los valores futuros de y_{t+k}.
  - Para $k < 0$ los coeficientes $\widehat{\gamma}_{xy}(k)$ representan cómo los valores de y_t influyen en los valores futuros de x_{t+k}.
- La principal deficiencia de esta medida es que mezcla los coeficientes de relación entre variables y la estructura de autocovarianzas de x_t, por lo que será complicada de interpretar.

Debido a las características mencionadas anteriormente, las covarianzas cruzadas no se utilizan para identificar el modelo.

Función de correlación cruzada

Definición 3.2.3. *Dados dos procesos $\{Y_t\}_{t\in\mathbb{R}}$ y $\{X_t\}_{t\in\mathbb{R}}$ conjuntamente estacionarios, la función de correlaciones cruzadas asigna a cada retardo k el valor $\rho_{X,Y}(k)$ dado por:*

$$\rho_{X,Y}(k) = \frac{\gamma_{xy}(k)}{\sigma_x\sigma_y}$$

La estimación de $\rho_{X,Y}(k)$ se corresponde con

$$r_{xy}(k) = \frac{\widehat{\gamma}_{xy}(k)}{S_x S_y}$$

donde $S_x = \widehat{\gamma}_x^{-1/2}(0)$ y $S_y = \widehat{\gamma}_y^{-1/2}(0)$.

La función de correlación cruzada, al ser la estandarización de la función de covarianzas cruzadas, presentará los mismos problemas para identificar los retardos que existen en la relación entre dos variables, es decir, mezcla los coeficientes de la relación con los de autocorrelación.

Detección de la dependencia lineal entre dos procesos estacionarios

Para comenzar destacaremos que si dos procesos $\{Y_t\}_{t\in\mathbb{R}}$ y $\{X_t\}_{t\in\mathbb{R}}$ son no estacionarios entonces es necesario el preblanqueo de los mismos para conseguir tal condición (ver Cryer, Jonathan D and Chan, Kung-Sik (2010, Sec. 11.4) o en Peña, Daniel (2010, Anexo 17.3)). De esta manera se evitan posibles problemáticas provocadas por la existencia de una regresión espuria entre las series. Una vez confirmado que las series siguen un proceso estacionario, deberemos calcular la función de correlación cruzada (FCC). A modo de ejemplo, la Figura 3.1 muestra un gráfico de la FCC muestral para un ejemplo concreto.

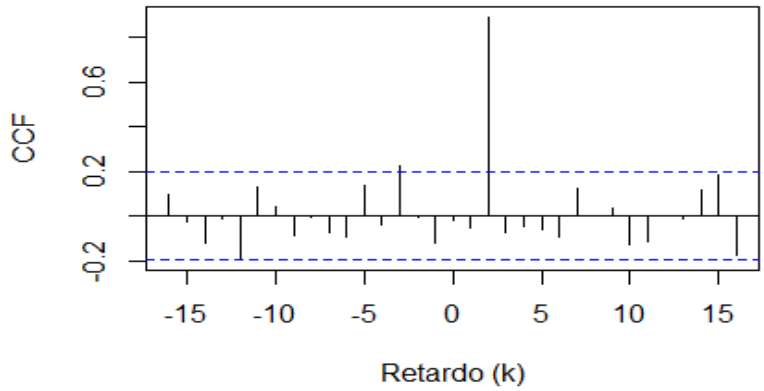


Figura 3.1: Función de correlación cruzada muestral (FCC).

En el ejemplo de la Figura 3.1, el retardo $k = 2$ es significativamente distinto de 0, de modo que la relación lineal entre los procesos estacionarios $\{Y_t\}_{t\in\mathbb{R}}$ y $\{X_t\}_{t\in\mathbb{R}}$ vendría determinada por cómo los valores de X_t influyen en los valores de $Y_{t+\mathbf{2}}$.

Con series de tiempo estacionarias, si la variables endógena y exógena del modelo mantienen algún tipo de dependencia lineal, entonces será factible detectarla mediante la función de correlación cruzada y, en general, el término error del modelo de regresión presentará autocorrelación. El modelo de regresión propuesto en el ejemplo coincidirá con:

$$Y_t = \beta_0 + \beta_1 X_{t-2} + Z_t,$$

donde Z_t se corresponderá con un proceso $ARIMA(p,d,q) \times (P,D,Q)_s$.

Diagnóstico del modelo

Consiste, como siempre, en comprobar si los residuos del modelo son una secuencia de variables aleatorias normales de media cero, con varianza constante e independientes. Por lo que debemos realizar los contrastes correspondientes, se pueden ver en la Subsección 3.1.1.

Procedimiento

Como se ha mencionado en capítulos anteriores, en la vida real las series temporales suelen ser series no estacionarias, por lo que el procedimiento sería el siguiente:

Paso 1 Elegir el retardo temporal basándose en el análisis de los coeficientes de correlación cruzada de las series previamente preblanquedas.

Paso 2 Una vez establecida la dependencia lineal entre las variables, ajustar un modelo de regresión lineal para las mismas suponiendo que los errores cumpliesen la hipótesis de incorrelación y observar los residuos resultantes de ajuste. Si son estacionarios ir al **Paso 3**, si no lo son volver al **Paso 1**.

Paso 3 Examinar los residuos para identificar un modelo $ARMA(p,q) \times (P,Q)_s$ (ver Capítulo 2).

Paso 4 Ajustar el modelo de manera conjunta mediante mínimos cuadrados o máxima verosimilitud.

Paso 5 Validación del modelo (ver Subsección 3.1.1). Si el modelo pasa los contrastes volver al **Paso 3** y ajustar otro modelo sobre los residuos, si siguen incumpliéndose las hipótesis volver al **Paso 1**.

Esta sección se ha construido siguiendo el capítulo 17 del libro "*Análisis de series temporales*" de Peña, Daniel (2010) y el capítulo 11 del libro "*Times Series Analysis with Applications in R*" de Cryer, Jonathan D and Chan, Kung-Sik (2010).

3.3. Modelo de regresión lineal generalizada

Los modelos de regresión lineal generalizada (GLM, por sus siglas en inglés *Generalized Linear Models*) son una denominación genérica que engloba otros métodos de regresión, como pueden ser, la regresión lineal simple, la regresión lineal múltiple, o la regresión logística entre otros.

En el modelo de regresión lineal múltiple expuesto en la Ecuación (3.1) se asume que una variable respuesta Y depende linealmente de p variables explicativas independientes $\mathbf{X}' = (X_1, \ldots, X_p)$. Específicamente, se asume que:

$$Y_i \mid \mathbf{X} = (X_1, \ldots, X_p) \sim N(\mu_i, \sigma) \qquad \text{con} \quad i = 1, \ldots, N, \tag{3.9}$$

y que la esperanza condicionada de la variable respuesta a los valores de las covariables del modelo μ_i satisface la relación:

$$\mu_i = \mathbb{E}[Y_i \mid \mathbf{X}_i] = \mathbf{X}_i'\beta \qquad \text{con} \quad i = 1, \ldots, N. \tag{3.10}$$

El modelo lineal generalizado también asume que las observaciones son condicionalmente independientes pero es mucho más versátil toda vez que no se impone: i) el comportamiento normal en distribución como en la Ecuación (3.9), ii) una relación lineal directa entre respuesta y explicativas, y iii) varianza constante. es factible por tanto emplear GLM para modelizar respuestas de tipo continuo, recuentos de observaciones, de tipo binario, etc.

En los GLM la distribución condicional de las Y_i se enmarca de manera más general en la familia de distribuciones de tipo exponencial que se definen a continuación.

Definición 3.3.1. *En un GLM la distribución de $Y_i \mid \mathbf{X}_i$ pertenece a la familia exponencial, y equivalentemente, su función de densidad admite la expresión*

$$f\big(Y_i \mid \theta_i, \xi\big) = exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\xi} + c(y_i, \xi)\right\},$$

donde θ_i se denomina parámetro natural, ξ es el parámetro de escala y, $b(\cdot)$ y $c(\cdot)$ son funciones conocidas que determinan el tipo de familia exponencial (ver Wood, Simon N. (2017, pág: 104)).

Como ya se ha mencionado, la esperanza condicional $\mu_i = \mathbb{E}[\mathbf{Y_i} \mid \mathbf{X}_i]$ no es necesariamente lineal. Denótese por $\eta_i = \mathbf{X}_i'\beta$ (a menudo referido como *predictor lineal*). Entonces la forma en la cual las covariables suministran información sobre la media μ_i se establece mediante una función *link* $g(\cdot)$, necesariamente monótona y diferenciable, que relaciona preditor lineal con media:

$$\eta_i = g(\mu_i) = \mathbf{X}_i'\beta \qquad \text{con} \quad i = 1, \ldots, N. \tag{3.11}$$

Denotando por $h(\cdot)$ a la inversa de la función link, $h = g^{-1}$, se tiene

$$\mu_i = h(\eta_i) = h(\mathbf{X}_i'\beta) \qquad \text{con} \quad i = 1, \ldots, N. \tag{3.12}$$

Por tanto, un GLM queda especificado una vez seleccionados el tipo de distribución de la familia exponencial para la distribución condicionada $Y_i \mid X_i$, la función link $g(\cdot)$ y el vector o matriz de diseño X_i.

Destacar que para cada familia exponencial existe una función link igual al parámetro natural θ_i con el predictor lineal $X_i'\beta$ de modo que:

$$\theta_i = w_1(\mu_i) = g(\mu_i) = \eta_i = \mathbf{X}_i'\beta \tag{3.13}$$

En el caso en que las $Y_i \mid X_i$ se distribuyan como normales $N(\mu_i, \sigma)$, $g(\mu) = \mu$ equivale a decir que la función link será igual a la identidad, por lo que nos encontraremos en el caso particular de la regresión lineal estándar vista en la Ecuación (3.1).

3.3.1. Estimación de los parámetros del modelo

La estimación de los parámetros del modelo lineal generalizado, así como los test de bondad del ajuste serán realizados siguiendo los métodos basados en máxima verosimilitud.

Para estos modelos, el sistema de ecuaciones de verosimilitud no va a tener habitualmente una solución analítica, por lo que deberá resolverse de forma numérica mediante un método iterativo. El más habitual es el método de mínimos cuadrados ponderados $IWLS$, por sus siglas en inglés *Iteratively Reweithed Least Squares*, también denominado método de las marcas de Fisher (*Fisher Scoring*). Otras alternativas son el método de Newton-Raphson o los métodos Cuasi-Newton (ver Wood, Simon N. (2017, pág: 107)).

El estimador de máxima verosimilitud $\widehat{\beta}$ obtenido mediante alguno de estos métodos anteriores tendrá una distribución asintótica normal multivariante tal que:

$$\widehat{\beta} \sim N(\beta, V)$$

donde V se corresponde con la matriz de covarianzas $(\mathbf{X'WX})^{-1}\xi$ (para más información ver Wood, Simon N. (2017, pág: 108), y Cabrero, Yolanda and García, Alfonso (2015, pág: 125)).

Es conveniente reflejar aquí que el estimador de máxima verosimilitud minimiza la *deviance*, un concepto de gran importancia para evaluar la calidad del ajuste y que se define como sigue.

Definición 3.3.2. *El estadístico deviance se obtiene mediante*

$$D = 2\{l_{max}(\widetilde{\beta}) - l(\widehat{\beta})\}\xi = \sum_{i=1}^{N} 2w_i\left\{y_i(\widetilde{\theta_i} - \widehat{\theta_i}) - b(\widetilde{\theta_i}) + b(\widehat{\theta_i})\right\},$$

donde $\widetilde{\beta}$ y $\widehat{\beta}$ son los estimadores máximo-verosímiles de β bajo el modelo saturado y el modelo ajustado respectivamente.

En definitiva, la deviance evalúa en qué medida la verosimilitud del modelo saturado (tantos parámetros como datos) supera a la verosimilitud del modelo propuesto. Si se ha realizado un buen ajuste se debería de obtener una deviance pequeña.

3.3.2. Hipótesis y diagnosis del modelo

Hipótesis modelo

Las hipótesis básicas del modelo GLM son:

- **Linealidad**: $g(\mathbb{E}[Y \mid X_1, \ldots, X_p]) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$.
- **Distribución**: La distribución condicional de las Y_i será de una familia de tipo exponencial con esperanza condicional $\mathbb{E}[Y_i \mid X_i] = \mu_i$.
- **Independencia**: Los residuos del modelo deben ser independientes.

En cuanto a las hipótesis de homocedasticidad e independencia de los residuos, en los GLM dependerá de la familia exponencial escogida para la variable respuesta/explicada, siendo necesario el cumplimiento de estas hipótesis cuando la familia exponencial escogida es la gaussiana.

Diagnosis del modelo

El análisis de los residuos o errores en los GLM se realiza a través de los *residuos de Pearson* y los *residuos deviance*.

Definición 3.3.3. *Residuos de Pearson:*

$$r_i^p = \frac{(y_i - \widehat{\mu}_i)^2}{\xi w_2(\widehat{\mu}_i)}$$

Definición 3.3.4. *Residuos deviance:*

$$r_i^d = signo(y_i - \widehat{\mu}_i)\sqrt{d_i}$$

donde d_i son los n sumandos del estadístico Deviance.

Los *residuos de Pearson* en el GLM deben de tener media cero y varianza ξ. En otros términos, no deben presentar tendencia, ni en media ni en varianza, cuando se representan frente a los valores ajustados o frente a alguna de las covariables del modelo. Dado que la distribución de los *residuos de Pearson* suele ser asimétrica, en la práctica se utilizan los *residuos deviance*, que se corresponden con los N sumandos del estadístico *Deviance*, (ver Definición 3.3.2).

Se admite que el modelo GLM es adecuado si los *residuos deviance* siguen una distribución normal tal que $r_i^d \sim N(0, \xi)$ (para más información ver Ruppert, David and Wand, M.P. and Carroll, Raymond J. (2009, pág: 113)). Si alguna de estas hipótesis no se cumpliese, deberíamos ajustar de nuevo el modelo. Pero, si detectamos que la independencia de los residuos del modelo, viene provocada por una dependencia temporal en los mismos, igual que en la Sección 3.2, ajustaremos un proceso $ARMA(p,q) \times (P,Q)_s$ sobre los mismos. De este modo conseguiremos modelizar la dependencia temporal existente entre la variable respuesta y las covariables, por lo que se podría decir que estamos ante un modelo GLM dinámico.

3.3.3. Criterios de selección y bondad del ajuste

El estadístico deviance (ver Definición 3.3.2) es una generalización de la suma de cuadrados de los residuos (RSS) en los modelos lineales. De hecho, si consideramos que $\xi = \sigma^2$, $b(\theta) = \frac{\theta^2}{2}$, $c(y,\xi) = -\frac{1}{2}\{\frac{y^2}{\xi} + log(2\pi\xi)\}$, y siguiendo la Ecuación (3.12) con $\theta = \mu = \eta$, obtenemos que:

$$D = \sum_{i=1}^{N}\left(2Y_i^2 - 2Y_i\widehat{\eta}_i - Y_i^2 + \widehat{\eta}_i^2\right) = \sum_{i=1}^{N}(Y_i - \widehat{\eta}_i)^2 = RSS(\widehat{\beta}) \tag{3.14}$$

Se denomina *Null Deviance* al estadístico deviance asociado a un modelo ajustado con solamente el intercepto (β_0) como única variable explicativa.

Definición 3.3.5. *Null Deviance*

$$D_0 = \sum_{i=1}^{N}(y_i - \widehat{\eta}_i)^2 = \sum_{i=1}^{N}(Y_i - \widehat{\beta}_0)^2 = SST$$

De la misma manera que el estadístico *Deviance* generaliza el estadístico RSS, el estadístico *Null Deviance* es la generalización de la suma de cuadrados totales (SST, por sus siglas en inglés *Total Sum of Squares*) del modelo lineal. Por consiguiente, *Null Deviance* y *Deviance* pueden ser usados conjuntamente para cuantificar el porcentaje de desviación explicada por el modelo ajustado. Para ello, el estadístico utilizado se corresponde con el R^2 resultante de la generalización del coeficiente de determinación de los modelos de regresión lineal estándar, ver Ecuación (3.7).

Su expresión es la siguiente:

$$R^2 = 1 - \frac{D}{D_0} \overset{LM}{=} 1 - \frac{SSE}{SST} \tag{3.15}$$

Destacar que, en los GLM la interpretación del estadístico R^2 es diferente a la interpretación en los LM. Es decir, no será el porcentaje de varianza explicada por el modelo, sino que indica lo bueno que es el modelo cuanto más se acerque el R^2 a la unidad, y por tanto, más próximo estuviesen los estimadores de máxima verosimulitud $l(\widehat{\beta})$ a los del modelo saturado $l(\widehat{\beta})_{max}$.

Además del criterio de selección R^2, también podremos utilizar los criterios de selección vistos en la Sección 2.6 adaptándolos según fuese necesario (para más información ver pag: 110 Wood, Simon N. (2017)).

Esta sección se ha construido siguiendo lo expuesto en Capítulo 3 del libro "*Generalized Additive Models*" de Wood, Simon N. (2017), el Capítulo 7 del libro "*Análisis estadístico de datos espaciales con QGIS y R*" de Cabrero, Yolanda and García, Alfonso (2015), y por último, el Capítulo 10 del libro "Semiparametric regression" de Ruppert, David and Wand, M.P. and Carroll, Raymond J. (2009).

3.4. Modelo de regresión aditiva generalizada

Hasta ahora hemos visto diferentes formas de abordar la rigidez de los modelos más simples, desde el modelo de regresión lineal múltiple de la Sección 3.1, hasta los modelos lineales generalizados de la Sección 3.3. Si queremos otorgar más flexibilidad, deberemos reemplazar el predictor lineal de los modelos (GLM) por un predictor aditivo, de esta forma nos encontraríamos ante un modelo (GAM), por sus siglas en inglés *Generalized Additive Model*. Siguiendo las palabras de Yee, Thomas W. (2015):

"*Los modelos GAM son la extensión no paramétrica de los modelos GLM*".

Destacar que de la misma manera que los GLM eran la generalización de los modelos de regresión lineal, los GAM son la generalización de los modelos de regresión aditivos, y a su vez de los modelos GLM, ya que los GLM son equivalentes a los GAM en el caso en el que las funciones η_i sean todas iguales a la función identidad.

A continuación, en esta sección se introducirán brevemente los modelos de regresión aditiva, para posteriormente explicar los modelos de regresión aditivos generalizados.

3.4.1. Modelos de regresión aditivos

Los modelos de regresión aditivos extienden los modelos de regresión lineal múltiple manteniendo el carácter aditivo de los efectos marginales de cada covariable X_i pero evaluando estos efecto de manera flexible, sin restringirlos a ser lineales. Es decir, los términos $\beta_i X_i$ se reemplazan por funciones no paramétricas y suaves $h(X_i)$ que proporcionan versatilidad para describir como influye marginalmente cada covariable en la respuesta. En consecuencia, la esperanza condicionada de la variable respuesta en un modelo aditivo sería la siguiente:

$$\mathbb{E}[Y \mid X] = h_0 + h_1(X_1) + \cdots + h_p(X_p)$$

siendo la expresión del modelo de regresión aditivo para una variable respuesta Y_i un conjunto de covariables X_j con $j = 1, \ldots, p$:

$$Y_i = \beta_0 + h_1(X_{1i}) + h_2(X_{2i}) + \cdots + h_p(X_{pi}) + \varepsilon_i \tag{3.16}$$

donde β_0 es el intercepto, h_j son funciones suaves, y los errores del modelo $\varepsilon_i \sim N(0, \sigma^2)$.

Los modelos aditivos tienen que verificar todas las suposiciones que exigíamos a los modelos de regresión lineal vistas en la Sección 3.1, (para más información ver Ruppert, David and Wand, M.P. and Carroll, Raymond J. (2009, Cap. 8) y Wood, Simon N. (2017, Sec. 4.3)).

3.4.2. Modelos de regresión aditivos generalizados

Dado un par de observaciones (x_i, y_i), para las cuales, la distribución condicional de y_i dada una x_i sigue una familia exponencial, en los GLM podíamos estimar $h(x) = \mathbb{E}(\mathbf{Y} \mid \mathbf{X})$ en base a criterios paramétricos, siendo la función de densidad de dicha distribución condicional (ver Definición 3.3.1).

En los modelos GAM la función de densidad de la función de distribución condicional, se corresponde con la función de densidad de cierta familia exponencial vista en la Definición 3.3.1. Pero a diferencia que en los DLM, en los GAM nosotros asumimos que $\eta_i = \eta(x_i)$, en donde $\eta(\cdot)$ es una función suave, de modo que:

$$\eta = [\eta(x_1), \ldots, \eta(x_N)]'$$

De esta forma, los modelos GAM extienden la Ecuación (3.11) a:

$$g(\mu(x_i)) = \eta_i = \beta + h_1(x_{i1}) + \cdots + h_p(x_{ip}) \tag{3.17}$$

Por tanto, los GAM pueden ser vistos como la suma de funciones suaves de cada una de las covariables, evitando la rigidez de la linealidad de los modelos GLM.

Estimación

Según Hastie, Trevor and Tibshirani, Robert (1990, pág: 140):

"*la estimación de β y $f_1, \ldots, f_p$ se logra reemplazando la regresión lineal ponderada en la regresión de la variable dependiente estimada mediante un algoritmo apropiado para ajustar un modelo aditivo ponderado.*"

Por tanto, los modelos GAM serán ajustados mediante "la maximización de la probabilidad penalizada", la cuál en la práctica se conseguirá a través de mínimos cuadrados penalizados iterativos (PIRLS), por sus siglas en inglés "*Penalized iterative least squares*", (para más información ver Wood, Simon N. (2017, pág: 180)).

De modo que, asumiendo una función de enlace canónica, y un parámetro de escala $\xi = 1$ para simplificar la demostración. La estimación de la función de suavizado mediante splines sería:

$$\widehat{\mathbf{f}} = (b')^{-1}(\widehat{\eta})$$

donde

$$\widehat{\eta} = \underset{\eta(\cdot)}{argmax}\{\mathbf{Y}'\eta - 1'b(\eta)\} - \frac{1}{2}\lambda^3 \int_{-\infty}^{\infty} \eta''(x)^2 dx \tag{3.18}$$

Para más información sobre distintas funciones de suavizado ver Ruppert, David and Wand, M.P. and Carroll, Raymond J. (2009, Cap. 3) y Wood, Simon N. (2017, Sec. 4.2).

Por último, destacar que el modelo será válido si los residuos son normales, homocedásticos e independientes. Si alguna de estas hipótesis no se cumpliese, deberíamos ajustar de nuevo el modelo. Pero, si detectamos que la independencia de los residuos del modelo, viene provocada por una dependencia temporal en los mismos, igual que en la secciones anteriores, ajustaremos un proceso $ARMA(p,q) \times (P,Q)_s$ sobre los mismos. De este modo conseguiremos modelizar la dependencia temporal existente entre la variable respuesta y las covariables, por lo que se podría decir que estamos ante un modelo GAM dinámico, (ver Sección 3.2).

Esta sección se ha construido siguiendo lo expuesto en capítulo 3 del libro "*Generalized Additive Models*" de Wood, Simon N. (2017), el capítulo 10 del libro "Semiparametric regression" de Ruppert, David and Wand, M.P. and Carroll, Raymond J. (2009), el libro "*Generalized additive models*" de Hastie, Trevor and Tibshirani, Robert (1990), y el libro "Vector generalized Linear and Additive Models" de Yee, Thomas W. (2015).

Capítulo 4

Corrección de series temporales

La mayoría de series económicas coyunturales son utilizadas como una herramienta para analizar el ciclo económico, permitiendo de esta forma, la adopción de estrategias adecuadas por parte de los expertos de la entidad ABANCA. Sin embargo, si estas series se encuentran influenciadas por efectos estacionales y efectos de calendario impedirán entender de forma clara el fenómeno económico subyacente detrás de las mismas.

Por tanto, en el presente trabajo se procederá a la corrección de las series temporales utilizadas como variables explicativas del modelo, pero la corrección de dichas series no está libre de riesgos ya que:

- Las series corregidas de estacionalidad y efectos de calendario dependerán del método utilizado, así como del software utilizado en dicho proceso.
- Un ajuste inapropiado puede generar resultados erróneos y señales falsas.
- La posible presencia de estacionalidad residual o un exceso de suavizado puede afectar de manera negativa a las series.

En el documento del INE "*Estándar del INE para la corrección de efectos estacionales y efectos de calendario en las series coyunturales*" INE (2019), se recoge una serie de recomendaciones para la corrección de series temporales en base a los estándares de unificación de metodologías propuestos por Eurostat en 2009[1] "*ESS guidelines on seasonal adjustment*" Boxall, Mark and Brown, Gary and Buono, Dario and Elliot, Duncan (2015). Estos estándares de unificación surgieron ante la necesidad de armonización de los tratamientos estadísticos de las series por parte de los institutos nacionales de estadística de los países miembros de la Unión Europea. Destacar que las recomendaciones no se restringen únicamente al proceso de ajuste estacional, si no que cubren otros aspectos sobre el pretratamiento de las series, (para más información ver INE (2019)).

Con el fin de adaptarnos a dicha armonización, para la corrección de las series temporales utilizaremos el paquete *seasonal* Sax and Eddelbuettel (2018), ya que se trata de la implementación del software utilizado por el INE para corregir la mayoría de sus series. Dicho paquete se basa en la interfaz X-13ARIMA-SEATS, software de corrección de series temporales del "*US Census Bureau*" que nos permite: corregir la estacionalidad, los atípicos y los posibles efectos de calendario presentes en las series temporales.

A continuación, se explicarán brevemente aspectos teóricos del ajuste de las series que nos permitirán entender mejor como se llevan a cabo los ajustes de estacionalidad y efectos de calendario, así como la detección y eliminación de atípicos.

Partiendo de la definición del proceso ARIMA estacional multiplicativo vista en la Subsección 2.2.4, una extensión útil de los modelos ARIMA surge de modelar la esperanza de Y a lo largo del tiempo mediante una regresión lineal respecto de otras variables regresoras observables en los mismos instantes de tiempo que Y.

[1] Estas recomendaciones han sido actualizadas en 2015.

Para una serie temporal $\{Y_t\}_{t\in\mathbb{R}}$ el modelo de regresión propuesto coincide con:

$$Y_t = \sum_{i=1} \beta_i X_{it} + z_t \tag{4.1}$$

donde Y_t es la serie temporal endógena del modelo de regresión propuesto, X_{it} son las variables de regresión observadas simultáneamente con Y_t, los β_i se corresponden con los parámetros de regresión, y z_t son los errores de la regresión para los cuales asumimos que siguen un proceso ARIMA.

Al utilizar conjuntamente la Ecuación (2.1) y la Ecuación (4.1), podemos obtener el modelo general utilizado por el software X-13ARIMA-SEATS, el cual puede escribirse como:

$$\phi(B)\Phi(B^s)(1-B)^d(1-B^s)^D\left(y_t - \sum_{i=1} \beta_i x_{it}\right) = \theta(B)\Theta(B^s)a_t \tag{4.2}$$

Este modelo puede considerarse como una generalización del modelo ARIMA especificado en la Ecuación (2.1) para permitir una *función media temporal* de regresión, o como generalización del modelo de regresión para permitir que los errores z_t sigan el modelo ARIMA. De cualquier modo, hay que tener en cuenta que este modelo general establecido en la Ecuación (4.2) implica que, primero se restan los efectos de regresión y_t para obtener la serie z_t de media cero, luego se diferencia la serie z_t de error para conseguir una serie estacionaria, llamémosle w_t, y por último se asume que w_t sigue el modelo ARMA estacional, $\phi(B)\Phi(B^s)w_t = \theta(B)\Theta(B^s)a_t$.

Por tanto, podemos reescribir el modelo general como[2]:

$$(1-B)^d(1-B^s)^D y_t = \sum_{i=1} \beta_i (1-B)^d(1-B^s)^D X_{it} + w_t \tag{4.3}$$

donde w_t sigue un modelo ARMA estacional.

A continuación, se enumerarán brevemente algunas de las variables regresoras incluidas en el software X-13ARIMA-SEATS:

- **Tendencia ("Trend Constant"):**

  Si el modelo ARIMA descrito en la Ecuación (4.3) no precisa ser diferenciado ($d = D = 0$), entonces se introduce una constante (el intercepto de un modelo de regresión) que coincide con la media de la serie. Si contiene diferencias, entonces se introduce una variable regresora (denominada çonstante de tendencia") que, tras diferenciarse de acuerdo con d y D, produce una columna de unos. Por tanto, el parámetro que acompaña a esta regresora proporciona modeliza una tendencia polinómica del mismo grado que el número de diferencias del modelo ($d + D$).

  $$(1-B)^{-d}(1-B^s)^{-D} I(t \geq 1), \text{ donde } I(t \geq 1) = \begin{cases} 1 \text{ para } t \geq 1 \\ 0 \text{ para } t < 1 \end{cases}$$

- **Estacionalidad ("Fixed Seasonal"):**

  Los efectos estacionales fijos de una serie mensual pueden modelarse utilizando 12 variables "*dummies*" o indicadoras, una para cada mes. Sin embargo, la suma de estas variables es igual a la unidad en un modelo sin diferenciar, o cero en el modelo diferenciado provocando un problema de colinealidad. Para evitarlo se utiliza una reparametrización apropiada que utiliza en su lugar 11 contrastes para cada una de las variables "*dummies*".

  $$M_{1,t} = \begin{cases} 1 \text{ en Enero} \\ -1 \text{ en Diciembre} \\ 0 \text{ en otro caso} \end{cases} \quad , \ldots, M_{11,t} = \begin{cases} 1 \text{ en Noviembre} \\ -1 \text{ en Diciembre} \\ 0 \text{ en otro caso} \end{cases}$$

[2]Nótese que en el modelo especificado las variables x_{it} afectan a y_t contemporáneamente. En la función X-13ARIMA-SEATS se permiten retardos en dicha relación que han sido obviados para la simplificación de la explicación.

- **Día laboral ("Traiding Day"):**

  Los "*Traiding Day*" o efectos del número de días de cotización, se producen cuando una serie se ve afectada por las diferentes composiciones del día de la semana del mismo mes natural en diferentes años. Los efectos del día de cotización se pueden modelar mediante el conocido como "*One Coefficient Traiding Day*":

  (número de días de la semana laborables) - $\frac{5}{2}$(número de sábados y domingos)

- **Vacaciones ("Holiday"):**

  Los efectos de vacaciones recogen aquellas vacaciones cuyas fechas varían en el tiempo, y que por tanto, afectan a la actividad económica provocando variaciones en las series. Dentro de estos efectos, el efecto de la Semana Santa es el más recurrente ya que sus fechas varían entre el 22 de marzo y el 25 de abril. El efecto de semana se puede modelizar mediante:

  $E(w,t) = \frac{1}{w} \times$[número de días antes del domingo de pascua que caen en el mes t] con $t = 1, \ldots, 12$.

  donde w se correponde con el número de días hasta Semana Santa.

- **Longitud del mes ("Length of Month"):**

  La variable regresora que recoge los efectos sobre las series derivadas de la variación de días en los meses del año tal que:

  $$m_t - \bar{m}$$

  donde $m_t =$ es la longitud del mes en t días y $\bar{m} = 30.4375$ es la longitud media de días en el mes.

- **Año bisiesto ("Leap Year"):**

  Variable regresora que recoge los efectos sobre las series derivados de la variación de los días provocada por los años bisiestos tal que:

  $$LY_t = \begin{cases} 0.75 \text{ cuando año bisiesto y t = febrero} \\ -0.25 \text{ cuando año no bisiesto y t = febrero} \\ 0 \text{ si } t \neq \text{febrero} \end{cases}$$

Además de las variables regresoras expuestas anteriormente para la corrección de posibles efectos de estacionalidad y calendario, en las series también se han tenido en cuenta diferentes tipos de datos atípicos. La interfaz X-13-TRAMO-SEATS tiene implementada la detección automática de datos atípicos aditivos (AO), cambios temporales (TC), y cambios de nivel (LS), siendo las variables de regresión asociadas a los mismos:

- **Atípicos aditivos ("Additive Outlier"):**

  Efectos transitorios que provocan un cambio en algunos valores de la serie, pero en el resto de instantes temporales el nivel de la serie no se ve afectado.

  $$AO_t^{(to)} = \begin{cases} 1 \text{ para } t = t_0 \\ 0 \text{ para } t \neq t_0 \end{cases}$$

- **Cambios de nivel ("Level Shift"):**

  Efectos permanentes que provoca un cambio de nivel a partir de un instante conocido.

$$LS_t^{(to)} = \begin{cases} -1 \text{ para } t < t_0 \\ 0 \text{ para } t \geq t_0 \end{cases}$$

- **Cambio temporal ("Temporary Change"):**

  Efectos transitorios que provocan un cambio en algunos valores de la serie, pero a la larga el efecto sobre el nivel no se ve afectado.

$$TC_t^{(to)} = \begin{cases} 0 \text{ para } t < t_0 \\ \alpha^{(t-t_0)} \text{ para } t \geq t_0 \end{cases}$$

La detección de datos atípicos implementada en el software X-13ARIMA-SEATS se basa en el trabajo de *Chang y Tiao (1983)*, siendo el enfoque general similar al "Stepwise" de los modelos (GLM), donde las variables de regresión candidatas a entrar en el modelo se corresponden con AO, LS, y/o TC para cada instante temporal de la serie. En resumen, este enfoque implica calcular para cada tipo de dato atípico, en cada instante temporal, el estadístico de contraste t para la significación del mismo, agregando al modelo las variables AO, LS, y/o TC correspondientes, (para más información ver US Census Bureau (2017, Sec. 7.11)).

En definitiva, en este capítulo hemos visto como podemos incorporar a la especificación de un modelo ARIMA estacional multiplicativo variables regresoras para corregir de esta forma la componente estacional, atípicos y los efectos de calendario en las series. El mismo se ha realizado siguiendo el manual del INE "*Estándar del INE para la corrección de efectos estacionales y efectos de calendario en las series coyunturales*" INE (2019), el libro de Eurostat "*ESS guidelines on seasonal adjustment*" Boxall, Mark and Brown, Gary and Buono, Dario and Elliot, Duncan 2015 y manual "*X-13ARIMA-SEATS Reference Manual*" US Census Bureau (2017).

Parte II

Caso práctico

Capítulo 5

Preparación de los datos

En este capítulo se mostrarán las modificaciones y correcciones aplicadas sobre la variable endógena de los modelos y las correspondientes variables explicativas del mismo. En concreto, en la Sección 5.1 se mostrará el procedimiento realizado para la mensualización de la variable endógena del modelo, es decir, la formación bruta de capital. Además, en la Sección 5.2 veremos las correcciones de atípicos, estacionalidad y efectos de calendario efectuadas sobre la batería de variables internas de la entidad ABANCA, ya que una vez corregidas serán las variables explicativas de los modelos de regresión propuestos en las siguientes secciones.

Cabe destacar que el objetivo de este trabajo se centra en construir un indicador para la `FBC` publicado por el IGE, para el cual una de las formas de seguimiento es mediante tasas de variación interanual. Es común el uso de dichas tasas dado que trabajar en cifras absolutas no nos permite ver la dinámica en la evolución de las series, pero el uso de tasas de variación interanual, al comparar siempre los mismos instantes temporales, nos permite ver dicha evolución y solventa la posible estacionaldad presente en las series. Además dado que las variables candidatas son de muy diversa naturaleza el uso de tasas variación nos facilita homogeneizarlas. Es por ello que a la hora de intentar modelizar la `FBC` todas la variables serán utilizadas en tasas de variación interanual.

5.1. Mensualización de la FBC

La variable endógena en todos los modelos de regresión propuestos en este trabajo es la `FBC`, cuya frecuencia de publicación es trimestral, pero las variables explicativas tienen frecuencia mensual ya que han sido construidas mediante las variables internas de negocio que se actualizan a cierre de mes. Por tanto, con el fin de no perder información trimestralizando las variables explicativas, procederemos a mensualizar la `FBC`, ya que de esta forma no perderemos información y tendremos un mayor conocimiento sobre la situación de la inversión en la economía gallega, aportando un mayor dinamismo a al análisis.

La mensualización de la `FBC` la realizaremos a través del paquete *tempdisagg* (Christoph Sax and Peter Steiner (2016)). En concreto, utilizaremos la función `td()` que nos permite desagregar o interpolar una serie de baja frecuencia a una serie de tiempo de frecuencia mayor, teniendo en cuenta, que el promedio de la serie de alta frecuencia resultante es consistente con la serie de baja frecuencia. Para esta tarea la función `td()` tiene implementados varios procedimientos, en concreto en este trabajo se han utilizado 3 métodos implementados en dicha función: el método *"denton-cholette"*, el método *"chow-lin-maxlog"*[1], y el método *"chow-lin-minrss-ecotrim"*.

En nuestro caso, mendiante la función `td()` estableceremos que la serie de baja frecuencia (`FBC`) vendrá determinada por un modelo de regresión, en donde la variable explicativa será una variable con mayor frecuencia. Para la elección de esta variable, nos hemos basado en Eurostat (2014, pág: 224), en donde se indica qué variables macroeconómicas con frecuencia más alta están relacionadas con la `FBC`. Además, las variables seleccionadas se han elegido en función a las componentes de la `FBC`. (ver Figura 1.2b).

[1] Es el método predeterminado (y recomendado),ya que produce buenos resultados para una amplia gama de aplicaciones.

Las series candidatas para mensualizar la FBC en base a sus componentes son:

- **Viviendas:** Número de compraventa de viviendas total (CVT).
- **Otros edificios y construcciones:** Producción o venta de productos de hormigón (PVH).
- **Maquinaria y bienes de equipo:** El índice de producción industrial (IPI).

Pese a que bajo el punto de vista económico la serie de producción y venta de hormigón es una buena opción para llevar a cabo la mensualización de la FBC, en la práctica no es posible disponer de los datos actualizados de la misma. Por tanto, se ha decidido no tenerla en cuenta en este análisis.

Para determinar cuál de las variables tiene mayor relación con el comportamiento de la FBC, se ha procedido de la siguiente manera:

1. En un primer lugar, corregimos las series candidatas de atípicos, estacionalidad y calendario.
2. En segundo lugar, mediante la función ta()[2] transformamos la frecuencia de las series candidatas a trimestral.
3. En tercer lugar, ajustamos un modelo de regresión lineal, uno por cada serie candidata, y elegimos aquella con mayor coeficiente de determinación (R^2) como serie que nos permita realizar la mensualización.

Destacar que debido a la naturaleza de los datos macroeconómicos utilizados, estos presentan demasiada volatilidad incluso cuando se han corregido de los efectos de estacionalidad y calendario. Por ejemplo, a efectos ilustrativos, las estadísticas de compraventas de viviendas se obtienen a partir de los datos de los Registros de la Propiedad, con un desfase temporal entre la fecha de compra efectiva y la fecha de inscripción, siendo esta última la que se tiene en cuenta a la hora de recopilar la información. Este retraso es desigual entre los distintos Registros creando una distorsión adicional. Por tanto, para la mensualización de la FBC tendremos en cuenta también la versión suavizada de las variables.

Esta problemática puede verse reflejada en la Figura 5.1 donde se representan las tasas de variación interanual de la compraventa de viviendas totales y su versión suavizada. Puede verse claramente, en línea con lo expuesto, que la serie sin suavizar presenta una elevada volatilidad, resultando poco apropiada para mensualizar la FBC ya que estaríamos introduciendo demasiado ruido en la serie mensualizada de la FBC.

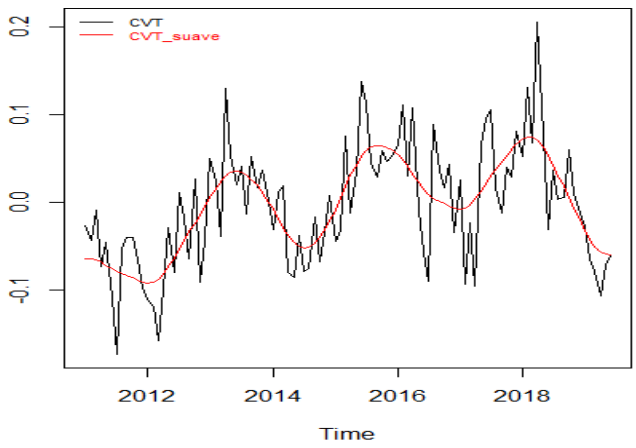


Figura 5.1: Tasas de variación interanual de la compraventa de viviendas total (CVT) y su versión suavizada (CVT_{suave}).

[2]Función implementada en el paquete *tempdisagg* (Christoph Sax and Peter Steiner (2016)), de funcionamiento similar a td(), pero en sentido inverso. Es decir, nos permite reducir la frecuencia de las series transformadas.

Siguiendo el procedimiento anterior, procedemos al ajuste de un modelo de regresión lineal, donde la variable explicada es la `FBC` y la variable explicativa, en cada caso, es una de las variables candidatas y su versión suavizada. Tras realizar el ajuste de cada uno de los modelos propuestos, obtenemos que para `CVT` el coeficiente de determinación es igual a 0.61, para la variable `IPI` obtenemos un $R^2 = 0.22$, y para la versión suave de las variables obtenemos respectivamente $R^2 = 0.76$ y $R^2 = 0.32$. Por tanto, la variable mensual con la que llevar acabo la mensualización de la `FBC` es la compra venta de viviendas suavizada $\texttt{CVT}_{suave}$.

A efectos ilustrativos se mostrará la función `td()` en un caso general con los parámetros a ajustar:

$$FBC_{mensualizada} \leftarrow td(FBC_{trimestral} \sim \textbf{\textit{variable}}, \textit{ to= "monthly", method= } \textbf{\textit{método}}\textit{, conversion= "average")}$$

Siguiendo la expresión anterior, en la Tabla 5.1 se muestran las diferentes mensualizaciones para la `FBC` resultado de las distintas combinaciones de variables y métodos[3].

Mensualización	Variable	Método	R^2_{ajust}
FBC_1	Constante	Denthon-Cholette	*
FBC_2	CVT	Denthon-Cholette	*
FBC_3	CVTs	Denthon-Cholette	*
FBC_4	CVT	Chow-lin-maxlog	0,211
FBC_5	CVTs	Chow-lin-maxlog	0,437
FBC_6	CVT	Chow-lin-minrss-ecotrim	0,254
FBC_7	CVTs	Chow-lin-minrss-ecotrim	0,642

Tabla 5.1: Resumen de las diferentes mensualizaciones de la `FBC` mediante la función `td()`.

En la Figura 5.2 se muestra el resultado de cada una de estas mensualizaciones. Como se puede observar, para el método 'Denthon-Cholette" tanto la mensualización realizada utilizando la CVT como la realizada mediante su versión suavizada presenta un comportamiento demasiado volátil. La que mejor ajuste proporciona en términos del coeficiente de determinación ajustado (R^2ajust) en la Tabla 5.1, y no presenta una volatilidad excesiva, es la mensualización derivada del método *"Chow-lin-minrss-ecotrim"* de la variable mensual de la CVT suavizada.

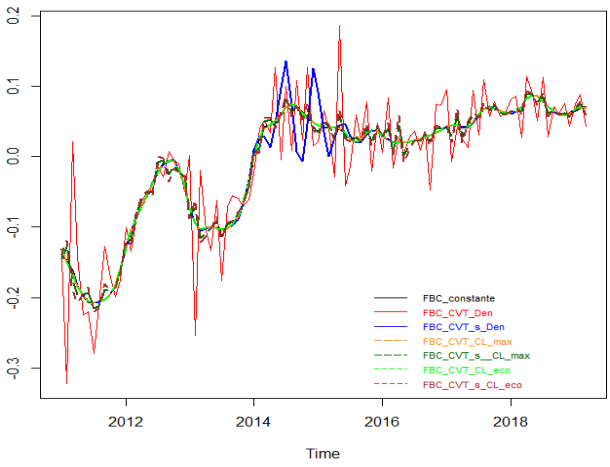


Figura 5.2: Tasas de variación interanual de la `FBC` mensualizada en sus diferentes versiones.

[3]Para el método "Denthon-Cholette" no están disponibles los valores del R^2_{ajust}.

Por tanto, la variable endógena con la que se trabajará será la mensualización de la formación bruta de capital mediante la suavización de la compraventa de viviendas bajo el método de *"Chow-lin-minrss-ecotrim"*.

5.2. Tratamiento variables explicativas

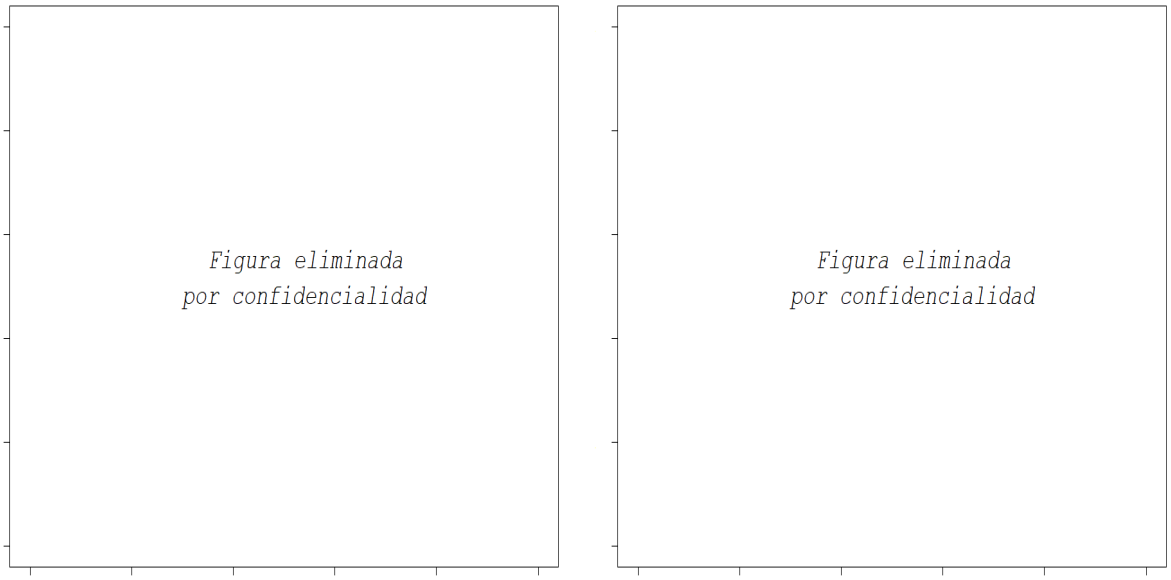
Antes de comenzar a modelizar procederemos a la corrección de las posibles variables explicativas ante la posible presencia de estacionalidad, datos atípicos y efectos de calendario. Esta corrección se realizará base a lo estudiado en el Capítulo 4 en el que se aborda la corrección de series temporales. Uno de los motivos por los que se realiza este análisis es para corregir el impacto que provocan las fusiones llevadas a cabo por la entidad, sobre las series candidatas a entrar en los modelos. Ya que cuando se produce una fusión, se intenta reconstruir la historia de las dos entidades en conjunto, y esto a veces no es posible para todas las series del banco, haciendo necesaria la corrección de esta sección. Además, las series también pueden verse afectadas por reprocesos de la información llevados a cabo en distintos instantes temporales.

Para llevar a cabo dichas correciones utilizaremos el paquete *seasonal* Sax and Eddelbuettel (2018) y el procedimiento a seguir sería el siguiente:

Paso 1 Corrección de los datos atípicos de las series.

Paso 2 Una vez corregidos los atípicos procedemos a corregir la estacionalidad y los posibles efectos de calendario de las series.

A continuación, se mostrarán algunos ejemplos de las series originales y una vez éstas han sido corregidas[4].



(a) Número de contratos para productos de circulante. (b) Número de efectos impagados en líneas de descuento.

Figura 5.3: Series corregidas de atípicos, estacionalidad y efectos de calendario.

Para la serie de **número de contratos de circulante** correspondiente a la Figura 5.3a, se puede observar cómo esta ha sido corregida de atípicos y estacionalidad. En concreto, la serie presentaba un atípico de nivel en agosto de 2011, atípico presente en la mayoría de las series debido a la fusión de las

[4]Los ejes de las figuras han sido eliminados por cuestiones de confidencialidad.

entidades bancarias Caixa Galicia y Caixanova. Además, la serie de número de contratos de circulante presentaba 2 atípicos aditivos en diciembre de 2015 y febrero de 2017, los cuales se correspondían con puntos que no son considerados como generados por el proceso al que corresponde dicha serie temporal. En cuanto a la estacionalidad, ha sido corregida mediante un modelo ARIMA multiplicativo estacional $ARIMA(1,1,2) \times (1,1,0)_{12}$.

Para la serie correspondiente al **número de efectos impagados en líneas de descuento**, se puede observar en la Figura 5.3b cómo la serie original se ve afectada también por la fusión bancaria presentando un cambio de nivel en agosto de 2011. En cuanto a la estacionalidad, ha sido corregida mediante un modelo ARIMA multiplicativo estacional $ARIMA(0,1,1) \times (0,1,1)_s$.

En la Figura 5.4 se observa el **número total de nóminas** por mes para empresas que tengan contratado el servicio de pagos nóminas con ABANCA. Se puede ver claramente que la serie original presenta una componente estacional representada por picos recurrentes en los meses de junio y diciembre para todo el período temporal correspondientes a las pagas extraordinarias realizadas en dichos meses. La componte estacional ha sido corregida mediante un modelo ARIMA multiplicativo estacional $ARIMA(0,1,1) \times (0,1,1)_{12}$.

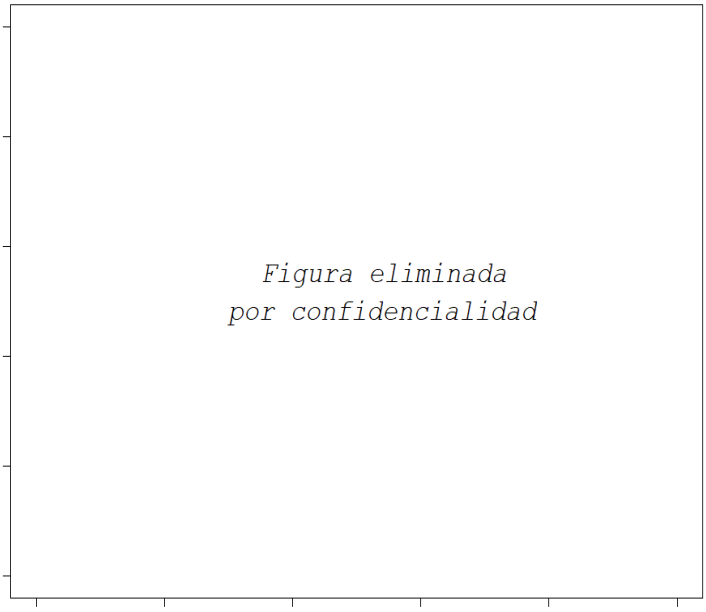


Figura 5.4: Número de nóminas pagadas por empresas mediante el servicio proporcionado por ABANCA.

Esta serie presenta la peculiaridad de tener 2 cambios de nivel, el primero provocado por la fusión con Caixanova en agosto de 2011, y el segundo provocado por la fusión con el banco Etcheverría en julio de 2014, presentando además un atípico aditivo en julio de 2016. En cuanto a efectos de calendario, la serie del **número total de nóminas** se ve afectada por el efecto provocado por la Semana Santa, ya que en esas fechas se produce un aumento del número de contrataciones vinculadas principalmente a la hostelería y al sector servicios.

En definitiva con estos ejemplos se ha intentado mostrar algunas de las correcciones más recurrentes en todas las series, las cuales, una vez corregidas serán utilizadas como variables explicativas en los modelos propuestos para intentar explicar el comportamiento de la `FBC` de la economía gallega.

Capítulo 6

Modelización de la formación bruta de capital

Una vez se han transformado y corregido las variables candidatas a entrar en los modelos, disponemos de 23 variables de frecuencia mensual que serán utilizadas como posibles variables explicativas del modelo, y la `FBC` transformada a frecuencia mensual como variable endógena. Para todas las variables disponemos de un período temporal comprendido entre enero de 2011 y marzo de 2019. Con el fin de ajustar los modelos y comprobar la calidad de los mismos a la hora de realizar predicciones se dividirá el conjunto de datos de la siguiente manera:

- Conjunto de datos de entrenamiento: comprendidos en el período enero-2011 y diciembre-2018.
- Conjunto de datos de predicción: comprendidos en el período enero-2019 y marzo-2019.

Empleado las variables mensuales obtenidas mediante las variables internas de ABANCA, se ajustarán distintos modelos con el objetivo de ajustar la `FBC` publicada por el IGE mensualizada. Una vez ajustados los modelos, mediante el conjunto de datos de predicción podemos: i) comparar los valores mensuales ajustados con los de la serie de la `FBC` mensualizada, ii) mediante la trimestralización de los valores ajustados realizar la comparación con los datos trimestrales publicados por el IGE.

En las siguientes secciones se mostrarán diferentes modelos de regresión empleados para dicha tarea, en concreto en la Sección 6.1 se ajustará un modelo de regresión lineal dinámica, y en la Sección 6.2 se ajustará un modelo de regresión aditivo generalizado.

6.1. Modelo de regresión dinámica (DLM)

Dado que nos encontramos en un contexto de series temporales, al realizar un ajuste de un modelo de regresión lineal múltiple es muy posible que los errores del mismo presenten una dependencia temporal que invaliden las hipótesis básicas del modelo, en concreto la de independencia de los residuos. Para solucionarlo, decidimos ajustar un modelo de regresión lineal dinámica como el estudiado en la Sección 3.2. Por tanto, en esta sección se intentará modelizar a través de un regresión lineal dinámica el comportamiento de la `FBC` de la economía gallega.

En un primer lugar, debemos de asegurarnos que las series sean estacionarias, ya que de no serlo podríamos estar cometiendo graves errores. Tal y como se ha visto en la Sección 3.2, es necesaria la condición de estacionariedad en las series para la correcta interpretación de los coeficientes de correlación cruzada (CCF), ya que de no cumplirse podríamos caer en la problemática de la regresión espuria. Por tanto, antes de ajustar el modelo procederemos a realizar el contraste de estacionariedad de "*Dickey - Fuller*". Si no rechazamos la hipótesis nula del mismo, no podremos asegurar que las series sean estacionarias y por tanto, procederemos al preblanqueo de las 23 variables explicativas con el fin de conseguir dicha estacionariedad.

Mediante el análisis gráfico de los coeficientes de correlación cruzada no se ha podido encontrar una dependencia temporal entre las variables candidatas y la `FBC` que nos permita ajustar un modelo de regresión entre las mismas. Por lo que para determinar qué variables resultarían significativas en el modelo y cuales serían los órdenes del proceso $ARMA(p,q) \times (P,Q)_s$ que modelizan la dependencia temporal entre las variables explicativas elegidas y la `FBC`, se han propuesto dos alternativas diferentes:

1. Ajuste de un modelo de regresión lineal múltiple, como el visto en la Sección 3.1. Este ajuste, a parte de darnos una idea sobre las variables que pueden resultar significativas, también nos da información sobre los órdenes del proceso $ARMA(p,q) \times (P,Q)_s$ que modelizan el comportamiento dinámico entre las variables.

2. Ajuste de un modelo de regresión lineal dinámica mediante la programación de una función en R realizando múltiples combinaciones entre las distintas variables explicativas del modelo, para intentar encontrar un ajuste que cumpla las hipótesis básicas del mismo.

1. Ajuste de un modelo de regresión lineal múltiple

Para seleccionar un modelo de regresión lineal múltiple, hemos procedido a realizar un "*stepwise backward*", partiendo de un modelo con todas las variables explicativas se ha ido eliminando sucesivamente la menos significativa. Realizando este proceso de manera iterativa hasta que todas las variables resulten significativas, se ha llegado a un modelo en el cual el signo de algunos de sus coeficientes estimados no se correspondía con la lógica económica. De nuevo, basándose en el p-valor del contraste de significación de los parámetros del modelo, se han ido eliminando aquellas variables menos significativas cuya influencia sobre la variable explicada fuese en contra de la lógica económica.

Mediante este procedimiento se encontró un modelo de regresión lineal múltiple para el cual todas las variables introducidas resultaban significativas, pero los residuos de dicho ajuste no cumplían la condición de estacionariedad imposibilitando modelizar la dependencia temporal presente en los residuos. Es por ello, que se ha decidido recurrir a las distintas combinaciones de las batería de 23 variables para encontrar un modelo estadísticamente válido.

2. Combinación de distintas variables explicativas

Dado que ni mediante el análisis de los coeficientes de correlación cruzada, ni mediante un ajuste de un modelo de regresión lineal múltiple se ha podido encontrar un modelo válido para explicar el comportamiento de la `FBC`, se ha procedido a la creación de una función que nos permita ajustar y comprobar para cada combinación de posibles variables un modelo de regresión lineal dinámico válido, permitiéndonos explicar el comportamiento de la `FBC`.

Esta función se compone de los siguientes pasos:

Paso 1 Mediante la función `combn()` se procede a realizar diferentes combinaciones de la batería de 23 variables construidas. En concreto se probarán combinaciones de 3, 4 , 5 y 6 variables.

Destacar que no se han probado combinaciones superiores a 6 variables debido al elevado coste computacional requerido, ya que para cada combinación se ajustarán diferentes combinaciones de parámetros del modelo $ARMA(p,q) \times (P,Q)_s$. Tampoco se realizarán combinaciones inferiores a 3 variables ya que, para la mayoría de los casos analizados, el número de parámetros necesarios para modelizar la dependencia temporal de los residuos se encuentra en el orden de 4-5, provocando que para combinaciones menores a 3 variables el peso de la componente temporal en los modelos fuese demasiado elevada.

Paso 2 Para cada una de las posibles combinaciones de variables se le ajustará un modelo de regresión lineal dinámica, permitiendo que el máximo de parámetros de la parte regular del proceso sea $ARMA(p=2,q=1)$, y el número máximo de parámetros de la parte estacional del proceso sea $ARMA(P=1,Q=1)_{12}$. Ya que elevar la búsqueda a órdenes mayores supone un gran coste computacional y el objetivo reside en encontrar una combinación de variables con un proceso

de dependencia poco elevado que permita al modelo detectar posibles cambios en el entorno macroeconómico.

Paso 3 Para cada uno de los posibles modelos de regresión lineal dinámica ajustados, se analizarán las siguientes hipótesis sobre los residuos de los mismos y la significatividad de todas las variables del modelo:

- Contraste de significatividad de todos los parámetros del modelo Student's t-Test.
- Contraste de estacionariedad de Dickey - Fuller sobre los residuos.
- Contraste de media 0 de Student's t-Test sobre los residuos.
- Contraste de incorrelación de Ljung-Box para los 30 primeros retardos temporales.

Paso 4 Por último, se almacenarán aquellas combinaciones cuyo modelo de regresión dinámica cumpla con todas las hipótesis comentadas en el Paso 3.

Mediante este método, se han obtenido diversos modelos que cumplen las hipótesis básicas exigidas a un modelo de regresión lineal dinámico y comentadas en el Paso 3. De estos posibles modelos, en función al criterio BIC de selección de modelos visto en la Sección 2.6 se ha elegido como mejor modelo el siguiente:

$$\begin{aligned} FBC_t = {} & \beta_1 \texttt{CIRC\_RATIO\_Mean}_t + \beta_2 \texttt{CIRC\_R\_NUMCONT}_t + \beta_3 \texttt{CIRC\_SALDO\_EXC\_Mean}_t \\ & + \beta_4 \texttt{CIRC\_RATIO\_EXC\_Mean}_t + \beta_5 \texttt{IMP\_TOTALES}_t + \varepsilon_t \\ \varepsilon_t = {} & \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \Phi_1 \varepsilon_{t-12} + a_t + \theta_1 a_{t-1} + \Theta_1 a_{t-12} \end{aligned} \tag{6.1}$$

donde β_1, β_2, β_3, β_4, β_5 son los parámetros del modelo, ε_t los errores del modelo, ϕ_1,ϕ_2, Φ_1, θ_1 y Θ_1 los parámetros del modelo del proceso Box-Jenkins, y a_t son variables i.i.d que se distribuyen como una $N(0,1)$.

Una vez estimado, y tras haber realizado los contrastes de significación de los parámetros (se puede observar en la Figura 6.1 que todos resultan significativos a un nivel de significación del 5 %), procedemos a realizar el contraste de las hipótesis básicas del modelo.

```
                   Estimate  Std. Error  z value  Pr(>|z|)
ar1              1.93067429  0.03462707  55.7562 < 2.2e-16 ***
ar2             -0.93170318  0.03442689 -27.0632 < 2.2e-16 ***
ma1              0.70107174  0.10669032   6.5711 4.995e-11 ***
sar1            -0.42343112  0.09675946  -4.3761 1.208e-05 ***
sma1            -0.99996668  0.15687225  -6.3744 1.837e-10 ***
CIRC_RATIO_Mean_1  -0.06222168  0.01595190  -3.9006 9.596e-05 ***
CIRC_NUMCONT_1     -0.01799586  0.00740057  -2.4317 0.0150287 *
CIRC_SALDO_EXC_Mean -0.00091874  0.00034357  -2.6741 0.0074941 **
CIRC_RATIO_EXC_Mean -0.00169911  0.00063786  -2.6638 0.0077273 **
IMP_TOTALES       0.00276565  0.00082125   3.3676 0.0007582 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

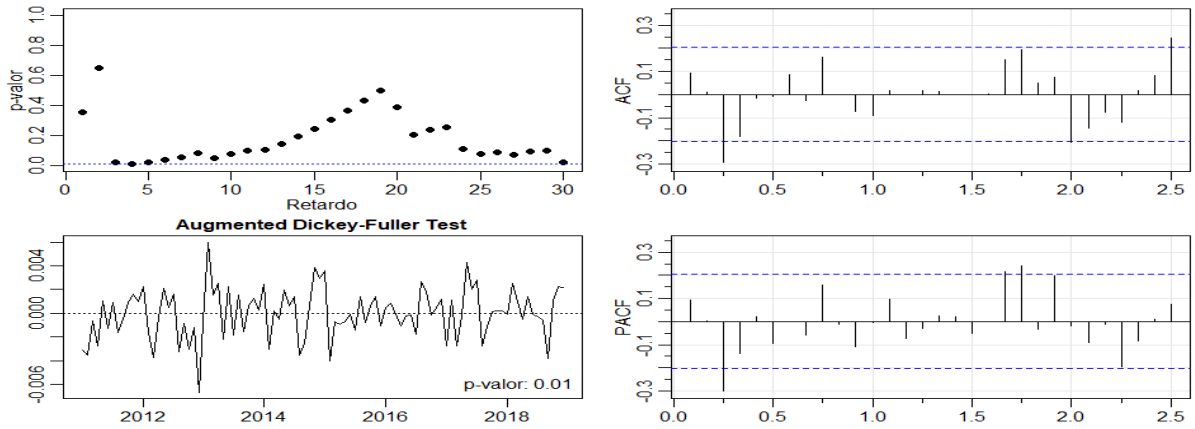
Figura 6.1: Coeficientes de las variables explicativas del modelo.

Contraste	Hipótesis	p-valor
Jarque Bera	Normalidad	0, 38
Shapiro-Wilk	Normalidad	0, 53
Student's t-Test	Media cero	0, 96
Dickey-Fuller Test	Estacionariedad	0, 01

Tabla 6.1: Contraste de hipótesis.

Como se puede observar en la Tabla 6.1, los residuos del modelo de la Ecuación (6.1) cumplen la hipótesis de normalidad para los contrastes de Jarque-Bera y Shapiro-Wilk. Además los residuos son estacionarios y de media cero. Para contrastar la hipótesis de independencia se ha calculado el estadístico de contraste de incorrelación de Ljung-Box para cada uno de los retardos temporales, los p-valores de

dicho contraste pueden verse en la Figura 6.2a. Por tanto, tras el contraste de hipótesis podemos decir que los residuos del modelo son estacionarios, de media cero e independientes, ya que, bajo normalidad, la incorrelación equivale a independencia, siendo el modelo válido para realizar predicciones.



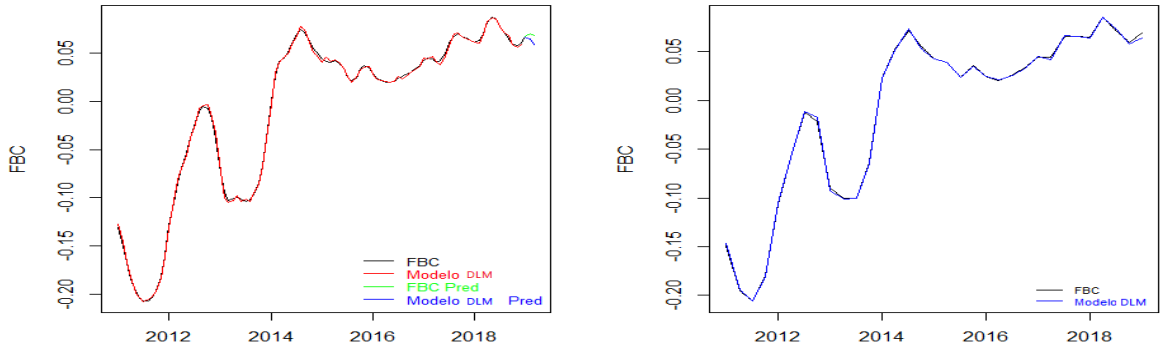
(a) P-valores del contraste de Ljung-Box para los diferentes retardos de los residuos del modelo.

(b) Coeficientes de correlación simple y parcial de los residuos del modelo.

Figura 6.2: Contrastes gráficos sobre los residuos del modelo.

Llegados a este punto, antes de realizar predicciones conviene aclarar que todas las variables introducidas en el modelo tienen una relación contemporánea con la `FBC`. Por tanto, a la hora de realizar predicciones, entendemos las mismas como el valor estimado por el modelo para los datos obtenidos a cierre de mes. Los cuales nos permiten conocer de manera adelanta a la publicación trimestral del IGE, cómo está evolucionando la inversión privada en Galicia. Además, para la comparación con el dato real para el primer trimestre del año publicado por el IGE, recurrimos a la trimestralización de las predicciones mensuales correspondientes a dicho trimestre.

Estas predicciones pueden ver el la Figura 6.3.



(a) Valores ajustados y predicciones mensuales.

(b) Valores ajustados y predicciones trimestrales.

Figura 6.3: Valores ajustados y predicciones para el modelo DLM.

En la Tabla 6.2 puede verse que el modelo de regresión lineal dinámica propuesto en la Ecuación (6.1) ofrece unas predicciones cercanas al valor real de la `FBC` para el primer trimestre del año, además en la Figura 6.3 se aprecia que el modelo propuesto tampoco ofrece un mal ajuste. Sin embargo, debemos destacar que la modelización de la `FBC` mediante un modelo de regresión dinámica presenta varios incon-

venientes:

- Las variables introducidas en el mismo no resultan del todo coherente con la lógica económica.
- El número de parámetros necesarios para modelizar la dependencia temporal de los residuos es elevado.

Es por ello por lo que decidimos modelizar la `FBC` mediante otro tipo de metodologías, en concreto mediante la regresión GAM.

	Enero	Febrero	Marzo	2019 Q1
Predicciones Modelo	0,066	0,065	0,059	0,064
Valor "real" `FBC`	0,067	0,069	0,068	0,069
Errores Predicción	−0,001	−0,004	−0,009	−0,005

Tabla 6.2: Predicciones mensuales y trimestrales modelo DLM.

6.2. Modelo de regresión aditivo generalizado (GAM)

Como hemos podido ver en la Sección 3.4, los modelos GAM son mucho más flexibles que los modelos DLM vistos en la Sección 3.2, permitiéndonos superar las limitaciones derivadas de la rigidez de estos últimos. Es posible que existan relaciones mucho más complejas que una simple relación lineal entre las covariables y la variable endógena del modelo, y es por ello que decidimos utilizar un modelo GAM para realizar el ajuste del modelo para la inversión empresarial en Galicia.

A continuación, se explicará cuál ha sido el procedimiento llevado a cabo para conseguir un modelo válido que nos permita realizar predicciones de la `FBC`.

En primer lugar, se ha realizado una adaptación del proceso "*stepwise backward*" de un modelo de regresión clásico, ya que para determinar cuáles son las variables explicativas del modelo, comenzamos con un modelo con las 23 métricas de la Tabla 1.5 expresadas como funciones suaves. Una vez ajustado, aquellas variables que no tengan una relación no lineal con la variable endógena, serán ajustadas de manera paramétrica para luego ir eliminando aquellas que no resulten significativas a un nivel de significación del 5 %. Una vez que hemos conseguido un modelo en el cual todas las variables son significativas y que la influencia de las mismas sobre la variable endógena es coherente con la teoría económica, procederemos a la comprobación de la condición de estacionariedad sobre los residuos del ajuste, además de contrastar las hipótesis básicas del modelo. Si se detecta dependencia entre los residuos del modelo, se procederá al ajuste de un modelo $ARMA(p,q) \times (P,Q)_s$ para los mismos. Para encontrar los órdenes de este proceso, se llevará a cabo un análisis gráfico de los residuos mediante las funciones de autocorrelación simple y parcial, y mediante la función `forecast()` del paquete *forecast* Hyndman et al. (2019). Por último, una vez comprobado que el modelo propuesto para los residuos es válido, procedemos al ajuste conjunto del modelo completo.

En resumen, los pasos a seguir se recogen en el siguiente esquema:

Paso 1 "*Stepwise backward*" con el fin de encontrar un modelo preliminar en donde las variables del mismo resultan significativas al nivel de significación del 5 % y la influencia de las mismas sobre la variable respuesta es coherente con la lógica económica.

Paso 2 Comprobación de la condición de estacionariedad de los residuos y la hipótesis de independencia de los mismos. Una vez comprobado que son estacionarios, de incumplirse la hipótesis de independencia se procede al ajuste de un modelo $ARMA(p,q) \times (P,Q)_s$ sobre los residuos.

Paso 3 Comprobación de la significación de los órdenes del modelo propuesto para los residuos y contraste de las hipótesis básicas del mismo.

Paso 4 Composición del modelo completo mediante el modelo de regresión GAM y el proceso Bos-Jenkins que modeliza la dependencia temporal de los residuos.

A efectos ilustrativos, a continuación se detalla cómo se compone el modelo completo a partir de un el modelo obtenido en Paso 1.

En primer lugar, mediante el procedimiento "*stepwise backward*" se procederá a la búsqueda de las variables que mediante funciones suaves formarán parte de la componente no paramétrica del modelo, y las que forman parte de la componente paramétrica del mismo. Una vez realizada, la expresión correspondiente al ajuste del modelo se muestra en la Ecuación (6.2):

$$\begin{aligned} FBC_t = \beta_1 \texttt{LARG\_NUMFORMA\_Sum}_t \; + \; \beta_2 \texttt{TRANSF\_R\_IMPT\_Mean}_t \; + \; \eta(\texttt{CIRC\_SALDO\_Mean}_t) \\ + \eta(\texttt{CIRC\_LIMITE\_Mean}_t) \; + \; \eta(\texttt{CIRC\_NUMCONT}_t) \\ + \eta(\texttt{NOM\_NUM\_Sum}_t) \; + \; \varepsilon_t \end{aligned} \tag{6.2}$$

donde β_1 y β_2 son los parámetros del modelo, η la función de suavizado, y ε_t los errores.

Una vez estimado, y tras haber realizado los contrastes de significación de los parámetros (se puede ver en la Figura 6.1 que resultan todos significativos a un nivel de significación del 5 %), procederemos a realizar el contraste de las hipótesis básicas del modelo.

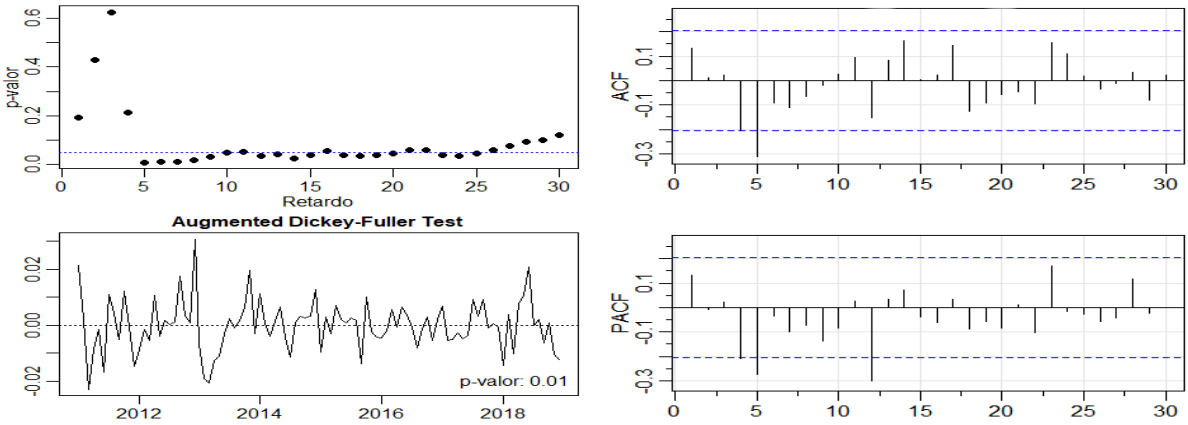
```
Parametric coefficients:
                 Estimate Std. Error t value Pr(>|t|)
LARG_NUMFORMA_Sum  0.02760    0.00501   5.509 6.51e-07 ***
TRANSF_R_IMPT_Mean 0.13031    0.01144  11.388  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                       edf Ref.df      F  p-value
s(CIRC_SALDO_Mean_1)  10.961 13.327 22.208  < 2e-16 ***
s(CIRC_LIMITE_Mean_1)  2.397  3.063  5.523  0.00176 **
s(CIRC_NUMCONT_1)      6.767  7.785  9.135 5.62e-09 ***
s(NOM_NUM_Sum)         8.346  8.835 66.281  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.984   Deviance explained = 98.9%
P-REML = -227.05  Scale est. = 0.00011833  n = 96
```

Figura 6.4: Coeficientes de las variables explicativas del modelo.

Contraste	Hipótesis	p-valor
Jarque Bera	Normalidad	0, 03
Shapiro-Wilk	Normalidad	0, 06
Student's t-Test	Media cero	0, 88
Dickey-Fuller Test	Estacionariedad	0, 01

Tabla 6.3: Contraste de hipótesis.

Como podemos ver en Tabla 6.3, para un nivel de significación del 5 % los residuos no son normales bajo el criterio del test de normalidad Jarque Bera, pero si para el test Shapiro-Wilk. Además los residuos tienen media 0 y son estacionarios. En cuanto a la hipótesis de independencia, ha sido contrastada mediante el test de Ljung-Box para cada uno de los retardos de la serie. En la Figura 6.5b puede verse que existe una estructura de correlación en los residuos del modelo, y además en la Figura 6.5a, que representa los p-valores del estadístico de contraste de Ljung-Box para distintos retardos temporales, vemos que a partir del quinto retardo la hipótesis de incorrelación deja de cumplirse. Por tanto, el modelo estimado no resultaría válido siendo necesaria la búsqueda de un modelo $ARMA(p,q) \times (P,Q)_s$ que nos ayude a modelizar la estructura de dependencia temporal presente en los residuos del modelo.



(a) P-valores del contraste de Ljung-Box para los diferentes retardos de los residuos del modelo.

(b) Coeficientes de correlación simple y parcial de los residuos del modelo.

Figura 6.5: Contrastes gráficos sobre los residuos del modelo.

A través de la función `auto.arima()` del paquete forecast (Hyndman et al. (2019)), procedemos a comprobar si existe algún proceso que pueda modelizar la estructura temporal de los residuos del modelo. Entre los diferentes modelos propuestos hemos elegido, en base al criterio BIC, el modelo $ARMA(1,0)\times(1,0)_{12}$. Se puede ver en la Figura 6.6 que los parámetros del proceso $ARMA(1,0)\times(1,0)_{12}$ resultan significativos a un nivel de siginificación del 5 %.

```
     Estimate Std. Error z value Pr(>|z|)
ar1     0.212      0.107   1.985    0.047 *
sar1   -0.280      0.119  -2.358    0.018 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

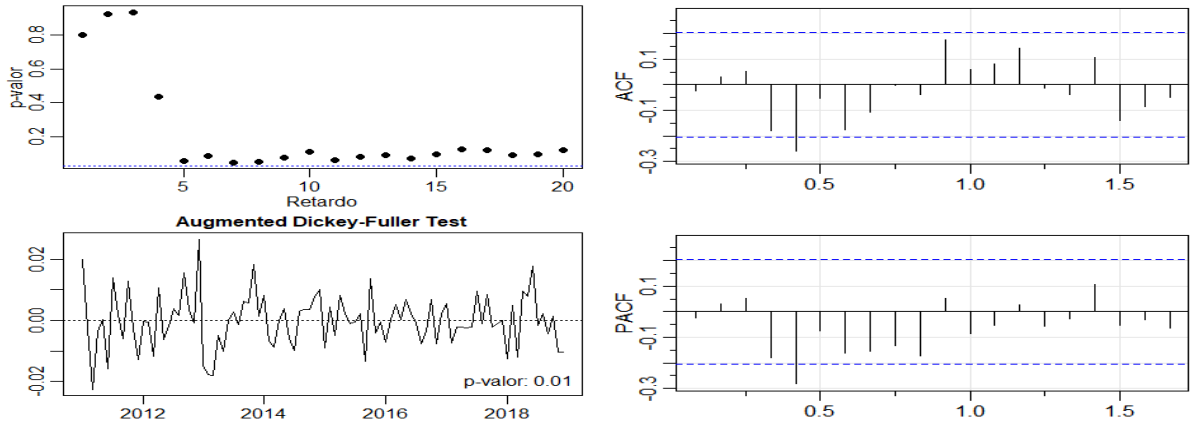
Figura 6.6: Coeficientes del proceso Box-Jenkins.

Una vez que se ha comprobado que los parámetros del modelo para los residuos del ajuste de la Ecuación (6.2) son todos significativos, procedemos a comprobar de nuevo las hipótesis básicas del mismo.

Contraste	Hipótesis	p-valor
Jarque Bera	Normalidad	0, 43
Shapiro-Wilk	Normalidad	0, 66
Student's t-Test	Media cero	0, 93
Dickey-Fuller Test	Estacionariedad	0, 01

Tabla 6.4: Contrastes de las hipótesis básicas del modelo GAM con residuos ARMA.

Como puede verse en la Tabla 6.4, se han incrementado los p-valores de ambos contrastes de normalidad, y los contrastes Dickey- Fuller y Student's t-Test nos confirman que los residuos son estacionarios y de media cero. En cuanto a la hipótesis de independencia en la Figura 6.7a, puede verse que los p-valores del contraste de incorrelación de Ljung-Box aumentan significativamente, pudiendo afirmar que los residuos del modelo son indenpendientes, ya que, bajo normalidad, la incorrelación equivale a independencia.



(a) P-valores del contraste de Ljung-Box para los diferentes retardos de los residuos del modelo.

(b) Coeficientes de correlación simple y parcial de los residuos del modelo.

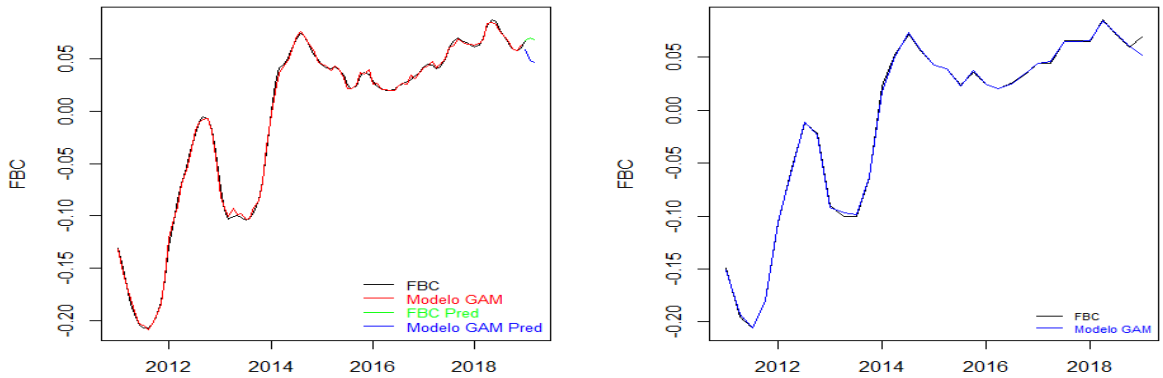
Figura 6.7: Contrastes gráficos sobre los residuos del modelo.

Por tanto, la nueva expresión del modelo completo con los parámetros de regresión correspondientes a las variables explicativas y del modelo ARMA propuesto para los residuos, se corresponde con la Ecuación (6.3).

$$\begin{aligned} FBC_t = {} & \beta_1 \texttt{LARG\_NUMFORMA\_Sum}_t \; + \; \beta_2 \texttt{TRANSF\_R\_IMPT\_Mean}_t \; + \; \eta(\texttt{CIRC\_SALDO\_Mean}_t) \\ & + \eta(\texttt{CIRC\_LIMITE\_Mean}_t) \; + \; \eta(\texttt{CIRC\_NUMCONT}_t) \; + \; \eta(\texttt{NOM\_NUM\_Sum}_t) \; + \; \varepsilon_t \\ \varepsilon_t = {} & \phi_1 \varepsilon_{t-1} + \Phi_1 \varepsilon_{t-12} + a_t \end{aligned} \tag{6.3}$$

donde β_1 y β_2 son los parámetros del modelo, η la función de suavizado, ε_t con los errores del modelo, ϕ_1, Φ_1 los parámetros del proceso Box-Jenkins y a_t son variables i.i.d que se distribuyen como una $N(0,1)$.

En resumen, se ha obtenido un modelo válido en el sentido de que cumple la validez estadística exigida a cualquier modelo, es decir, todos los parámetros del mismo resultan significativos y cumplen las hipótesis básicas del mismo, y además, el sentido económico de las variables introducidas es el que cabría esperar. Podemos afirmar entonces, que tenemos un modelo con el cual realizar predicciones de la `FBC`:



(a) Valores ajustados y predicciones mensuales.

(b) Valores ajustados y predicciones trimestrales.

Figura 6.8: Valores ajustados y predicciones para el modelo GAM con estructura ARMA para los residuos.

Por último, en la Tabla 6.5 se muestran las predicciones mensuales del modelo correspondiente a la Ecuación (6.3) y su versión trimestral, junto con los errores de predicción calculados como la diferencia entre el valor de la predicción y valor real en el dato de 2019 Q1, y los obtenidos mediante el proceso de mensualización explicado en la Sección 5.1 en el caso de datos mensuales.

	Enero	Febrero	Marzo	2019 Q1
Predicciones Modelo	0,059	0,048	0,047	0,052
Valor "real" FBC	0,067	0,069	0,068	0,069
Errores Predicción	−0,008	−0,021	−0,021	−0,017

Tabla 6.5: Predicciones mensuales y trimestrales modelo GAM.

Selección de modelos

Siguiendo el procedimiento anterior, se han encontrado varios modelos válidos reflejados en la Tabla 6.6, entendiendo por modelo válido aquel modelo que cumple los requisitos anteriores, es decir, todas sus covariables resultan significativas a un nivel de confianza del 95 %, se cumplen las hipótesis básicas del mismo, y las variables introducidas en el modelo son coherentes con la lógica económica.

Modelos	Error Predicción	R^2_{ajust}	Deviance
Modelo 1	−0,027	0,982	98,7 %
Modelo 2	−0,017	0,986	98,8 %
Modelo 3	−0,012	0,982	98,7 %
Modelo 4	−0,025	0,960	95,8 %
Modelo 5	−0,025	0,968	97,6 %
Modelo 6	−0,025	0,965	97,2 %
Modelo 7	−0,016	0,951	95,1 %

Tabla 6.6: Contrastes hipótesis básicas modelo GAM.

Para establecer un criterio de selección entre los distintos modelos válidos encontrados nos basamos en los siguientes criterios:

- Validez estadística: entendida como el cumplimiento de las hipótesis básicas del modelo y la significación de todas las variables incluidas en el mismo.
- Sentido económico: las variables introducidas en el modelo deben de ser coherentes con la lógica económica.
- Medida del grado de ajuste: comparación de modelos en términos del coeficiente de determinación ajustado (R^2_{ajust}).
- Buen nivel predictivo: ejercicio de backtesting realizado a través del cálculo del error de predicción de los modelos propuestos.

De los modelos vistos en la Tabla 6.6, el que mejor comportamiento reporta en base a los criterios de selección comentados es el **Modelo 2**, para el cual el procedimiento de ajuste realizado sobre el mismo se ha explicado en esta sección. Destacar que el modelo elegido, a parte de ser el que mejor se ajustaba a los criterios de selección propuestos, es el que menor variabilidad presenta para los valores ajustados, siendo ésta una cualidad a destacar cuando se trabaja con modelos de predicción, en particular, en un contexto de variables macroeconómicas en el que se pretende obtener una previsión de un indicador de la evolución de la economía.

Capítulo 7

Comparativa entre las distintas metodologías

En el Capítulo 6 se ha trabajado con dos metodologías con un objetivo en común, la predicción de la FCB de la economía gallega. En este capítulo se llevará a cabo una comparación de ambas metodologías con el fin de determinar qué modelo se ajusta mejor al comportamiento de la FCB gallega. Para ello, los modelos válidos seleccionados en la Sección 6.1 y en la Sección 6.2, correspondientes a Ecuación (6.1) y Ecuación (6.3) respectivamente, se compararán en función a:

1. Grado de bondad de ajuste: medido a través del coeficiente de determinación ajustado (R^2_{ajust}) y el criterio de selección BIC.
2. Capacidad predictiva: aproximada mediante los errores de predicción.
3. Sensibilidad de los modelos ante cambios en el escenario macroeconómico: aproximado mediante la simulación de distintas situaciones sobre las variables.

Los puntos 1 y 2 los encontramos en los ajustes de ambos modelos y en las predicciones de los mismos del capítulo anterior y resumidos en la Tabla 7.1.

Modelos	Bondad de Ajuste		Errores de predicción			
	R^2_{ajust}	BIC	ene-19	feb-19	mar-19	2019 Q1
Modelo DLM	0.994	-818.077	-0,001	-0,004	-0,009	-0,005
Modelo GAM	0.986	-471,569	-0,008	-0,021	-0,021	-0,017

Tabla 7.1: Diferentes criterios para comparación de modelos.

Cabe destacar que el fin último de este trabajo no reside principalmente en buscar un modelo que nos ofrezca las mejores predicciones, si no que también se busca que los modelos posean la capacidad de detectar cambios en el entorno macroeconómico. Para ello, realizaremos un análisis de sensibilidad correspondiente al punto 3 de la comparación.

Para llevar a cabo este análisis se compararán los resultados dados por los modelos ante distintos escenarios macroeconómicos. En concreto:

- **Escenario base:** Se corresponde con los datos reales de las variables internas de la entidad a cierre de mes, los cuales tomaremos como punto de referencia.
- **Generación de un escenarios optimista:** Para simular una situación favorable en la economía gallega, se tomarán los datos de las variables explicativas del escenario base y se multiplicarán por

un coeficiente. En concreto, simularemos una situación en donde las variables del escenario base sean un 20 % superiores. Destacar que dependiendo del sentido de cada una de las variables sobre la `FBC` supondrá un aumento o reducción de las mismas. De esta manera se estaría simulando un shock positivo de la economía gallega.

- **Generación de un escenario adverso:** procedemos de la misma forma que la utilizada para la generación de un escenario optimista, pero la situación a simular se corresponderá con una reducción del 20 % de las variables respecto al escenario base. De algún modo estaríamos simulando un shock negativo de la economía gallega.

A continuación, se mostrarán las variables explicativas utilizadas por los modelos en cada uno de los diferentes escenarios comentados.

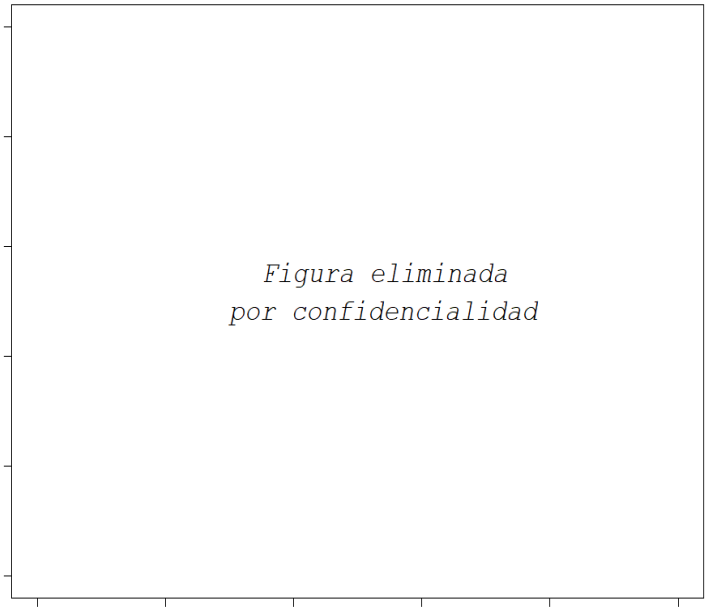


Tabla 7.2: Escenarios macroeconómicos simulados.

En los distintos distintos escenarios, se procederá a analizar la sensibilidad de cada modelo ante los mismos. Para ello, con los datos que se muestran en cada una de las tablas se realizarán las predicciones para cada uno de los escenarios planteados. Siguiendo la línea de exposición del trabajo, empezaremos por el modelo DLM correspondiente a la Ecuación (6.1), para luego continuar con el modelo GAM correspondiente a la Ecuación (6.3).

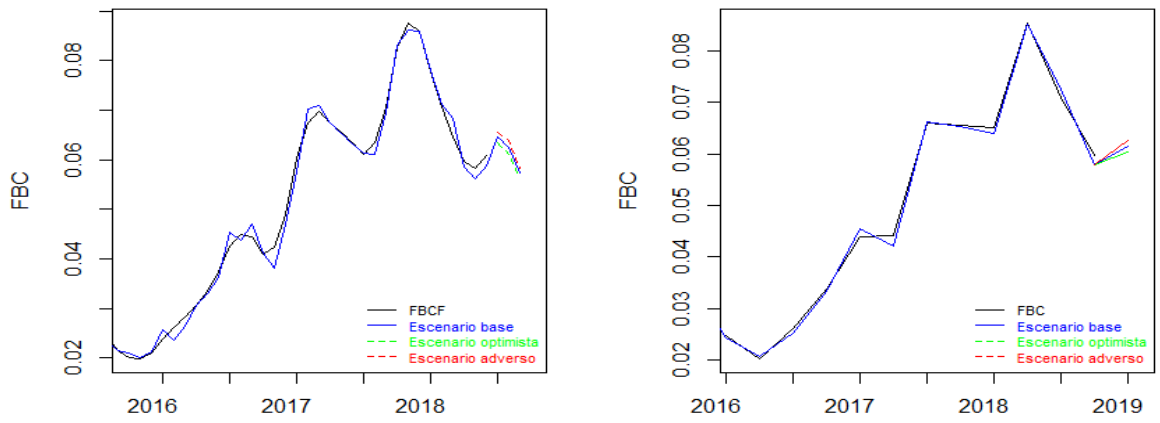
Modelo DLM

En la Tabla 7.3 se muestran las predicciones del modelo correspondiente a la Ecuación (6.1) para cada uno de los distintos escenarios:

	Enero	Febrero	Marzo	2019 Q1
Predicciones escenario base	0,065	0,062	0,057	0,061
Predicciones escenario optimista	0,066	0,064	0,058	0,065
Predicciones escenario adverso	0,066	0,064	0,058	0,063

Tabla 7.3: Predicciones mensuales y trimestrales ante los distintos escenarios modelo DLM.

En la Figura 7.1 se puede observar gráficamente los valores de las predicciones ante los distintos escenarios



(a) Predicciones mensuales para distintos escenarios. (b) Predicciones trimestrales para distintos escenarios.

Figura 7.1: Análisis de sensibilidad ante distintos escenarios macroeconómicos.

Como se puede apreciar en Tabla 7.3, no parece que el modelo ajustado tenga sensibilidad ante cambios en las variables, además la predicción del mismo para cada escenario carece de sentido económico, ya que ofrece una mejor evolución cuando se plantea un escenario económicamente desfavorable que cuando se simula una situación favorable. Como ya se ha comentado, el sentido de los coeficientes ajustados por el mismo no era el esperado según la lógica económica y, además, la estructura de correlación de los residuos mas compleja de lo deseado. Todo en su conjunto está provocando que el modelo no reaccione bien a los distintos escenarios y que la sensibilidad del mismo es limitada.

Modelo GAM

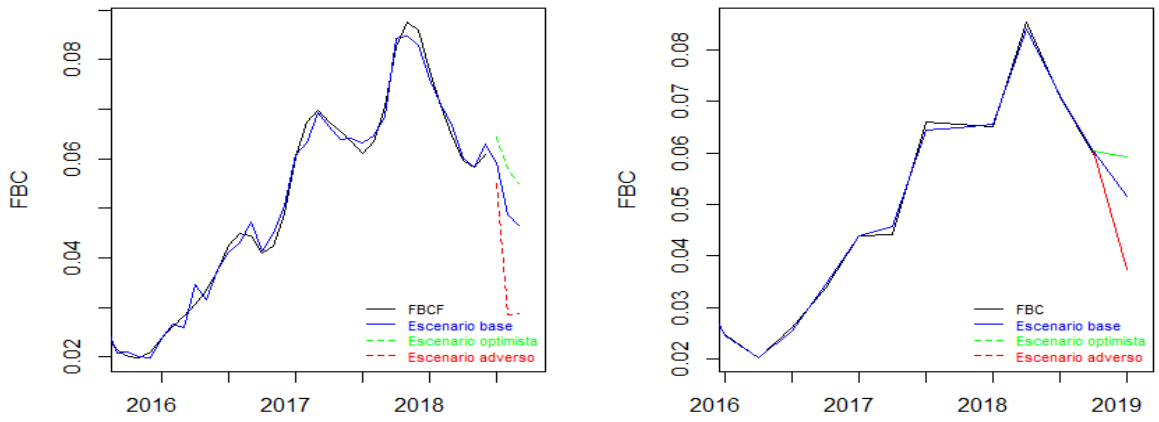
A continuación, para cada uno de los escenarios propuestos se realizarán predicciones con el modelo visto en la sección Sección 6.2, en concreto con el modelo correspondiente a la Ecuación (6.3).

En la Tabla 7.4 se muestran las predicciones del modelo correspondiente a la Ecuación (6.3) para cada uno de los distintos escenarios:

	Enero	Febrero	Marzo	2019 Q1
Predicciones escenario base	0,059	0,049	0,047	0,052
Predicciones escenario optimista	0,064	0,058	0,055	0,059
Predicciones escenario adverso	0,055	0,028	0,029	0,037

Tabla 7.4: Predicciones mensuales y trimestrales ante los distintos escenarios modelo GAM.

En la Figura 7.2 observan los valores de la predicción del modelo GAM para cada uno de los diferentes escenarios planteados



(a) Predicciones mensuales para distintos escenarios. (b) Predicciones trimestrales para distintos escenarios.

Figura 7.2: Análisis de sensibilidad ante distintos escenarios macroeconómicos.

Como se puede apreciar en la Figura 7.2 y en la Tabla 7.4, el modelo propuesto es sensible ante los diferentes escenarios planteados. Además la evolución de la dirección de las predicciones del mismo es coherente con la lógica económica.

En resumen, y en línea de los expuesto a comienzo del capítulo, el modelo elegido para predecir la FCB de la economía gallega será el modelo GAM, ya que además de proporcionar un mejor ajuste, las predicciones del mismo son más sensibles a posibles shocks y coherentes con la teoría económica, y por tanto, nos permitirán detectar mejor cambios que afecten a la `FBC` en Galicia.

Capítulo 8

Conclusiones

El objetivo de este trabajo se ha centrado en la obtención de un indicador para la formación bruta de capital de la economía gallega a partir de variables internas de ABANCA. Para tal fin, se han construido 23 variables vinculadas con distintos productos financieros relacionados con el comportamiento de las empresas en el corto plazo, las inversiones que estas realizan en el medio y largo plazo, así como el riesgo o deterioro de la actividad empresarial de la comunidad autónoma de Galicia. Estas series fueron construidas a partir de variables internas de negocio que nos permiten obtener información a cierre de cada mes, ofreciéndonos la ventaja de aportar un mayor dinamismo al indicador de la `FBC` publicado por el IGE, cuya publicación tiene carácter trimestral y presenta un retraso en su publicación de 53 días respecto al trimestre en cuestión.

Para construir este indicador se ha recurrido a distintas metodologías. Dado que las 23 variables de las que disponíamos eran series temporales, se ha estimado un modelo mediante regresión lineal dinámica que, a pesar de proporcionar un buen ajuste, ha mostrado escasa sensibilidad en la predicción de shocks de la economía y dudosa lógica económica en la interpretación de los efectos de las variables explicativas. Por ello, se ha recurrido a un modelo de regresión no paramétrica, en concreto, el modelo GAM, que como se ha podido ver proporciona un buen ajuste y sus predicciones permiten detectar cambios ante diferentes escenarios macroeconómicos. Cabe destacar que el indicador construido resulta de gran utilidad para ABANCA, y sigue una línea de trabajos que permiten disponer a la entidad de una modelización para cada una de los diferentes componentes del `PIB` vistos en el primer capítulo.

Por último, como futura línea de investigación, dada la riqueza de la información interna de ABANCA, se está estudiando la posibilidad de crear un indicador desagregado por las diferentes tipologías de empresas vistas en este trabajo. De esta forma, la entidad dispondría de un indicador que reflejaría como se estaría comportando el tejido empresarial gallego, siendo ésta una opción de gran interés, ya que permitiría realizar un análisis más exhaustivo sobre el comportamiento de la inversión la comunidad gallega para cada uno de los agentes económicos implicados en dicho proceso. Además, se está trabajando en la mejora del método de ajuste de modelos GAM mediante un procedimiento automático para la selección de las variables explicativas.

Bibliografía

Boxall, Mark and Brown, Gary and Buono, Dario and Elliot, Duncan (2015). *ESS guidelines on seasonal adjustment*. Luxembourg: Publications Office of the European Union.

Cabrero, Yolanda and García, Alfonso (2015). *Análisis Estadístico de Datos Espaciales Con Qgis y R*. UNED.

Chan, K.-S. and Ripley, B. (2018). *TSA: Time Series Analysis*. R package version 1.2.

Chiang, Alan Y (2009). Generalized Additive Models: An Introduction With R. *Technometrics*.

Christoph Sax and Peter Steiner (2016). *tempdisagg: Methods for Temporal Disaggregation and Interpolation of Time Series*. R package version 0.25.0.

Comisión Europea (2015). Guía del Usuario Sobre la Definición del Concepto de Pyme. *Oficina de Publicaciones de la Unión Europea*, pages 1–54.

Cryer, Jonathan D and Chan, Kung-Sik (2010). *Times Series Analysis with Applications in R*. Springer USA.

Eurostat (2014). *Fundamentos de SCN: Formulación de los elementos básicos*. Publications Office.

Hastie, Trevor and Tibshirani, Robert (1990). *Generalized additive models*. Chapman and Hal, New York.

Hastie, Trevor and Tibshirani, Robert and Friedman, Jerome (2009). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York, New York, NY.

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E., and Yasmeen, F. (2019). *forecast: Forecasting functions for time series and linear models*. R package version 8.8.

INE (2019). *Estándar del INE para la correción de efectos estacionales y efectos de calendario en las series coyunturales*.

Lequiller, François and Blades, Derek (2018). *Comprendiendo las Cuentas Nacionales*. OECD.

Peña, Daniel (2010). *Análisis de series temporales*. Alianza Editorial.

Petris, Giovanni and Petrone, Sonia and Campagnoli, Patrizia (2009). Dynamic linear models. In *Dynamic Linear Models with R*, pages 1–247. Springer New York.

Ruppert, David (2004). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. *Journal of the American Statistical Association*, 99(466):1–567.

Ruppert, David and Wand, M.P. and Carroll, Raymond J. (2009). Semiparametric regression. *Electronic Journal of Statistics*, 3:1193–1256.

Sax, C. and Eddelbuettel, D. (2018). Seasonal adjustment by X-13ARIMA-SEATS in R. *Journal of Statistical Software*, 87(11):1–17.

Tsay, Ruey S. (2010). *Analysis of Financial Time Series*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, USA.

US Census Bureau (2017). *X-13ARIMA-SEATS Reference Manual*. U.S Census Bureau.

Wood, Simon N. (2017). *Generalized Additive Models*. Chapman and Hall/CRC.

Yee, Thomas W. (2015). *Vector Generalized Linear and Additive Models*. Springer Series in Statistics. Springer New York, New York, NY.