



FACULTAD DE ESTUDIOS ESTADÍSTICOS

GRADO EN ESTADÍSTICA APLICADA

Curso 2019/2020

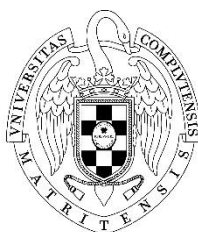
Trabajo de Fin de Grado

TÍTULO: Análisis de la sostenibilidad del sistema público de pensiones español, mediante una evaluación comparativa de técnicas univariantes y multivariantes de predicción de series temporales de agregados económicos y demográficos

Alumno: Gustavo Rodríguez Rodríguez

Tutor: José Luis Valencia Delfa

Septiembre de 2020



UNIVERSIDAD COMPLUTENSE
MADRID

Contenido

RESUMEN	3
1. INTRODUCCIÓN	4
1.1. El Origen de los Sistemas de Pensiones de Reparto	4
1.2. El Sistema Público de Pensiones Español	4
1.3. La Situación Actual	6
2. ESTADO DEL ARTE	13
2.1. Introducción	13
2.2. Enfoques de Estudio de la Sostenibilidad de un Sistema de Pensiones	13
3. HIPÓTESIS Y OBJETIVOS.....	16
3.1. Hipótesis.....	16
3.2. Objetivo Principal del Estudio	16
3.3. Objetivos Secundarios del Estudio	16
4. MATERIAL Y METODOLOGÍA	17
4.1. Ámbito y Periodo del Estudio.....	17
4.2. Fuentes de Datos.....	17
4.3. Definiciones.....	17
4.4. Pasos a Seguir en el Análisis de los Datos	19
4.5. Técnicas Estadísticas Utilizadas.....	21
• 4.5.1. Análisis Univariante.....	21
➤ 4.5.1.1. Métodos de Suavizado o Alisado Exponencial	21
➤ 4.5.1.2. Metodología Box-Jenkins	22
• 4.5.2. Análisis Multivariante.....	23
➤ 4.5.2.1. Análisis de Componentes Principales.....	23
➤ 4.5.2.2. Modelos Vectoriales.....	24
• 4.5.3. Variables de Intervención.....	26
4.6. Software	26
5. ANÁLISIS Y RESULTADOS	27
5.1. Procesamiento de las Series.....	27
5.2. Análisis.....	28
4.3. Comparativa y Selección del Mejor Modelo	52
4.4. Predicción.....	54
6. DISCURSIÓN Y CONCLUSIONES	58
7. BIBLIOGRAFÍA.....	60

RESUMEN

En las próximas páginas se evaluará la sostenibilidad futura de nuestro sistema público de pensiones, mediante el análisis de series temporales de agregados macroeconómicos y demográficos. La motivación de este trabajo es doble: económica y estadística.

Por un lado, el gasto en pensiones es uno de los mayores desembolsos que realiza el Estado y, actualmente, el sistema se encuentra en una situación de dificultad e incertidumbre, debido principalmente a la grave situación económica y a la dinámica demográfica.

Por otro lado, a pesar de que los modelos multivariantes son muy usados en economía para modelizar las interdependencias, no lo son tanto a la hora de predecir series temporales. El motivo es que a menudo los métodos univariantes superan a los multivariantes a causa de su simplicidad. Por ello en este trabajo comprobaremos que método nos proporciona una mayor precisión para predecir el porcentaje de Producto Interior Bruto que se destinará al gasto en pensiones en España.

Al final determinaremos que la mejor aproximación la obtendremos mediante modelos univariantes. También concluiremos que el sistema público de pensiones necesita cambios estructurales, puesto que la crisis producida por el COVID-19 ha acelerado su deterioro y pasará a estar en una situación insostenible en los próximos años.

Palabras clave: Pensiones, Sostenibilidad, Predicción, Series Temporales, Suavizado Exponencial, Box-Jenkins, Univariante, Multivariante, Variables de Intervención, COVID-19.

1. INTRODUCCIÓN

1.1. El Origen de los Sistemas de Pensiones de Reparto

En las sociedades previas a la Revolución Industrial los ancianos solo podían recibir ayuda a través de los recursos informales de apoyo social, principalmente los miembros de la familia y las amistades, lo que conllevaba altas tasas de morbilidad y mortalidad.

La industrialización y el capitalismo implicaron una transformación de la estructura familiar y una mayor esperanza de vida. Todos estos procesos impulsarían la demanda de creación de sistemas estructurados de ayuda para la jubilación, enfermedad, accidentes o discapacidad.

El primer sistema de Seguridad Social fue creado en 1881 por Otto von Bismarck con la finalidad de proporcionar una renta en situaciones vulnerables como la vejez o la incapacidad. Al terminar la Primera Guerra Mundial, los sistemas de seguros sociales se extendieron por las distintas naciones hasta llegar a la situación en la que se encuentran actualmente.

Durante su desarrollo surgió una pugna sobre cómo debía articularse un verdadero sistema de pensiones. Unos argumentaban que las aportaciones del trabajador y su pensión debían mantener una relación directa (*Sistemas de Reparto Contributivo o Bismarckianos*). Por el contrario, otros defendían un sistema de mínimos, en el cuál las pensiones tuvieran igual cuantía para cada trabajador (*Sistemas Asistenciales o Beveridge*).

Hoy en día la mayoría de los países europeos cuenta con lo que se denomina un sistema público de pensiones de reparto. En esencia, este tipo de sistemas de pensiones se basan en la idea del pacto intergeneracional para la redistribución de la renta, es decir, los trabajadores contribuyen con una parte de su salario a pagar las pensiones actuales con la esperanza de que el sistema se prolongue el tiempo suficiente como para que las generaciones venideras aporten sus futuras pensiones.

1.2. El Sistema Público de Pensiones Español

En nuestro país, la Ley General de la Seguridad Social de 1966 fue la encargada de implantar un modelo unitario e integrado de protección social con una base financiera de reparto, participación del Estado en la financiación y gestión pública. Con esta ley se establecía la edad de jubilación a los 65 años, se necesitaban 8 años de cotización para tener derecho a la misma y se tenían en cuenta los 2 años anteriores a la fecha de la jubilación para calcular la pensión.

A lo largo de las últimas décadas del siglo XX y las primeras del siglo XXI se han realizado sucesivas reformas con el objetivo de adecuar el sistema a las necesidades sociales, económicas y demográficas del país. A continuación, se muestra una tabla con los principales hitos y medidas que se tomaron en distintas etapas.

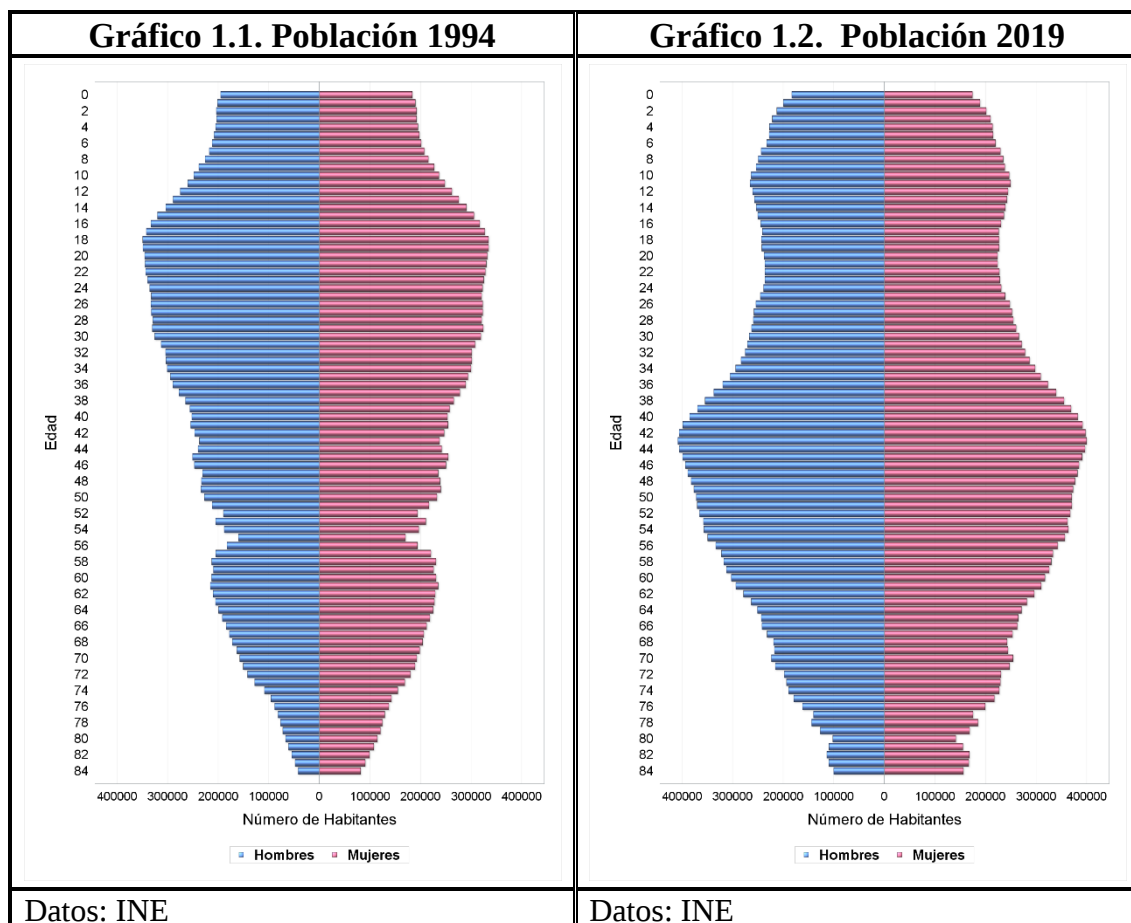
Tabla 1.1. Principales Reformas en el Sistema Público de Pensiones		
Década de los 80	Finales de los 90 y principios del 2000	Década del 2010
• Equiparación paulatina de las bases de cotización con los salarios reales. • Revalorización de las pensiones en función de la variación del IPC. • Extensión de la protección a más colectivos no cubiertos. • Aumento del período mínimo de cotización de 10 a 15 años. • Aumento desde los 2 años hasta los 8 años previos a la jubilación para el cómputo de la base reguladora de la pensión.	• Regulación de las pensiones no contributivas. • Firma del Pacto de Toledo para asegurar la estabilidad financiera y las prestaciones futuras. • Aumento de 8 a 15 en el número de años aplicados al cálculo de la base reguladora de la pensión. • Creación del Fondo de Reserva para asegurar la sostenibilidad del Sistema y atender posibles necesidades futuras. • Incentivación de la jubilación más allá de los 65 años.	• Aumento de la edad de jubilación de 65 a 67 años. • Aumento de la base reguladora de la pensión de 15 a 25 años. • Aumento de 35 a 37 años cotizados para alcanzar el 100% de la base reguladora de la pensión. • Introducción del Factor de Sostenibilidad. • Desvinculación de la revalorización de las pensiones al IPC y vinculación al Índice de Revalorización.
Fuente: Secretaría de Estado de la Seguridad Social		

Muchas de estas medidas se crearon con el propósito de aumentar el número de trabajadores afiliados a la Seguridad Social, disminuir el número de pensionistas y rebajar la cuantía de la pensión media. Sin embargo, a pesar de que en el momento de su aplicación supusieron un alivio para el sistema, se mantuvo la tendencia creciente tanto en la cantidad de pensiones como en el importe de la pensión media.

Todo ello, nos lleva a la siguiente pregunta: ¿En qué circunstancias nos encontramos ahora?

1.3. La Situación Actual

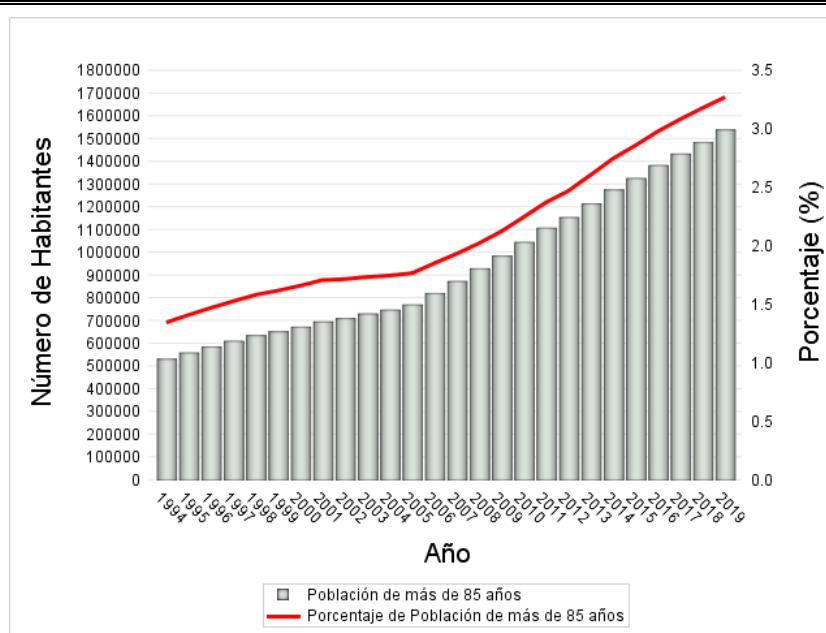
El progresivo envejecimiento de la población, debido al aumento de la esperanza de vida, ha modificado de manera sustancial la estructura de la pirámide poblacional española.



Cada vez ganan más peso los colectivos más longevos, en especial los de más de 65 años y más de 85 años.

Tabla 1.2. Población de más de 65 años (1994-2019)				
Año	Mayores de 65 años	Población Total	Porcentaje Población Mayor de 65 años	Incremento
1994	5.857.730	39.547.353	14,81%	
1999	6.616.982	40.369.666	16,39%	1,58%
2004	7.154.259	42.859.173	16,69%	0,30%
2009	7.727.117	46.367.550	16,66%	-0,03%
2014	8.510.477	46.455.123	18,32%	1,65%
2019	9.180.401	47.100.395	19,49%	1,17%
Datos: INE				

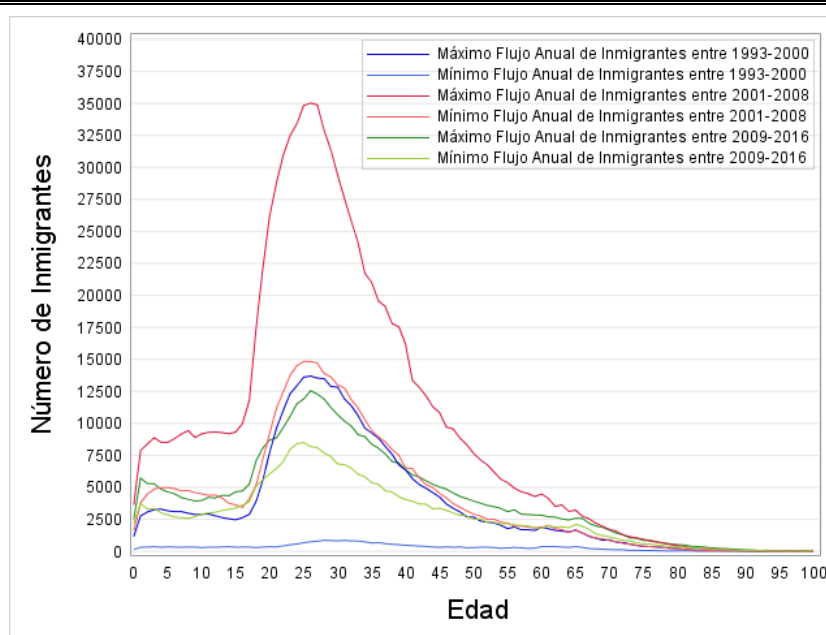
Gráfico 1.3. Población de más de 85 años (1994-2019)



Datos: INE

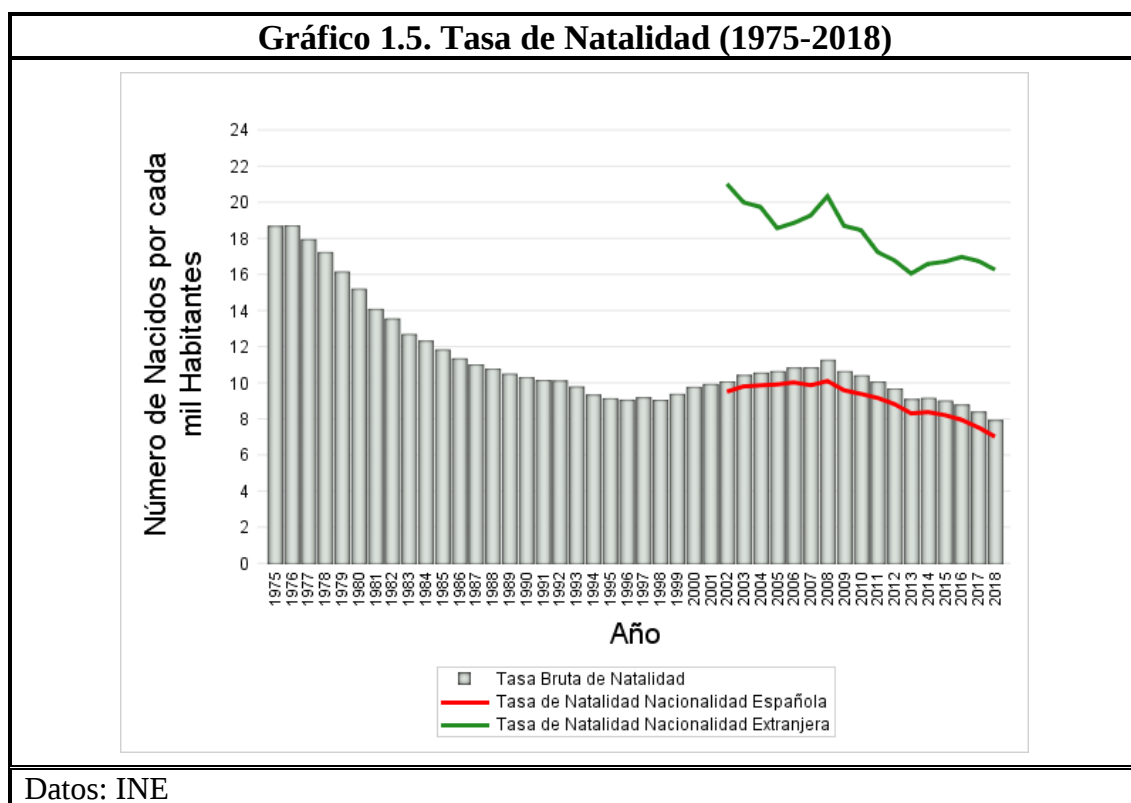
En la Tabla 1.2 y el Gráfico 1.3 vemos como el engrosamiento de las cohortes de más edad ha sido continuo durante los últimos 25 años. Tan solo la llegada de inmigrantes jóvenes, motivados por el crecimiento económico de comienzos de siglo XXI, pudo ralentizar la progresión del incremento en el peso específico de estos grupos sobre el total de la población.

Gráfico 1.4. Flujos Máximos y Mínimos de Inmigración Anual por Edades en los Periodos 1993-2000, 2001-2008 y 2009-2016



Datos: EUROSTAT

Mientras, las generaciones más jóvenes van mermando año tras año. Así, en el primer semestre de 2019 en España solo han nacido 170.074 personas, lo que supone la cifra más baja registrada desde 1941, primer año del que hay datos. También la tasa bruta de natalidad ha decrecido a lo largo de las últimas décadas, situándose en su mínimo desde que se empezaron a contabilizar datos en 1975.

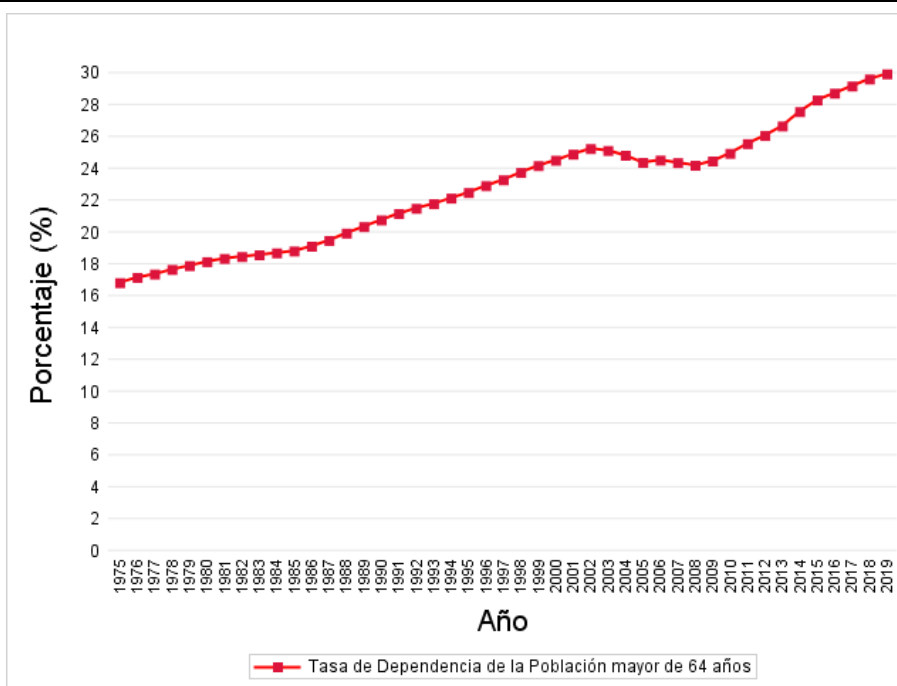


Este fuerte descenso de la natalidad, motivado por causas culturales, sociales y económicas, se produce esencialmente por el retraso en la edad a la que se tienen hijos.

Tabla 1.3. Edad Media a la Maternidad (1994-2018)	
Año	Edad Media
1994	29,74
1999	30,68
2004	30,87
2009	31,04
2014	31,78
2018	32,20
Datos: INE	

La unión de todos estos fenómenos demográficos produce un peligroso aumento de la tasa de dependencia de la población mayor de 64 años, entendida como el cociente entre la población mayor de 64 años y la población entre 16 y 64 años.

Gráfico 1.6. Tasa de Dependencia de la Población mayor de 64 años (1975-2019)



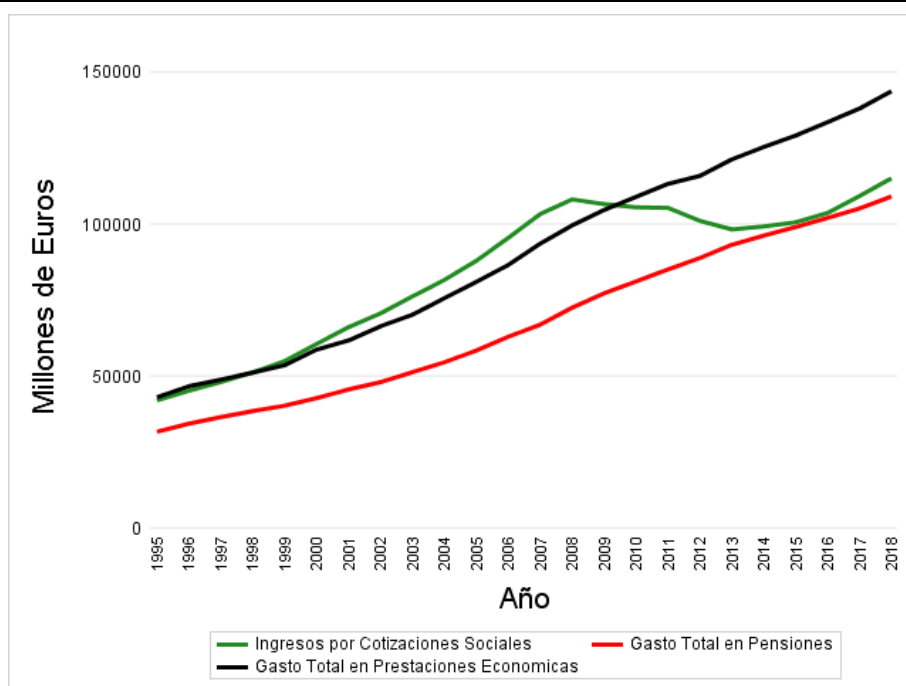
Datos: INE

Aunque este índice de dependencia sea, estrictamente hablando, un indicador demográfico, se utiliza ampliamente como indicador económico. Puesto que contrapone al grupo que es principal generador de riqueza, frente al grupo que es mayor consumidor de recursos del estado.

La tendencia claramente creciente que nos muestra el Gráfico 1.6 es una señal de que, si no se producen cambios estructurales, en el futuro tendremos desequilibrios económicos.

Por otro lado, la evolución del saldo del Sistema de Seguridad Social tampoco arroja resultados halagüeños. Aunque los ingresos de los últimos años mantienen una tendencia positiva gracias al aumento en las cotizaciones sociales, la partida de gastos crece a tasas todavía más altas.

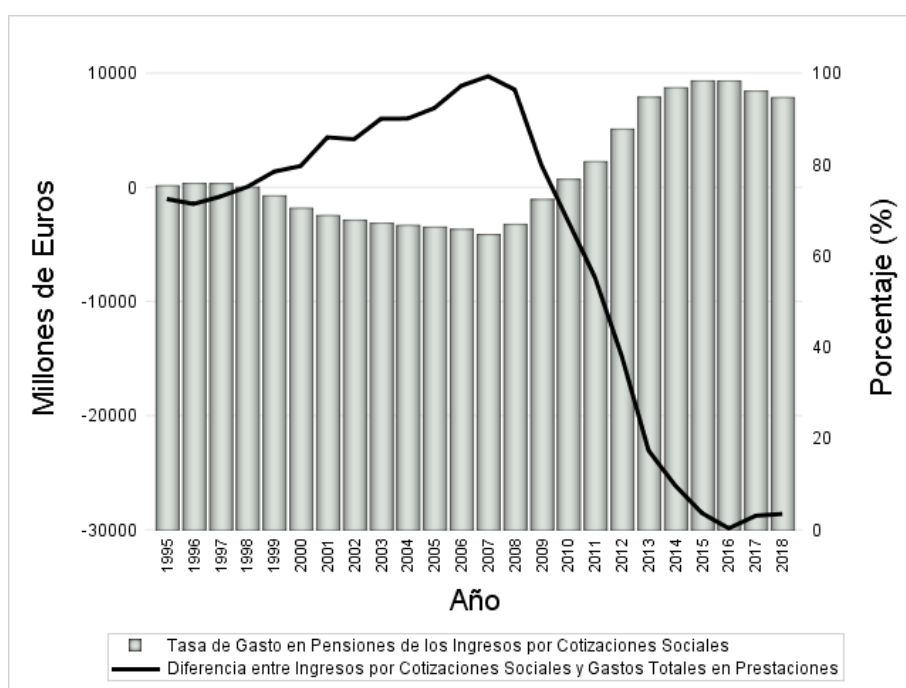
Gráfico 1.7. Serie de Ingresos y Gastos de la Seguridad Social (1995-2018)



Datos: Seguridad Social

Esta disparidad entre gastos e ingresos se refleja de forma clara y evidente en los dos indicadores que vemos en el Gráfico 1.8.

Gráfico 1.8. Tasa de Gasto en Pensiones de las Cotizaciones Sociales vs Diferencia entre Ingresos y Gastos (1995-2018)



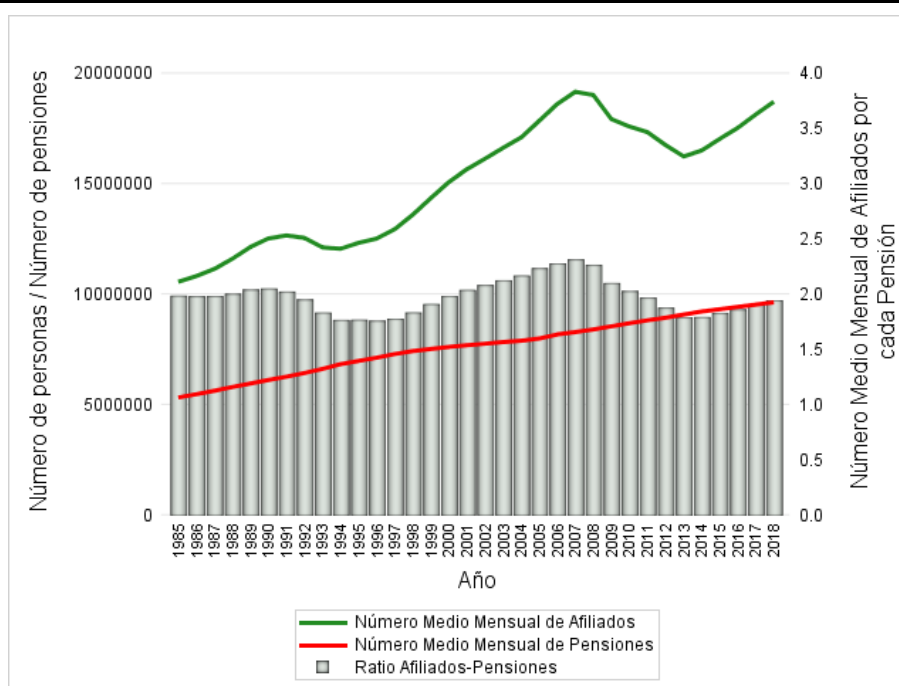
Datos: Seguridad Social

En primer lugar, la diferencia entre los ingresos por cotizaciones sociales y los gastos derivados de todas las prestaciones económicas que abarca la seguridad social pasó de una situación de superávit a otra de déficit en el año 2009. Y, a día de hoy, todavía está muy lejos de poder recuperar el equilibrio financiero.

En segundo lugar, el cociente entre el gasto en pensiones y los ingresos por cotizaciones sociales nos muestra que actualmente casi el 100% de lo cotizado por empleados y empleadores se destina a cubrir el gasto en pensiones. Esto deja muy poco margen para poder abordar otros grandes gastos que cubre nuestro sistema de seguridad social.

Por último, es esencial subrayar que nuestro sistema de reparto necesita de la existencia de un equilibrio entre el número de afiliados o contribuyentes al sistema y el número de beneficiarios o pensionistas. Una buena medida de esta armonía es la ratio entre el número de afiliados y el número de pensiones. Este indicador nos permite saber cuántos cotizantes soportan los gastos derivados de cada pensión. En la actualidad, existe cierto consenso entre los economistas sobre que la ratio afiliados-pensiones debe ser superior a 2 para que el sistema pueda ser sostenible.

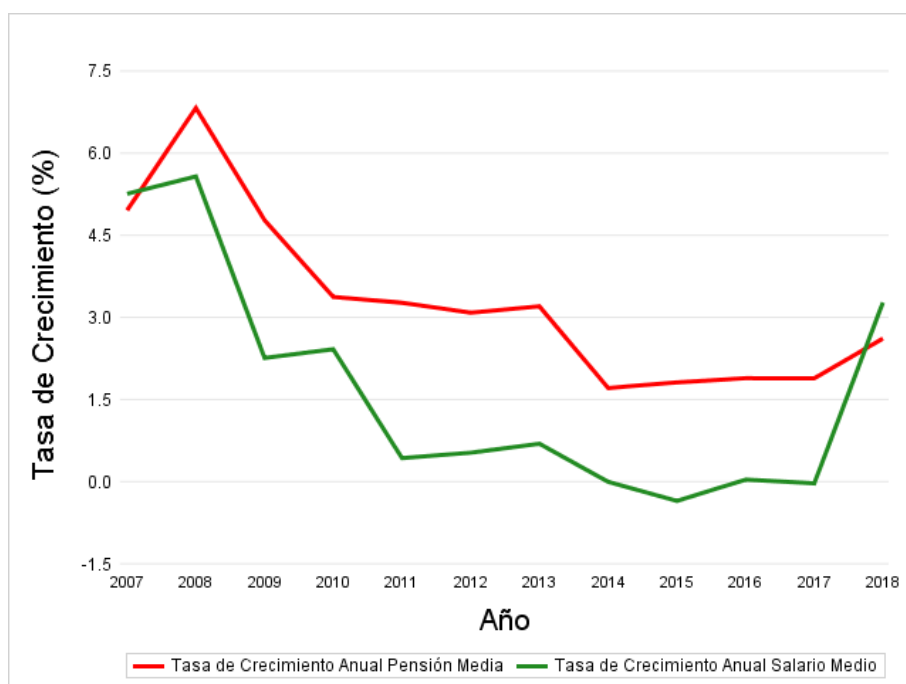
Gráfico 1.9. Número Medio Mensual de Pensiones y Afiliados (1995-2018)



Datos: Seguridad Social

En estas décadas hemos pasado de tener una ratio cercana a 4 trabajadores por pensión en los años 70, a menos de 2 trabajadores por pensión. Además, en los últimos años la cuantía la pensión media no ha parado de incrementarse a un ritmo mucho mayor que el del salario medio.

Gráfico 1.10. Tasa de Crecimiento Anual de la Pensión Media y del Salario Medio (2007-2018)



Datos: INE y Seguridad Social

En resumen, el impacto de las medidas tomadas en las últimas décadas no parece suficiente para compensar el incremento de la esperanza de vida, la disminución de la inmigración, la caída de la natalidad y el deterioro general de la economía y el empleo producido por la Gran Recesión.

Además, a todo lo anterior se suma la actual crisis en la que estamos inmersos debido al COVID-19, la cual está golpeando todos los sectores socioeconómicos y cuyos efectos son todavía difíciles de evaluar.

Luego, a la luz de los datos expuestos, parece evidente que existe la necesidad de estudiar en profundidad la sostenibilidad de nuestro sistema de pensiones.

2. ESTADO DEL ARTE

2.1. Introducción

Los sistemas públicos de pensiones de reparto han sido objeto de multitud de críticas desde su nacimiento. Desde una perspectiva ética o filosófica, muchos consideran que la obligatoriedad del sistema limita la libertad del individuo. También desde el punto de vista de la política, se señalan las complicaciones que provoca la acumulación de más poder por parte del Estado. Pero las críticas más feroces han sido desde la economía, argumentando sus problemas de eficacia y sostenibilidad.

En este trabajo no nos detendremos en evaluar los fundamentos filosóficos del sistema o si es el más eficiente de todos los posibles, tan solo estudiaremos si es económicamente viable en el tiempo. Por lo tanto, en este apartado haremos una breve recopilación de la información publicada por otros investigadores con el objetivo de establecer un conocimiento sobre cómo abordar el análisis de la sostenibilidad de los sistemas públicos de pensiones.

2.2. Enfoques de Estudio de la Sostenibilidad de un Sistema de Pensiones

Desde la publicación “*Averting the old crisis: policies to protect the old and promote the growth*” (World Bank, 1994), cada vez han aparecido más investigaciones que analizan la robustez de los sistemas de pensiones. Muchos de estos estudios se han centrado en analizar un tipo particular de sistemas de pensiones, los conocidos como sistemas de pensiones reparto.

Según Devesa et Devesa (2009), la mayor parte de estos trabajos se centran en uno de los siguientes tipos de estudio:

- Flujos de pensiones. En este método se presupone que los ingresos por cotizaciones son siempre constantes a un determinado porcentaje del PIB (Producto Interior Bruto). Por lo que lo único que se tiene que estimar son los gastos en pensiones.
- Flujos de pensiones y cotizaciones. Además de estimar cuál es el flujo de gasto en pensiones, también se calcula la proyección de ingresos por cotizaciones.
- Deuda implícita. Esta metodología se centra en determinar cuál es la deuda que ha contraído el sistema ante sus actuales pensionistas, y también ante sus actuales afiliados, por el hecho de que desde ese mismo momento están adquiriendo un derecho a recibir en el futuro unas prestaciones. Luego trata de determinar la cantidad que debería de tener el sistema para poder hacer frente a sus compromisos futuros.
- Tasa Interna de Rendimiento (TIR). Trata de determinar cuál es la rentabilidad que el sistema da a los participantes en el mismo. En este caso, la viabilidad del sistema viene dada por la relación entre el valor obtenido y el nivel máximo de rentabilidad que podría ofrecer teóricamente el sistema, que, a su vez, está relacionado con el crecimiento del PIB.

Una variante de este procedimiento fue propuesta por Brío Carretero (2007), a través de la cual se modela el sistema de la Seguridad Social mediante una cola de cotizantes y otra de pensionistas. Luego se utiliza la Teoría de Colas para calcular los valores medios de permanencia en el sistema, pensiones medias, etcétera.

- Coste de la pensión unitaria. En este caso, se trata de estimar cual es el coste para el sistema de entregar una unidad monetaria de pensión. Evidentemente, el sistema se halla en situación de desequilibrio cuando el coste es superior a uno.

Los dos primeros métodos se denominan “Métodos financieros”, y, debido a su sencillez y potencia, han sido los más utilizados para estudiar la viabilidad del sistema de pensiones en España.

El resto de los métodos se encuadran dentro de los “Métodos financiero-actuariales”, pues emplean la técnica actuarial. La principal desventaja de estos últimos se encuentra en la complejidad de los cálculos a la hora de abordar el solapamiento de múltiples generaciones, por lo que al final solo se aplican sobre sistemas cerrados.

Nosotros en este trabajo nos centraremos en los métodos financieros. La principal virtud de esta metodología es que podemos saber cuál es la situación en cada instante. Dentro de este enfoque existen multitud de técnicas a utilizar, desde la regresión más simple a métodos muy sofisticados.

Unos estudios interesantes son aquellos en los que se utilizan microsimulaciones, es decir, se crean modelos usando información a nivel de agente económico individual (personas, hogares, empresas, etcétera).

En España, un buen ejemplo de este tipo de trabajos es el publicado por la AIREF (Autoridad Independiente de Responsabilidad Fiscal) y realizado por Pastor et Vila (2019). En este artículo se utilizan los microdatos de la MCVL¹ (Muestra Continua de Vidas Laborales) para estimar un modelo con todas las interacciones entre los componentes demográficos y macroeconómicos. Una vez determinado el modelo, pronostican cual será la demografía y el PIB a largo plazo, y, con ello, la proporción sobre el PIB del gasto en Seguridad Social.

¹ La MCVL es un conjunto de microdatos individuales anonimizados procedente de las bases de datos de la Seguridad Social, además de otros que se toman del Registro municipal continuo (INE) y del resumen anual de retenciones e ingresos por Impuesto a la Renta Personal (Modelo 190) de la AEAT (Agencia Tributaria Española).

Este conjunto de microdatos permite reconstruir la historia laboral de los individuos extraídos de la muestra en un año determinado, es decir, el 4% de los individuos que tuvieron relación con la Seguridad Social (afiliados en alta laboral, perceptores de prestaciones o subsidios por desempleo y perceptores de pensiones contributivas de la Seguridad Social).

En otros países de Europa también se han llevado a cabo investigaciones muy similares, como la realizada en 2018 por el *Regeringskansliet* (Gobierno de Suecia) para estimar el comportamiento del sistema de pensiones sueco hasta 2070.

Pese al indudable valor de los métodos que realizan microsimulaciones, nosotros profundizaremos en los trabajos que utilizan datos de series de agregados macroeconómicos y demográficos, pues serán los datos que usaremos a lo largo de todo este estudio.

La metodología que utiliza datos de contabilidad agregada o de crecimiento para obtener la proyección del gasto en pensiones tiene una larga tradición en la literatura. Como vemos en el trabajo de Doménech et Melguizo (2009), la principal ventaja de estos métodos es que, en base a las proyecciones de un número relativamente pequeño de variables, es simple obtener unas predicciones razonablemente buenas sobre el gasto en pensiones como parte del PIB.

También es posible comprobar la sensibilidad del modelo, utilizando diferentes escenarios demográficos y económicos. Estos escenarios se consiguen modificando los flujos de inmigración, tasas de mortalidad, fertilidad, productividad, etcétera; y se clasifican habitualmente en escenario estándar, escenario favorable y escenario desfavorable.

Para terminar vamos a destacar el trabajo de Tian et Zhao (2016), donde se centran en realizar un pronóstico de la sostenibilidad financiera de las pensiones en China utilizando: la matriz de Leslie¹ combinada con simulaciones de Monte Carlo para calcular la evolución de la población; y un modelo de series temporales ARMA² para las predicciones de las tasas de fertilidad y mortalidad de los distintos grupos de edad.

¹ El modelo matricial de Leslie (Leslie, 1945) se usa frecuentemente para describir la dinámica de una población estructurada por grupos de edad. La definición general es:

$$\begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_{i-1} \\ x_i \end{bmatrix}_{t+1} = \begin{bmatrix} f_1 & f_2 & \dots & f_i \\ s_1 & 0 & \dots & 0 \\ 0 & s_2 & \dots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & s_{i-1} & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_{i-1} \\ x_i \end{bmatrix}_t,$$

donde f_i representa la tasa de fecundidad o reproductividad de la clase i , s_i representa la tasa de supervivencia de la clase i a la clase $i+1$, cada elemento del vector $\mathbf{X}_t$ es el número de individuos de cada clase i en el instante t y cada elemento del vector $\mathbf{X}_{t+1}$ es el número de individuos de cada clase i en el instante $t+1$.

² El análisis de series temporales mediante la metodología Box-Jenkins se explica con detenimiento en el apartado 4.

3. HIPÓTESIS Y OBJETIVOS

3.1. Hipótesis

La hipótesis de partida es que existe una relación subyacente en los datos de las series temporales de gasto en pensiones como porcentaje del PIB que puede ser modelizada mediante técnicas estadísticas. Y, por tanto, podemos pronosticar, con diferentes grados de confianza, el comportamiento de la serie en el futuro.

3.2. Objetivo Principal del Estudio

La finalidad principal de este estudio es comprobar si el sistema de pensiones, tal y como está estructurado actualmente, es viable en un horizonte temporal de 10 años. Para ello calcularemos el porcentaje del PIB que supone el gasto en pensiones durante los años 2020-2030.

3.3. Objetivos Secundarios del Estudio

Entre los propósitos secundarios están:

- Comparar distintas metodologías para predecir series temporales, tanto univariantes como multivariantes, y determinar cuál es la que proporciona mejores resultados.
- Estimar el posible impacto que tendrá la crisis del COVID-19 en nuestro sistema de pensiones.

4. MATERIAL Y METODOLOGÍA

4.1. Ámbito y Periodo del Estudio

El trabajo se ha desarrollado en el ámbito del territorio español.

Los datos pertenecen al periodo 1995-2020 y como unidad de tiempo básica se ha seleccionado el mes.

4.2. Fuentes de Datos

Los datos para poder llevar a cabo los objetivos de este estudio se han seleccionado siguiendo, esencialmente, tres criterios: su disponibilidad, su fiabilidad y su potencial capacidad explicativa.

Teniendo en cuenta el primer y el segundo criterio se han empleado bases de datos procedentes de Eurostat, el Instituto Nacional de Estadística y la Seguridad Social. El uso de fuentes de datos que han sido elaboradas por instituciones que se encuentran bajo el paraguas del Sistema Estadístico Europeo (SEE) garantiza que los datos son fiables y comparables, ya que han sido tratados de manera adecuada a lo largo de todo el proceso.

En cuanto al tercer criterio, nos guiaremos por los trabajos publicados por otros investigadores y por nuestra intuición. Como hemos visto en las secciones anteriores los flujos de inmigración, las tasas de mortalidad y fecundidad, el nivel de empleo y afiliación, la estructura de la población, etcétera juegan un papel determinante en las fluctuaciones a lo largo del tiempo del porcentaje del PIB destinado al gasto en pensiones.

Por todo lo anterior las variables que se han incluido en este estudio son:

- Número de pensiones pagadas.
- Valor de la pensión media a precios corrientes.
- Número de inmigrantes llegados a España por edad.
- Producto Interior Bruto de España a precios corrientes.
- Número de habitantes en España por edad y sexo.
- Número de defunciones por cada mil habitantes en España por edad y sexo.
- Número de personas afiliadas a la Seguridad Social.

El motivo principal de no introducir los datos relativos a la natalidad es que el impacto de los nuevos nacimientos solo será de utilidad en estudios a más largo plazo (20 o más años).

4.3. Definiciones

Como paso previo al análisis y la construcción de modelos, es esencial definir un indicador que pueda medir fehacientemente el grado de sostenibilidad del sistema de pensiones y un determinado umbral en el que establecer si el sistema es o no viable.

La variable indicadora de la sostenibilidad será el porcentaje de PIB que se dedica a cubrir el gasto en pensiones, que, a partir de ahora, denominaremos PGP (Porcentaje de Gasto en Pensiones). Esta variable queda definida por la siguiente expresión:

$$PGP_t = \frac{GP_t}{PIB_t} = \frac{PM_t \times NP_t}{PIB_t}$$

donde GP_t representa el gasto en pensiones en el instante t, PIB_t representa el Producto Interior Bruto en el instante t, PM_t representa la pensión media en el instante t y NP_t representa el número de pensiones en el instante t.

Puesto que las únicas variables que se miden en unidades monetarias son el PIB y la pensión media, a lo largo del estudio usaremos precios corrientes. La razón es obvia, ya que dividir numerador y denominador por el deflactor nos daría los mismos resultados.

El establecimiento de un umbral a partir del cual determinar la sostenibilidad o no de nuestro sistema de pensiones es un asunto más complejo. En la literatura no existe un consenso, por lo que tomaremos esta decisión en base a los datos de España y de países de nuestro entorno.

Tabla 4.1. Porcentaje del PIB Destinado a Pensiones (2015)	
País	% PIB
Grecia	17.8
Italia	16.5
Francia	15.0
Portugal	14.9
Austria	14.6
Dinamarca	13.5
Finlandia	13.1
Países Bajos	13.0
UE-28	12.8
Bélgica	12.7
España	12.6
Alemania	11.8
Polonia	11.8
Suecia	11.4
Reino Unido	11.4
Eslovenia	10.9
Datos: EUROSTAT	

Observamos que el gasto medio en los países de la UE en pensiones se situó en 2015 en el 12.8% del PIB, para España los datos son muy similares a la media europea (tan solo unas décimas por debajo). Dentro de ese 12.6% del PIB que gasto nuestro país en pensiones se engloban diferentes tipos de pensiones, como las de la Seguridad Social (aproximadamente el 9.1% del PIB) o las de las clases pasivas (alrededor del 1.2% del PIB).

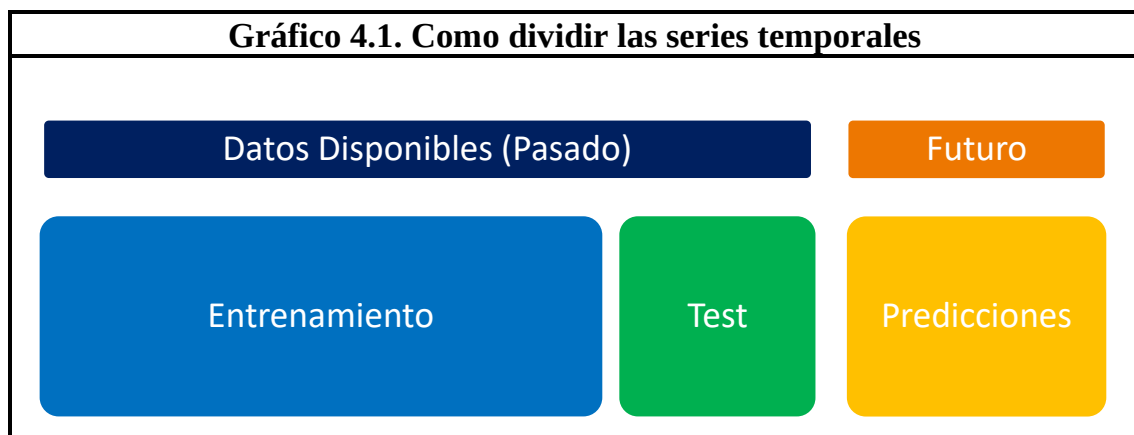
Teniendo en cuenta que los datos que manejamos en este estudio son los relativos a las pensiones pagadas por la Seguridad Social, podemos establecer a través de una relación de proporcionalidad que el umbral de gasto en pensiones para nuestro estudio debería situarse como mucho en una banda entre el 13% y el 15% del PIB. Ya que, en caso de superar ese umbral, el gasto total de España en pensiones se situaría muy por encima del resto de países de la UE.

En resumen, determinamos que el sistema se encuentra en una situación sostenible mientras nuestras predicciones de gasto del PIB en pensiones estén por debajo del 13%; en situación de riesgo cuando estén entre el 13% y el 15%; y en situación claramente insostenible cuando estén por encima del 15%.

4.4. Pasos a Seguir en el Análisis de los Datos

En primer lugar, uniremos las distintas series temporales en un único conjunto de datos y calcularemos las variables que sean necesarias. En muchos de los casos para poder trabajar tendremos que realizar un procesamiento de las series que incluye: su desestacionalización, ya que es fundamental para poder analizar de manera correcta su evolución y tendencia a lo largo del tiempo; y la interpolación de datos faltantes, ya que algunas de las series estadísticas disponibles no tienen la periodicidad adecuada (trimestrales, semestrales o anuales).

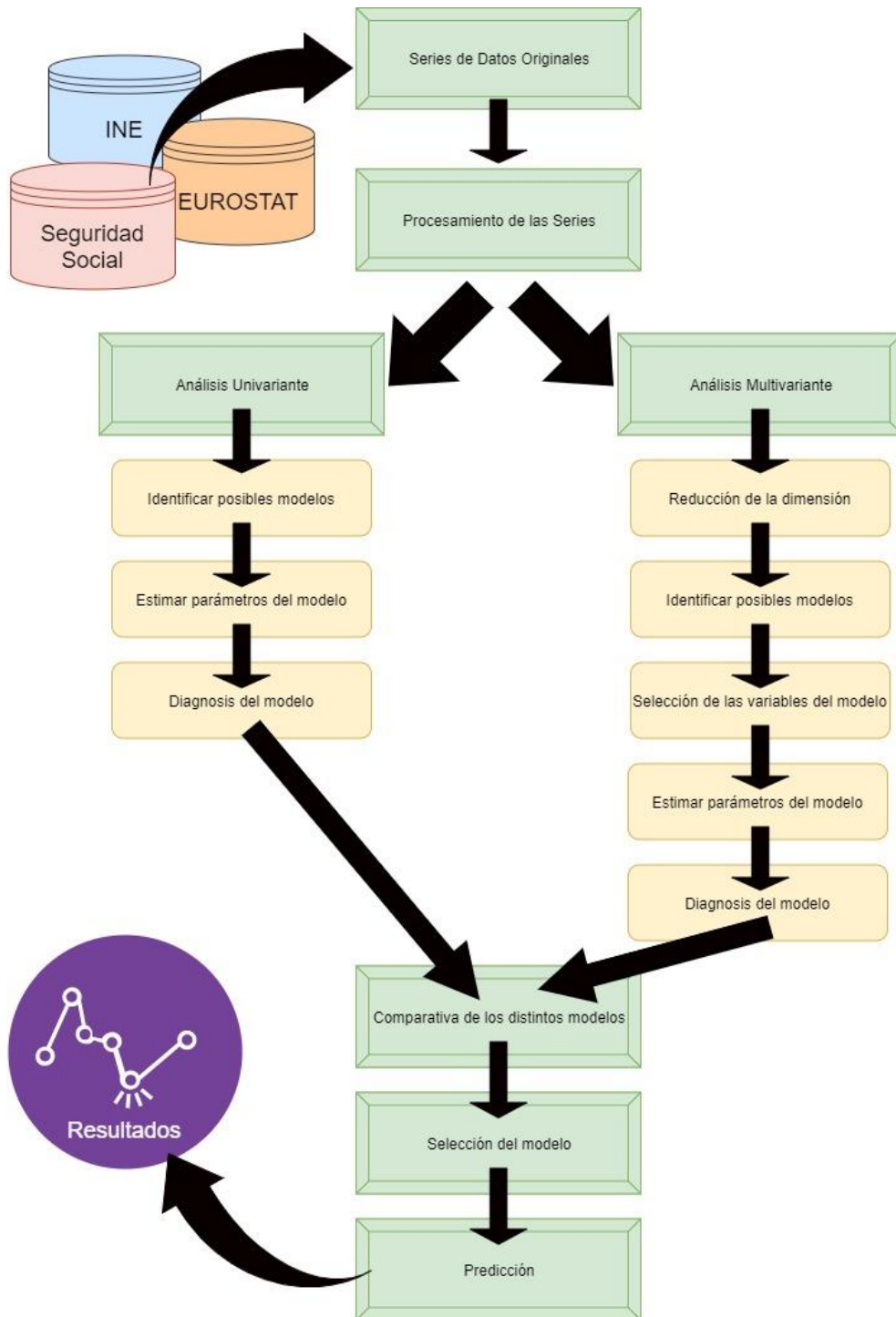
A continuación, procederemos a la parte del análisis y la creación de modelos. En este apartado usaremos dos tipos de metodología: univariante y multivariante. Un paso esencial en la fase de análisis es el de dividir nuestro conjunto de datos en dos partes: una para la estimación de los parámetros (entrenamiento) y otra para comprobar la calidad de nuestras predicciones (test).



Cuando ya tengamos nuestros modelos estimados y validados procederemos a ver cuál es el que nos proporciona una mejor aproximación a la realidad. Seleccionaremos el modelo que nos proporcione unas mejores predicciones.

Por último, usaremos todas las observaciones de las que disponemos para predecir el comportamiento del porcentaje de PIB que se destina a cubrir el gasto en pensiones. Con estas últimas predicciones veremos si el sistema es o no sostenible en el tiempo.

Gráfico 4.2. Esquema general de trabajo



4.5. Técnicas Estadísticas Utilizadas

Para llevar a cabo el análisis de nuestra base de datos, vamos a utilizar diversas técnicas estadísticas.

- **4.5.1. Análisis Univariante**

- **4.5.1.1. Métodos de Suavizado o Alisado Exponencial**

Los métodos de suavizado o alisado exponencial fueron desarrollados por Brown, Holt y Winters entre 1950 y 1960. Estos modelos ajustan la evolución de la serie calculando cada nueva observación como una media ponderada de las anteriores, considerando que los pesos de las mismas se reducen exponencialmente. Pese a su simplicidad, varios autores han demostrado que pueden tener mejor desempeño que otros modelos más complejos, particularmente cuando los datos no son normales.

Existen multitud de métodos de suavizado, pero los más conocidos y usados son el método de alisado simple, el método de alisado doble de Holt, el método aditivo de Holt-Winters y el método multiplicativo de Holt-Winters. Estos dos últimos serán los que utilizaremos nosotros y los que detallamos a continuación.

El método de Holt-Winters se puede aplicar en series temporales con tendencia lineal y estacionalidad. El modelo aditivo y el multiplicativo se definen matemáticamente de la siguiente forma:

Holt-Winters Aditivo	Holt-Winters Multiplicativo
$x_t = (L_t + b_t) + S_t + \epsilon_t$	$x_t = (L_t + b_t) \times S_t + \epsilon_t$
donde	donde
$L_t = \alpha(x_t - S_{t-s}) + (1 - \alpha)(L_{t-1} + b_{t-1})$	$L_t = \alpha \frac{x_t}{S_{t-s}} + (1 - \alpha)(L_{t-1} + b_{t-1})$
$b_t = \beta(L_t - L_{t-1}) + (1 - \beta)b_{t-1}$	$b_t = \beta(L_t - L_{t-1}) + (1 - \beta)b_{t-1}$
$S_t = \gamma(x_t - L_t) + (1 - \gamma)S_{t-s}$	$S_t = \gamma \frac{x_t}{L_t} + (1 - \gamma)S_{t-s}$
$\epsilon \equiv N(0, \sigma)$	$\epsilon \equiv N(0, \sigma)$

Los tres parámetros α , β y γ toman valores que oscilan entre 0 y 1.

Las predicciones para estos modelos se calculan como:

Holt-Winters Aditivo

$$\hat{x}_{t+1} = (L_t + b_t) + S_{t+1-s}$$

Holt-Winters Multiplicativo

$$\hat{x}_{t+1} = (L_t + b_t) \times S_{t+1-s}$$

Tan solo señalaremos que en estos modelos cuando el valor de los parámetros está próximo a 0, indica que las predicciones se modifican muy poco con la nueva información, por lo que son muy similares entre sí (mucha memoria). Por el contrario, cuando toman un valor cercano a 1 la predicción se va adaptando según el valor que toman las últimas observaciones, por lo que se dice que tiene poca memoria (los valores alejados prácticamente no tienen influencia en la predicción).

➤ 4.5.1.2. Metodología Box-Jenkins

La publicación de “*Time Series Analysis: Forecasting and Control*” por Box y Jenkins a comienzos de los 70 supuso la integración del conocimiento previo existente sobre la noción de estocasticidad en series de tiempo, así como la creación de una metodología coherente sobre cómo abordar la identificación, estimación y validación de los modelos.

La caracterización principal de este enfoque es la consideración de que la serie temporal viene determinada por la realización de un proceso estocástico, siendo esta una realización particular del mismo. Sobre la base de esta simple idea se han ido desarrollando multitud de modelos: autorregresivos (AR), de medias móviles (MA), mixtos autorregresivos-medias móviles (ARMA) y autorregresivos integrados de medias móviles (ARIMA).

La estructura de los modelos AR, MA y ARMA solo permite ser usada en series de tiempo estacionarias¹, pero la mayoría de las series no lo son. Gracias a la introducción de los modelos ARIMA se realiza la integración o diferenciación que permite abordar series no estacionarias.

Se dice que un proceso z_t es $ARIMA(p, d, q)(P, D, Q)_s$ cuando p es el orden de la parte autorregresiva, d el orden de integración, q el orden de la parte móvil, P el orden de la parte autorregresiva periódica o estacional, D el orden de integración de la parte periódica o estacional, Q el orden de la parte móvil periódica o estacional y s es la periodicidad de la serie². Matemáticamente se define como:

$$\Phi_P(L^s)\phi_p(L)(1 - L^s)^D(1 - L)^d z_t = c + \Theta_Q(L^s)\theta_q(L)\epsilon_t$$

donde:

¹ En términos prácticos, una serie es estacionaria (en el sentido débil) cuando la media, la varianza y la covarianza permanecen constantes a lo largo del tiempo.

² La periodicidad de la serie viene determinada por la naturaleza de los datos (días, meses, trimestres, etcétera), luego no hay que calcular s .

$\Phi_P(L^S) = (1 - \Phi_1 L^S - \dots - \Phi_P L^{SP})$ es el operador AR estacional de orden P,

$\phi_p(L) = (1 - \phi_1 L - \dots - \phi_p L^p)$ es el operador AR regular de orden p,

$\Theta_Q(L^S) = (1 - \Theta_1 L^S - \dots - \Theta_Q L^{SQ})$ es el operador MA estacional de orden Q,

$\theta_q(L) = (1 - \theta_1 L - \dots - \theta_q L^q)$ es el operador MA regular de orden q,

$(1 - L^S)^D$ son las diferencias estacionales,

$(1 - L)^d$ son las diferencias regulares y

ϵ_t es un proceso de ruido blanco.

También es conveniente definir el operador polinomial de retardos L, que aplicado a la variable z_t se define de la siguiente manera:

$$L(z_t) = z_{t-1}$$

$$L^k(z_t) = z_{t-k}$$

$$(1 - L)z_t = z_t - z_{t-1}$$

• 4.5.2. *Análisis Multivariante*

Primeramente, debemos asegurarnos de que no existen problemas debido a que existan variables que son combinación lineal directa de otras. Una solución sencilla para este problema consiste en eliminar alguna de las variables que sea combinación lineal de las otras.

Una vez solucionado el problema anterior, debemos comprobar si existen variables que tengan información redundante, es decir, si hay una relación de dependencia lineal fuerte entre algunas de las variables predictoras (Multicolinealidad). En caso afirmativo, procederemos a realizar una reducción de la dimensión mediante componentes principales.

El último paso será utilizar los datos resultantes de la reducción de la dimensión para poder estimar el comportamiento de la serie multivariante.

➤ 4.5.2.1. *Análisis de Componentes Principales*

El Análisis de Componentes Principales describe la variación de un conjunto de datos multivariante en términos de un conjunto incorrelado de variables, cada una de las cuales es una combinación lineal particular de las variables originales.

Estas nuevas variables reciben el nombre de Componentes Principales y se obtienen en orden decreciente a su importancia. Esto quiere decir que la primera Componente Principal recoge la máxima información de los datos originales, y como consecuencia, la

segunda Componente recoge la mayor cantidad de información que no haya sido recogida por la primera Componente y así sucesivamente con el resto de cada una de las Componentes Principales.

Debido a que las Componentes Principales se utilizan para resumir los datos originales con la mínima pérdida de información, esto dará lugar a importantes simplificaciones en los análisis posteriores que vamos a realizar. Elegiremos el número de Componentes en función de la proporción de variabilidad explicada, de manera que dicha proporción se considere suficiente (con resultados de encuestas reales un 70% es suficiente, si son pruebas de laboratorio se debe ser más exigente).

Desde un punto de vista geométrico las Componentes Principales representan la selección de un nuevo sistema de coordenadas obtenido mediante la rotación del sistema original. Los nuevos ejes representan las direcciones con máxima variabilidad y producen una descripción más simple y ordenada de la estructura de la matriz de Varianzas-Covarianzas.

Luego, partiendo de la matriz de Varianzas-Covarianzas, se hallará su descomposición en función de sus valores propios y la matriz formada por sus autovectores correspondientes. Matemáticamente queda expresado de la siguiente forma:

$$\Sigma = P\Lambda P^{-1}$$

siendo P la matriz que contiene a los autovectores y Λ una matriz diagonal formada por los autovalores.

De este modo, la primera componente principal será aquella variable cuyo vector dirección coincide con el del autovector correspondiente al mayor autovalor de la matriz de Varianzas-Covarianzas.

Finalmente hay que destacar que cuando las variables no se midan en unidades semejantes, se aplicará el análisis sobre la matriz de correlaciones (en vez de sobre la matriz de Varianzas-Covarianzas).

➤ 4.5.2.2. Modelos Vectoriales

La generalización de la metodología Box-Jenkins a las series temporales multivariantes no tardó mucho en llegar. Los procesos vectoriales autorregresivos de medias móviles (VARMA) fueron descritos por primera vez por Quenouille en 1957, aunque el software para poder implementarlos solo estaría disponible a principios de los años 80. Fue a partir de esa década cuando el uso de este tipo de modelos se convirtió en la piedra angular de la predicción macroeconómica, gracias a la publicación de más trabajos teóricos y prácticos.

Los primeros modelos que se utilizarían fueron los modelos vectoriales autorregresivos (VAR), que constituyen un caso particular de los VARMA. A pesar de que la teoría económica no está de acuerdo con el modelado de procesos que usan modelos VAR (ya que los términos de medias móviles no pueden excluirse), la dificultad de la metodología VARMA impulsó la selección de modelos VAR.

En esencia, un modelo $VAR(p)$ se define de la siguiente forma:

$$\mathbf{z}_t = \mathbf{A}_1 \mathbf{z}_{t-1} + \cdots + \mathbf{A}_p \mathbf{z}_{t-p} + \boldsymbol{\epsilon}_t$$

donde

$$\begin{aligned} \mathbf{z}_t &= (z_{1t}, z_{2t}, \dots, z_{kt}) \forall k = 1, \dots, n \quad (\text{Vector de variables}) \\ \mathbf{A}_i &\in \mathcal{M}_{k \times k} \forall i = 1, \dots, p \wedge k = 1, \dots, n \quad (\text{Matriz de coeficientes}) \\ \mathbb{E}[\boldsymbol{\epsilon}_t] &= 0 \wedge \mathbb{E}[\boldsymbol{\epsilon}_t \boldsymbol{\epsilon}_t'] = \Sigma_{\epsilon} \text{ definida positiva} \quad (\text{Ruido blanco } k - \text{dimensional}) \end{aligned}$$

Por tanto, la expresión matricial de un modelo $VAR(1)$ con un vector de dos variables es:

$$\begin{bmatrix} z_{1,t} \\ z_{2,t} \end{bmatrix} = \begin{bmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{bmatrix} \begin{bmatrix} z_{1,t-1} \\ z_{2,t-1} \end{bmatrix} + \begin{bmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \end{bmatrix}$$

En general, los modelos VAR proporcionaban unos resultados bastantes buenos y supusieron una mejora sobre los modelos anteriores. Con el tiempo se descubriría que tienden a sufrir un ligero sobreajuste (proporcionan predicciones deficientes fuera de la muestra, aunque el ajuste sea muy bueno dentro de la muestra). A pesar de ello se mantuvieron como uno de los principales modelos de referencia durante casi dos décadas.

La falta de interés en los modelos multivariantes de medias móviles se debió principalmente a la dificultad a la hora de implementarlos. Sería necesario una nueva ampliación de la metodología y, sobre todo, de la capacidad de cómputo de los ordenadores para que se extendiera su uso a finales de los 90.

Utilizando una mixtura de la notación usada para describir los modelos univariantes $ARIMA(p, d, q)(P, D, Q)_s$ y los multivariantes $VAR(p)$, tenemos que un modelo $VARMA(p, q)$ se define matemáticamente como:

$$\mathbf{z}_t - \boldsymbol{\phi}_1 \mathbf{z}_{t-1} - \cdots - \boldsymbol{\phi}_p \mathbf{z}_{t-p} = \mathbf{c} + \boldsymbol{\epsilon}_t - \boldsymbol{\theta}_1 \boldsymbol{\epsilon}_{t-1} - \cdots - \boldsymbol{\theta}_q \boldsymbol{\epsilon}_{t-q}$$

Estos modelos también permiten la introducción de variables ficticias o *dummy* para poder modelar la estacionalidad (*seasonal factors*). Por ejemplo, con datos trimestrales el modelo se especificaría de la forma:

$$\mathbf{z}_t - \boldsymbol{\phi}_1 \mathbf{z}_{t-1} - \cdots - \boldsymbol{\phi}_p \mathbf{z}_{t-p} = \mathbf{c} + \mathbf{D}_{Q1} \boldsymbol{\beta}_{Q1} + \cdots + \mathbf{D}_{Q3} \boldsymbol{\beta}_{Q3} + \boldsymbol{\epsilon}_t - \boldsymbol{\theta}_1 \boldsymbol{\epsilon}_{t-1} - \cdots - \boldsymbol{\theta}_q \boldsymbol{\epsilon}_{t-q}$$

donde:

- $\mathbf{c}$ es el vector correspondiente a la parametrización del cuarto trimestre (Q4).
- $\mathbf{D}_{Qi} \forall i = 1, \dots, 3$ son matrices dummy con todos sus valores iguales a 0 salvo que el trimestre t sea el correcto. Es decir, si el trimestre t es Q1 entonces $\mathbf{D}_{Q1}$ es la matriz identidad, en otro caso es una matriz nula.
- $\boldsymbol{\beta}_{Qi} \forall i = 1, \dots, 3$ son vectores que contienen los parámetros de los efectos de cada uno de los trimestres.

Por último, destacamos que los modelos multivariantes que hemos expuesto asumen que las series son simultáneamente estacionarias.

• 4.5.3. Variables de Intervención

En ciertas ocasiones los modelos no son capaces de obtener unas predicciones adecuadas debido a la existencia de un hecho extraordinario que impacta sobre nuestra serie temporal. Dentro de este tipo de actos singulares se encuentran las huelgas, un accidente o una crisis financiera. Para que nuestros modelos puedan seguir prediciendo con calidad necesitamos incluir más información a través de las variables de intervención.

Las principales variables para describir los efectos anómalos son:

- Impulso: es siempre cero, excepto en el instante F , donde toma el valor uno.
- Escalón: toma el valor cero antes del instante F y uno posteriormente.
- Tendencia: toma el valor cero antes del instante F y otro valor distinto de cero a partir de entonces (habitualmente valores consecutivos).

Matemáticamente se definen:

$$I_t = \begin{cases} 0 & \text{si } t \neq F \\ 1 & \text{si } t = F \end{cases}$$

$$E_t = \begin{cases} 0 & \text{si } t < F \\ 1 & \text{si } t \geq F \end{cases}$$

$$T_t = \begin{cases} 0 & \text{si } t < F \\ r & \text{si } t \geq F \end{cases} \forall r \in \mathbb{R}$$

Para poder utilizar esta técnica es necesario que el analista conozca que tipo de intervención es necesaria, el tipo de efecto que producirá y, sobre todo, su ubicación temporal.

4.6. Software

Para realizar todo este estudio se ha utilizado SAS Software (versión 9.4, SAS Institute, Cary, NC, USA), y en especial los módulos específicos para analizar series temporales.

5. ANÁLISIS Y RESULTADOS

5.1. Procesamiento de las Series

Comenzaremos creando nuestro conjunto de datos. Siempre que estén disponibles usaremos series que ya hayan sido desestacionalizadas previamente, y en caso contrario las desestacionalizaremos nosotros mediante el procedimiento X11 de SAS. Este procedimiento permite la eliminación de la componente estacional, lo que facilita la comparación entre meses o trimestres consecutivos de una serie económica.

Otro problema a la hora de unir las series es la diferencia entre sus intervalos de tiempo (meses, trimestres, semestres o años). Como ya dijimos anteriormente, nosotros tomaremos el mes como nuestra unidad temporal básica, luego debemos interpolar los datos de algunas de nuestras series para expandir su frecuencia. Por ello utilizaremos el procedimiento EXPAND que está específicamente diseñado para este propósito. El único cuidado que debemos de tener al expandir la frecuencia de nuestras series es entender que tipo de dato es el que estamos tratando, ya que el procedimiento es distinto cuando los datos son agregados de un periodo, como el PIB trimestral o los flujos de inmigración anual; si son un promedio del periodo, como las tasas de mortalidad; o si son datos puntuales, como la población de hombres y mujeres en un determinado momento.

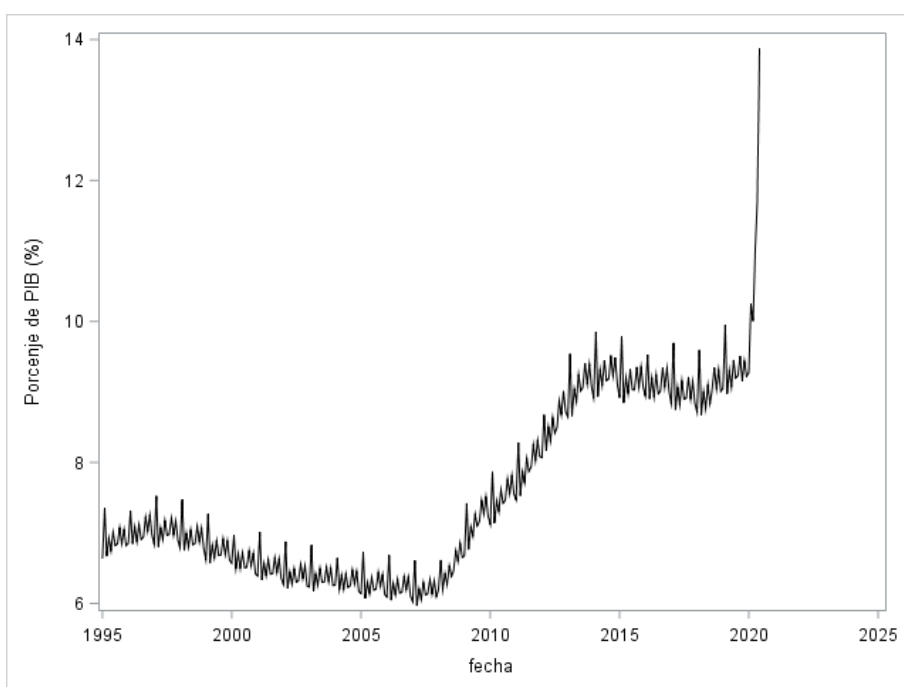
Tabla 5.1. Series de Datos		
Series	Fuente	Frecuencia Original
Afiliados	Seguridad Social	Mensual
Pensiones	Seguridad Social	Mensual
Inmigración	Eurostat	Anual
Mortalidad	INE	Anual
PIB Desestacionalizado	INE	Trimestral
Población	INE	Semestral

Una vez que ya tenemos todas nuestras variables desestacionalizadas y con la misma frecuencia procederemos al cálculo de una nueva variable: el porcentaje de gasto del PIB en pensiones (PGP).

En el Grafico 5.1 vemos que antes del 2009 el PGP se mantuvo en una horquilla de entre el 6% y el 8%; la crisis económica hizo que pasara a moverse entre el 8% y el 10%; y la nueva crisis producida por el COVID-19 ha causado que el gasto en pensiones se dispare hasta casi el 14% de nuestro producto interior bruto en tan solo unos meses.

Como todavía tenemos pocos datos disponibles de 2020 y son muy atípicos en nuestros primeros análisis no tendremos en cuenta esos meses, ya que ningún modelo podía predecir ese comportamiento previamente.

Gráfico 5.1. Serie de Porcentaje de Gasto del PIB en Pensiones



5.2. Análisis

Los dos primeros modelos que crearemos serán los de Holt-Winters aditivo y multiplicativo. Con este tipo de modelos no tenemos que preocuparnos acerca de si los datos no son estacionarios, de hecho, la utilización de este tipo de métodos sobre datos estacionarios produce resultados subóptimos.

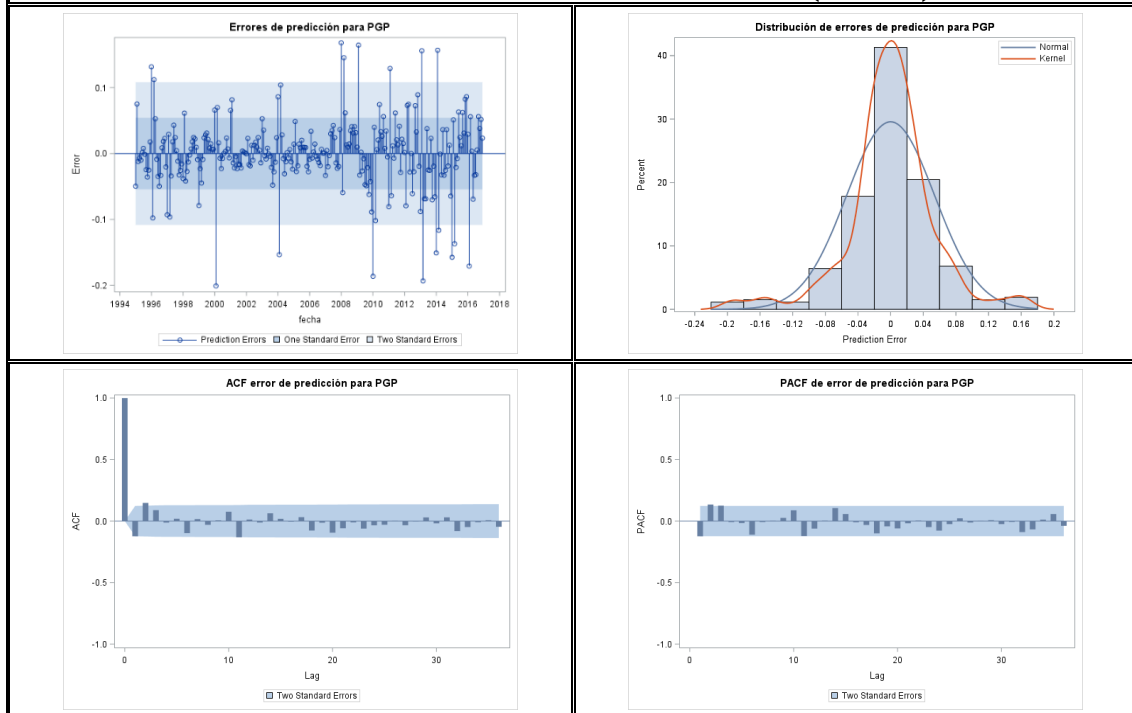
En ambos casos tomaremos datos hasta diciembre de 2016 para crear el modelo y guardaremos el resto para evaluar la calidad de nuestras predicciones.

Tabla 5.2. Método de Winters (Aditivo)

Parámetro	Estimador	Error estándar	Valor t	Aprox Pr > t
Peso de nivel	0.30821	0.02759	11.17	<.0001
Peso de tendencia	0.32757	0.04802	6.82	<.0001
Peso estacional	0.26549	0.03692	7.19	<.0001

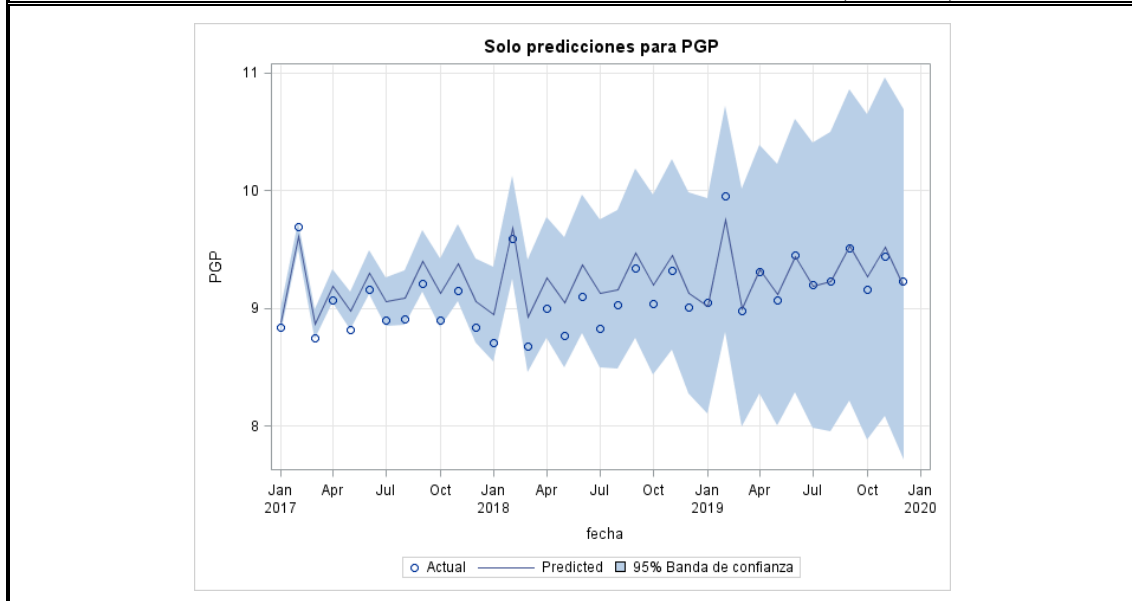
Vemos que todos los coeficientes son estadísticamente significativos para el modelo aditivo. También, mediante el análisis gráfico de los residuos observamos que el modelo es correcto, ya que los residuos son incorrelados y están normalmente distribuidos.

Gráfico 5.2. Residuos Método de Winters (Aditivo)



En cuanto a las predicciones, vemos que son en general muy buenas.

Gráfico 5.3. Predicciones Método de Winters (Aditivo)



Por último, mostraremos los estadísticos asociados a la calidad del modelo tanto para la muestra de entrenamiento como para la del conjunto test.

Tabla 5.3. Estadísticas de ajuste para la variable PGP	
Estadístico	Valor
Grados de error de libertad	261
Número de observaciones utilizadas	264
Número de parámetros modelo	3
Error cuadrático medio	0.00290176
Error cuadrático medio de la raíz	0.05386795
Error absoluto medio	0.03695032
R-cuadrado	0.99746856
R-cuadrado ajustado	0.99744916
Criterio de información de Akaike	-1536.404
Criterio de información bayesiano de Schwarz	-1525.6761
Cálculos basados en el rango de ajuste de datos.	

Tabla 5.4. Estadísticas de rendimiento para la variable PGP	
Estadístico	Valor
Grados de error de libertad	36
Número de observaciones utilizadas	36
Error cuadrático medio	0.0257795
Error cuadrático medio de la raíz	0.16055996
Error absoluto medio	0.13433175
R-cuadrado	0.68659986
R-cuadrado ajustado	0.69530542
Criterio de información de Akaike	-131.69433
Criterio de información bayesiano de Schwarz	-131.69433
Cálculos basados en la región de predicción de datos.	

A continuación, observamos que la información relativa al modelo de Holt-Winters multiplicativo es muy similar.

Tabla 5.5 Método de Winters (Multiplicativo)				
Parámetro	Estimador	Error estándar	Valor t	Aprox Pr > t
Peso de nivel	0.28367	0.02354	12.05	<.0001
Peso de tendencia	0.36503	0.04906	7.44	<.0001
Peso estacional	0.13793	0.01773	7.78	<.0001

Gráfico 5.4. Residuos Método de Winters (Multiplicativo)

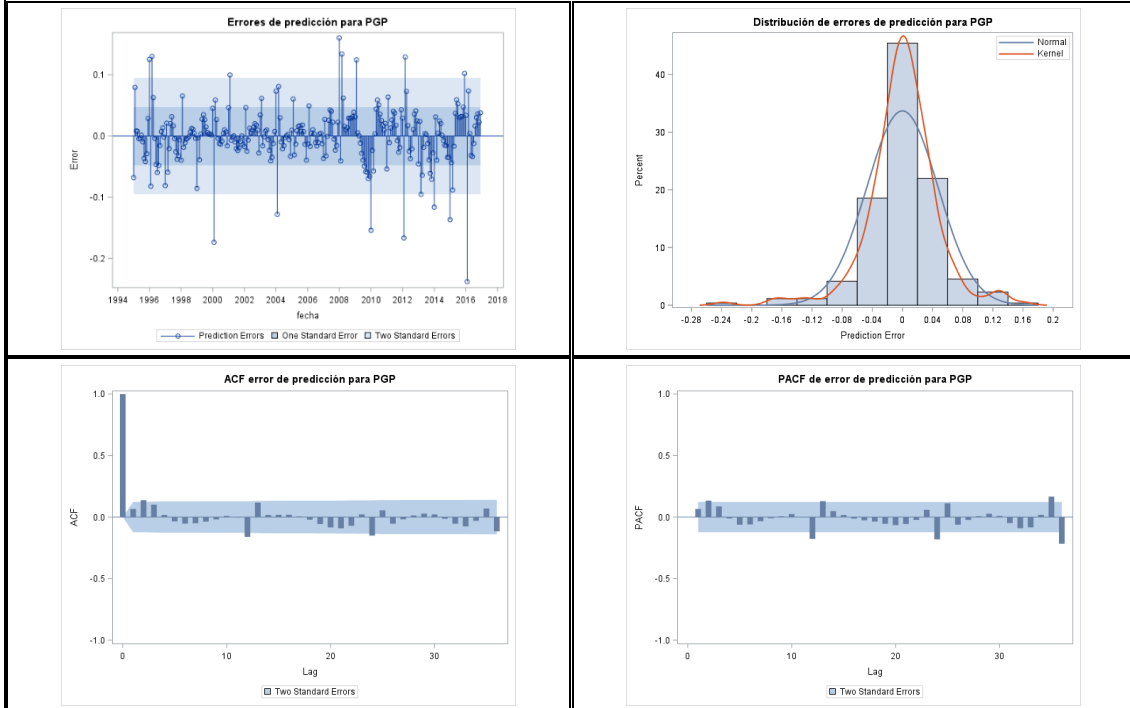


Gráfico 5.5. Predicciones Método de Winters (Multiplicativo)

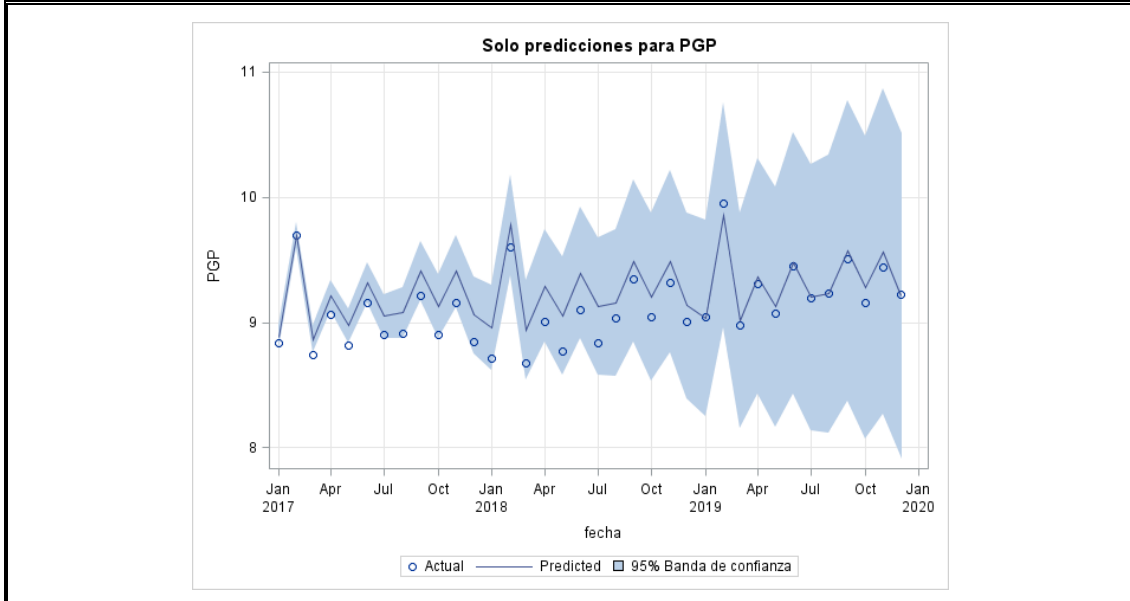


Tabla 5.6. Estadísticas de ajuste para la variable PGP

Estadístico	Valor
Grados de error de libertad	261
Número de observaciones utilizadas	264
Número de parámetros modelo	3

Tabla 5.6. Estadísticas de ajuste para la variable PGP	
Estadístico	Valor
Error cuadrático medio	0.0022329
Error cuadrático medio de la raíz	0.04725362
Error absoluto medio	0.03238464
R-cuadrado	0.99805205
R-cuadrado ajustado	0.99803712
Criterio de información de Akaike	-1605.5754
Criterio de información bayesiano de Schwarz	-1594.8475
Cálculos basados en el rango de ajuste de datos.	

Tabla 5.7. Estadísticas de rendimiento para la variable PGP	
Estadístico	Valor
Grados de error de libertad	36
Número de observaciones utilizadas	36
Error de suma de cuadrados	1.00272543
Error cuadrático medio	0.02785348
Error cuadrático medio de la raíz	0.16689363
Error absoluto medio	0.13999174
R-cuadrado	0.66138652
R-cuadrado ajustado	0.67079245
Criterio de información de Akaike	-128.9087
Criterio de información bayesiano de Schwarz	-128.9087
Cálculos basados en la región de predicción de datos.	

En definitiva, tanto el modelo aditivo como el multiplicativo son adecuados para modelizar el porcentaje de gasto del PIB en pensiones. La principal virtud de estos métodos es que a pesar de su sencillez proporcionan unas predicciones muy ajustadas a la realidad.

Ambos modelos quedan definidos matemáticamente de la siguiente manera:

Holt-Winters Aditivo

$$x_t = (L_t + b_t) + S_t$$

donde

$$L_t = 0.30831(x_t - S_{t-s}) + (1 - 0.30821)(L_{t-1} + b_{t-1})$$

$$b_t = 0.32757(L_t - L_{t-1}) + (1 - 0.32757)b_{t-1}$$

$$S_t = 0.26549(x_t - L_t) + (1 - 0.26549)S_{t-s}$$

Holt-Winters Multiplicativo

$$x_t = (L_t + b_t) \times S_t$$

donde

$$L_t = 0.28367 \frac{x_t}{S_{t-s}} + (1 - 0.28367)(L_{t-1} + b_{t-1})$$

$$b_t = 0.36503(L_t - L_{t-1}) + (1 - 0.36503)b_{t-1}$$

$$S_t = 0.13793 \frac{x_t}{L_t} + (1 - 0.13793)S_{t-s}$$

El siguiente paso será la creación de un modelo ARIMA mediante el uso de la metodología Box-Jenkins. La principal dificultad de esta metodología es que, debido a la infinidad de modelos que pueden ser creados, no siempre es fácil saber cuál es el correcto para explicar la estructura de nuestra serie. Por ello a lo largo de los años se ha desarrollado un procedimiento iterativo que permite identificar el modelo más apropiado (Gráfico 5.6). El problema de este esquema de trabajo es que no siempre conduce a resultados únicos, por lo que dos investigadores podrían llegar a resultados diferentes y validos con los mismos datos. Paralelamente a esta forma manual de actuar casi todos los programas estadísticos actuales han desarrollado algoritmos que permiten el cálculo del mejor modelo de forma automática, aunque los resultados no siempre son los óptimos.

Aquí también trabajaremos con los datos desde 1995 hasta 2016 para la creación del modelo y los posteriores para comprobar la calidad de las predicciones.

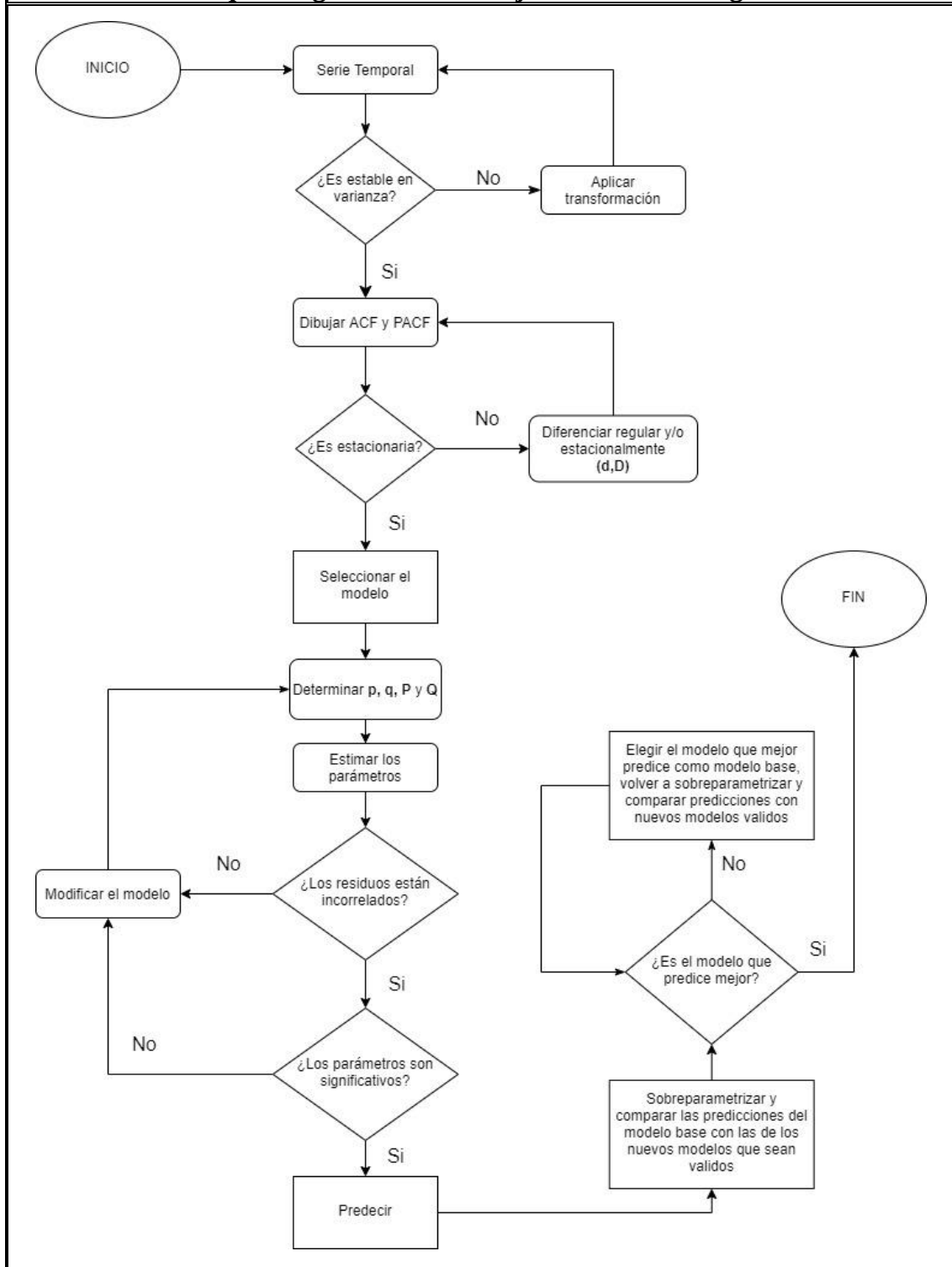
En este caso sí que debemos comprobar si nuestra serie es estacionaria, pues en caso contrario deberemos de diferenciar regular y/o estacionalmente. Para que una serie $\{z_t\}_{t \geq 1}$ sea estacionaria debe cumplir las siguientes propiedades:

$$\mu = E[z_t] \forall t \in \mathbb{Z}^+$$

$$\sigma^2 = Var[z_t] \forall t \in \mathbb{Z}^+$$

$$\gamma_{t,t+h} = Cov[z_t, z_{t+h}] \forall t, h \in \mathbb{Z}^+$$

Gráfico 5.6. Esquema general de trabajo de la metodología Box-Jenkins



En primer lugar, analizaremos si es necesario realizar transformaciones para estabilizar la varianza. Para ello utilizaremos principalmente el test de Levene, donde se pone a prueba la hipótesis nula de igualdad de varianzas en la población.

Gráfico 5.7. Box-Plot variable PGP

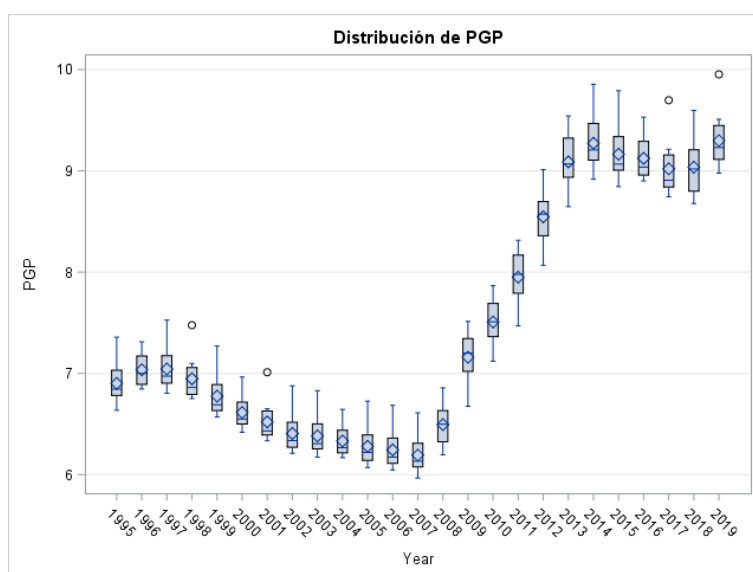


Tabla 5.8. Test de Levene para homogeneidad de la varianza PGP
ANOVA de desviaciones absolutas de las medias de grupo

Fuente	DF	Suma de cuadrados	Cuadrado de la media	F-Valor	Pr > F
Year	24	0.2813	0.0117	0.69	0.8558
Error	275	4.6383	0.0169		

Teniendo en cuenta el Gráfico 5.7 y la Tabla 5.8, determinamos que no hay evidencias estadísticamente significativas para rechazar la homocedasticidad. Luego los datos son constantes en varianza y no hace falta realizar ninguna transformación.

Observamos ahora que la serie no es estacionaria, lo cual se pone de manifiesto en el lento decrecimiento de su ACF, tanto regular como periódica, y en que no podemos rechazar la hipótesis nula de la existencia de raíces unitarias mediante la prueba de Dickey-Fuller.

Gráfico 5.8. ACF y PACF variable PGP

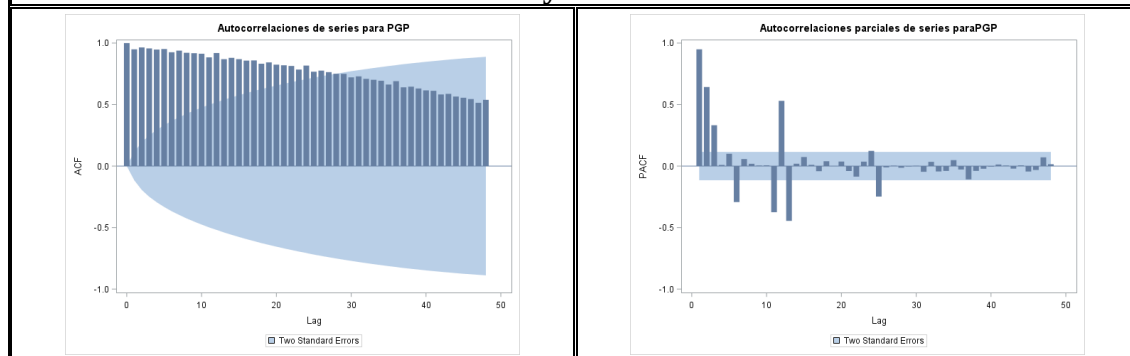


Tabla 5.9. Pruebas aumentadas de la raíz unidad de Dickey-Fuller variable PGP						
Tipo	Rho	Pr < Rho	Tau	Pr < Tau	F	Pr > F
Zero Mean	0.0249	0.6882	0.03	0.6923		
Single Mean	-13.0799	0.0617	-2.50	0.1172	3.21	0.2493
Tendencia	-36.2181	0.0016	-4.40	0.0025	9.71	0.0010

Por tanto, tomamos la decisión de que debemos realizar primero una diferenciación ordinaria, y luego una diferenciación periódica. Si hubiéramos procedido diferenciando primero estacionalmente, habríamos llegado también a la conclusión de que teníamos que diferenciar después regularmente.

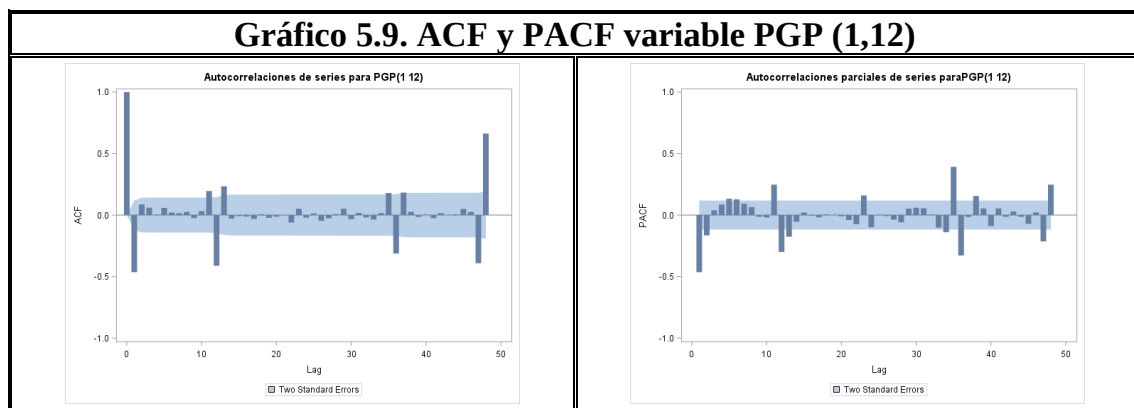


Tabla 5.10. Pruebas aumentadas de la raíz unidad de Dickey-Fuller variable PGP (1,12)						
Tipo	Rho	Pr < Rho	Tau	Pr < Tau	F	Pr > F
Zero Mean	-419.378	0.0001	-28.55	<.0001		
Single Mean	-419.368	0.0001	-28.50	<.0001	406.13	0.0010
Tendencia	-419.377	0.0001	-28.45	<.0001	404.67	0.0010

Ahora que ya podemos decir que la serie PGP (1,12) es estacionaria, procedemos a estimar el mejor modelo posible que logre identificar la estructura de la serie. Para avanzar en la búsqueda hacia ese modelo ideal, utilizaremos la ACF, PACF, la significatividad de los coeficientes estimados, el estadístico de Box.Ljung y los estadísticos de información.

Tabla 5.11. Elección del Mejor Modelo				
MODELO	Box-Ljung	¿Son significativos todos los coeficientes?	SBC	¿Es válido el modelo?
PASO 1. MODELO DE PARTIDA $ARIMA(0, 1, 0)(0, 1, 0)_{12}$				
$ARIMA(0, 1, 0)(0, 1, 0)_{12}$	<.0001	SI	-607.134	NO
$ARIMA(1, 1, 0)(0, 1, 0)_{12}$	<.0001	SI	-671.506	NO
$ARIMA(0, 1, 1)(0, 1, 0)_{12}$	<.0001	SI	-657.04	NO
$ARIMA(0, 1, 0)(1, 1, 0)_{12}$	<.0001	SI	-669.244	NO
$ARIMA(0, 1, 0)(0, 1, 1)_{12}$	<.0001	SI	-695.577	NO
PASO 2. MODELO DE PARTIDA $ARIMA(0, 1, 0)(0, 1, 1)_{12}$				
$ARIMA(1, 1, 0)(0, 1, 1)_{12}$	<.0001	SI	-733.635	NO
$ARIMA(0, 1, 1)(0, 1, 1)_{12}$	<.0001	NO	-701.087	NO
$ARIMA(0, 1, 0)(1, 1, 1)_{12}$	<.0001	SI	-713.115	NO
PASO 3. MODELO DE PARTIDA $ARIMA(1, 1, 0)(0, 1, 1)_{12}$				
$ARIMA(2, 1, 0)(0, 1, 1)_{12}$	0.0005	SI	-738.519	NO
...	...	...	...	...
PASO 4. MODELO DE PARTIDA $ARIMA(2, 1, 0)(0, 1, 1)_{12}$				
$ARIMA(3, 1, 0)(0, 1, 1)_{12}$	0.0030	SI	-738.592	NO
...	...	...	...	...
PASO 5. MODELO DE PARTIDA $ARIMA(3, 1, 0)(0, 1, 1)_{12}$				
$ARIMA(4, 1, 0)(0, 1, 1)_{12}$	0.0316	SI	-739.256	NO
...	...	...	...	...
PASO 6. MODELO DE PARTIDA $ARIMA(4, 1, 0)(0, 1, 1)_{12}$				
$ARIMA(5, 1, 0)(0, 1, 1)_{12}$	0.2876	SI	-744.263	SI
...	...	...	...	...
PASO 7. MODELO DE PARTIDA $ARIMA(5, 1, 0)(0, 1, 1)_{12}$				
$ARIMA(6, 1, 0)(0, 1, 1)_{12}$	0.7103	SI	-748.083	SI
...	...	...	...	...
PASO 8. MODELO DE PARTIDA $ARIMA(6, 1, 0)(0, 1, 1)_{12}$				
$ARIMA(7, 1, 0)(0, 1, 1)_{12}$	0.9158	SI	-749.69	SI
$ARIMA(6, 1, 1)(0, 1, 1)_{12}$	0.6675	NO	-743.176	NO
$ARIMA(6, 1, 0)(1, 1, 1)_{12}$	0.6691	NO	-742.424	NO
PASO 9. MODELO DE PARTIDA $ARIMA(7, 1, 0)(0, 1, 1)_{12}$				
$ARIMA(8, 1, 0)(0, 1, 1)_{12}$	0.9039	NO	-746.079	NO
$ARIMA(7, 1, 1)(0, 1, 1)_{12}$	0.8702	NO	-745.088	NO
$ARIMA(7, 1, 0)(1, 1, 1)_{12}$	0.9093	NO	-744.103	NO
PASO 10. MODELO FINAL $ARIMA(7, 1, 0)(0, 1, 1)_{12}$				

A lo largo de este proceso iterativo determinamos que el mejor candidato para modelizar nuestra serie es el $ARIMA(7,1,0)(0,1,1)_{12}$. A continuación, se muestra la información asociada a este modelo.

Tabla 5.12. $ARIMA(7, 1, 0)(0, 1, 1)_{12}$					
Parámetro	Estimador	Error estándar	Valor t	Aprox Pr > t	Retardo
MU	-0.0000551	0.0023473	-0.02	0.9813	0
MA1,1	0.63297	0.05053	12.53	<.0001	12
AR1,1	-0.53578	0.05962	-8.99	<.0001	1
AR1,2	-0.10337	0.06629	-1.56	0.1201	2
AR1,3	0.12539	0.06410	1.96	0.0514	3
AR1,4	0.20318	0.06350	3.20	0.0015	4
AR1,5	0.29959	0.06448	4.65	<.0001	5
AR1,6	0.26517	0.06676	3.97	<.0001	6
AR1,7	0.16100	0.06040	2.67	0.0081	7

Tabla 5.13. Comprobación de autocorrelación de residuales (Box-Ljung)			
Para retardo	Chi-cuadrado	DF	Pr > ChiSq
12	4.41	4	0.3533
18	6.79	10	0.7453
24	10.46	16	0.8415
30	12.45	22	0.9473
36	17.37	28	0.9409
42	23.33	34	0.9158

Tabla 5.14. Estadísticas de ajuste para la variable PGP	
Error Std Estimación	0.060942
AIC	-782.625
SBC	-749.69
Número de residuales	287

Se observa que los parámetros del modelo tienen un p-valor asociado muy bajo (salvo la constante que no importa si es o no estadísticamente significativa); sus residuos son normales e incorrelados; y sus predicciones se acercan mucho a los datos reales. Por tanto, en este caso también concluimos que el modelo $ARIMA(7,1,0)(0,1,1)_{12}$ es adecuado para estimar el porcentaje de gasto del PIB en pensiones.

Nuestro modelo queda definido matemáticamente como:

$$(1 + 0.53578L + 0.10337L^2 - 0.12539L^3 - 0.20318L^4 - 0.29959L^5 - 0.26517L^6 - 0.1600L^7)(1 - L^{12})(1 - L)z_t = -0.00006 + (1 - 0.63297L^{12})\epsilon_t$$

Gráfico 5.10. Residuos $ARIMA(7, 1, 0)(0, 1, 1)_{12}$

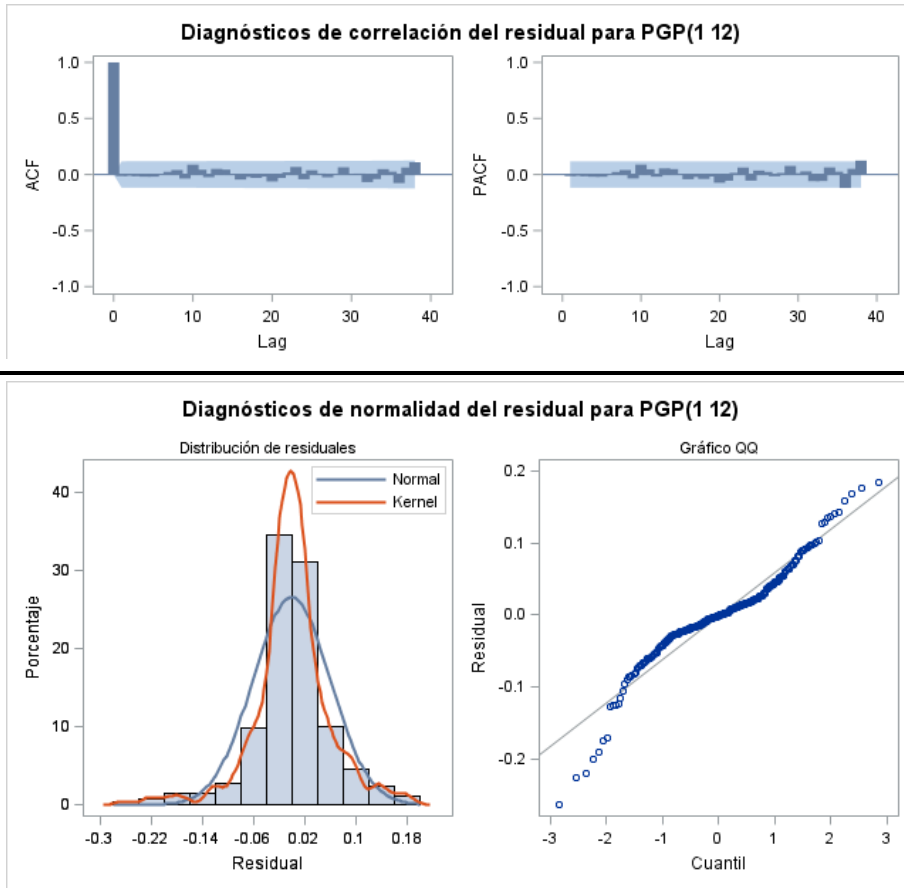
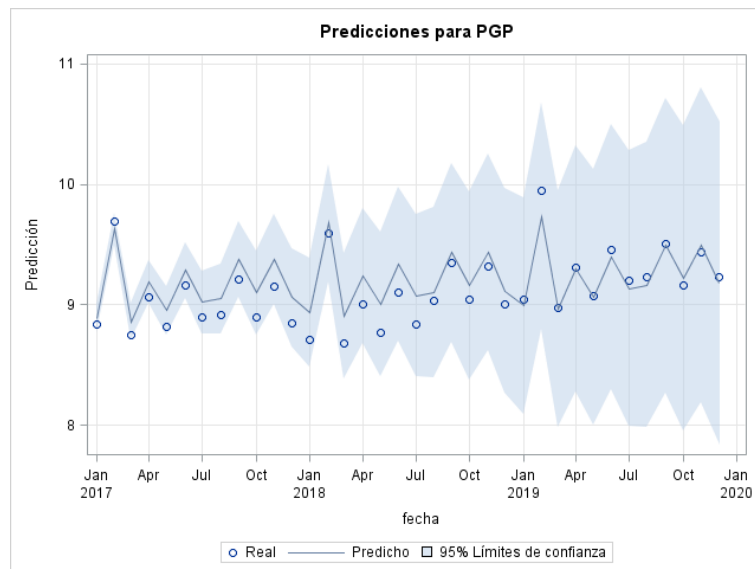


Gráfico 5.11. Predicciones $ARIMA(7, 1, 0)(0, 1, 1)_{12}$



Pasemos ahora a cambiar de paradigma, para adentrarnos en la metodología multivariante. El primer punto a tratar será la reducción del número de variables. En nuestro caso partimos de las siguientes series:

Tabla 5.15. Número de Variables		
Series	Número de Variables	Descripción
Afiliados	1	Número de afiliados a la Seguridad Social
PGP	1	Porcentaje del PIB destinado a pensiones
Inmigración	100	Flujo de inmigrantes por edad
Mortalidad	200	Tasa de mortalidad por edad y sexo
Población	172	Población por edad y sexo

Debido a que muchas de estas series están muy correlacionadas, procederemos a realizar una reducción mediante componentes principales. El ACP se hará en tres fases independientes: una para las series de inmigración, otra para las de mortalidad y una última para las de población.

En el caso de las series de inmigración, vemos que al analizar los autovalores debemos de reducir a dos componentes principales, ya que con ello conseguimos explicar más del 90% de la variabilidad de nuestros datos.

Tabla 5.16. Serie Inmigración - Autovalores de la matriz de correlación				
	Autovalor	Diferencia	Proporción	Acumulada
1	85.6728872	74.0499011	0.8482	0.8482
2	11.6229861	10.5404390	0.1151	0.9633
3	1.0825471	0.5065657	0.0107	0.9740

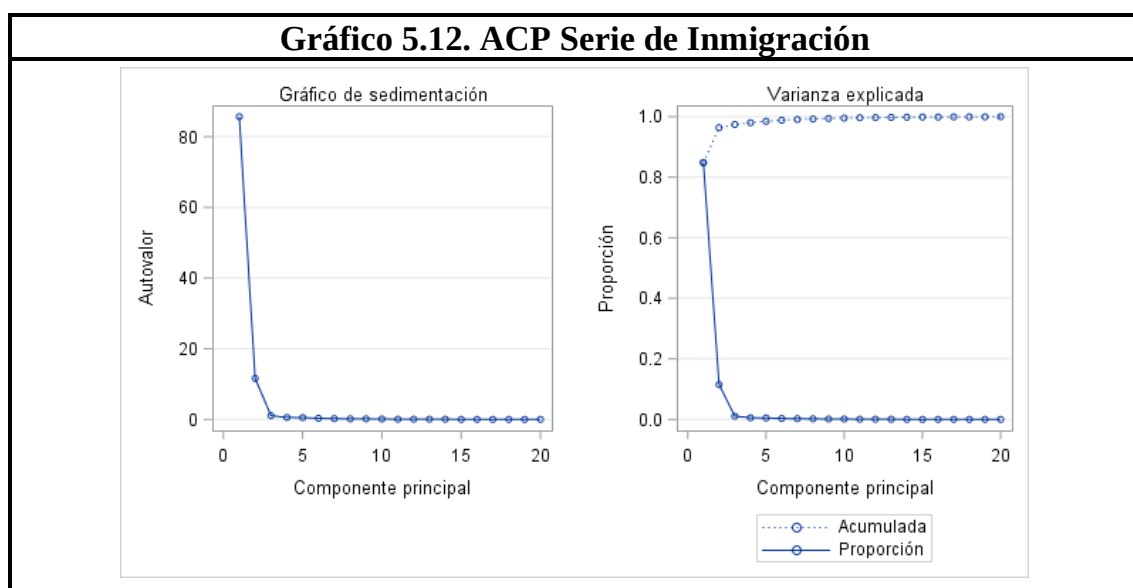
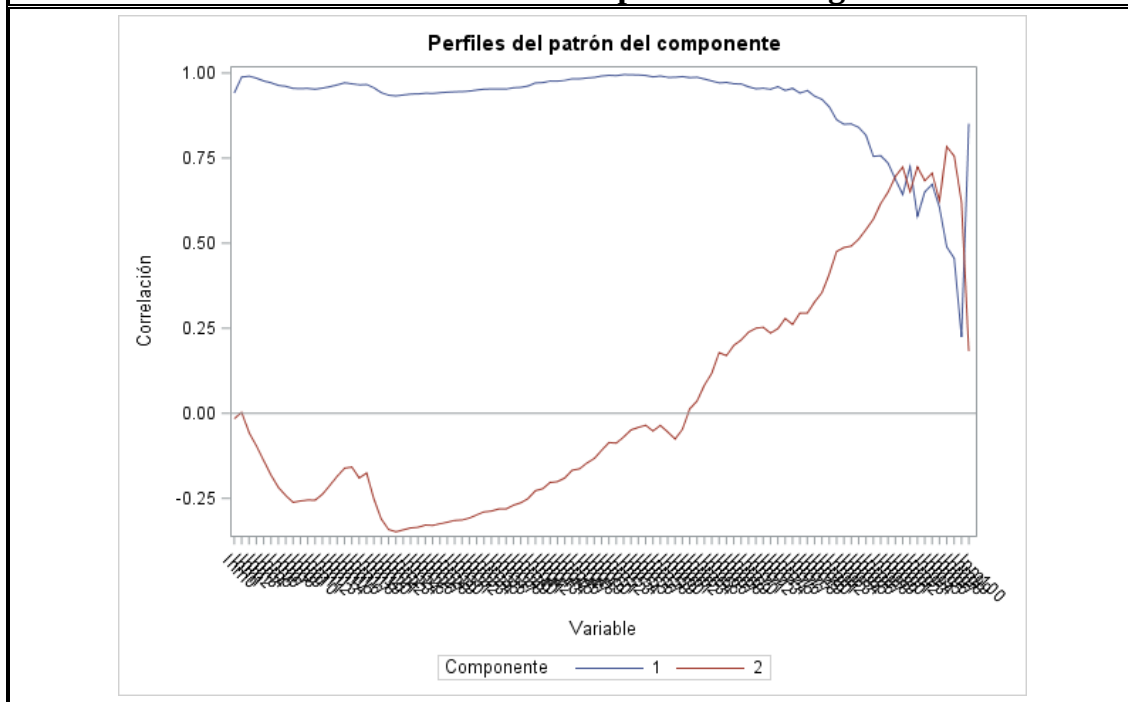


Gráfico 5.13. Patrones Componentes Inmigración



Las representaciones gráficas nos ayudan a entender como están recogidas nuestras variables iniciales en las nuevas componentes. Por lo que podríamos decir que, en general, la primera Componente toma valores cercanos a uno cuando son personas menores de 65 años y valores inferiores a medida que aumenta su edad. La segunda Componente toma valores negativos hasta los 60 años aproximadamente (con el mínimo cercano a los 20 años), para luego crecer rápidamente.

Procediendo de manera similar estipulamos que la serie de mortalidad puede ser reducida a tres componentes principales y la serie de población a cuatro componentes principales. En ambos casos conseguimos explicar más del 90% de la variabilidad de nuestros datos.

Tabla 5.17. Serie Mortalidad - Autovalores de la matriz de correlación				
	Autovalor	Diferencia	Proporción	Acumulada
1	169.999371	162.356933	0.8416	0.8416
2	7.642438	3.098831	0.0378	0.8794
3	4.543607	1.884282	0.0225	0.9019
4	2.659325	0.495731	0.0132	0.9151

Gráfico 5.14. ACP Serie de Mortalidad

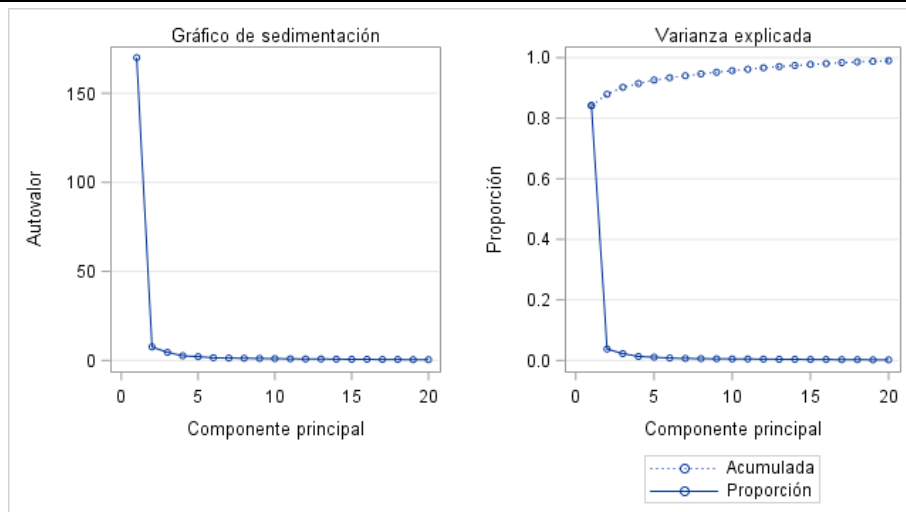
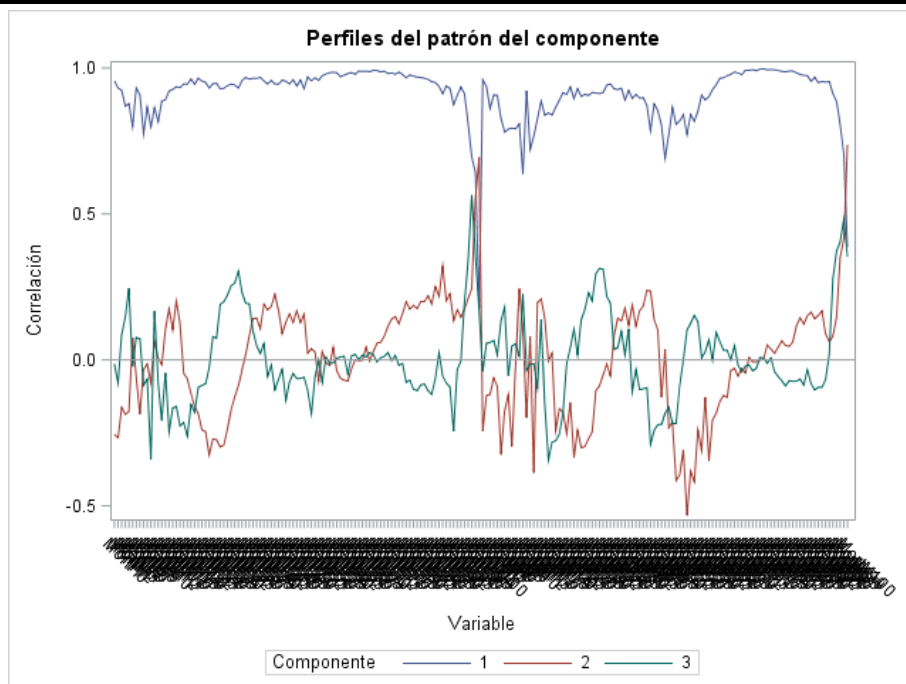


Gráfico 5.15. Patrones Componentes Mortalidad



Ahora podríamos identificar la primera Componente como un factor masa. La segunda Componente y la tercera Componente reúnen mucha información (variabilidad), pero muy compleja.

Tabla 5.18. Serie Población - Autovalores de la matriz de correlación

	Autovalor	Diferencia	Proporción	Acumulada
1	108.706558	70.911947	0.6248	0.6248
2	37.794611	26.373957	0.2172	0.8420
3	11.420653	6.202805	0.0656	0.9076
4	5.217849	3.045380	0.0300	0.9376
5	2.172468	0.844531	0.0125	0.9501

Gráfico 5.16. ACP Serie de Población

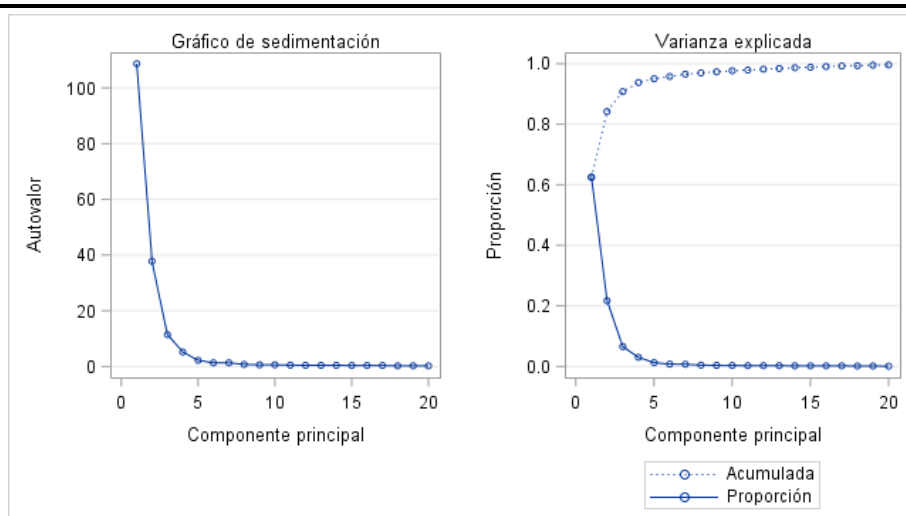
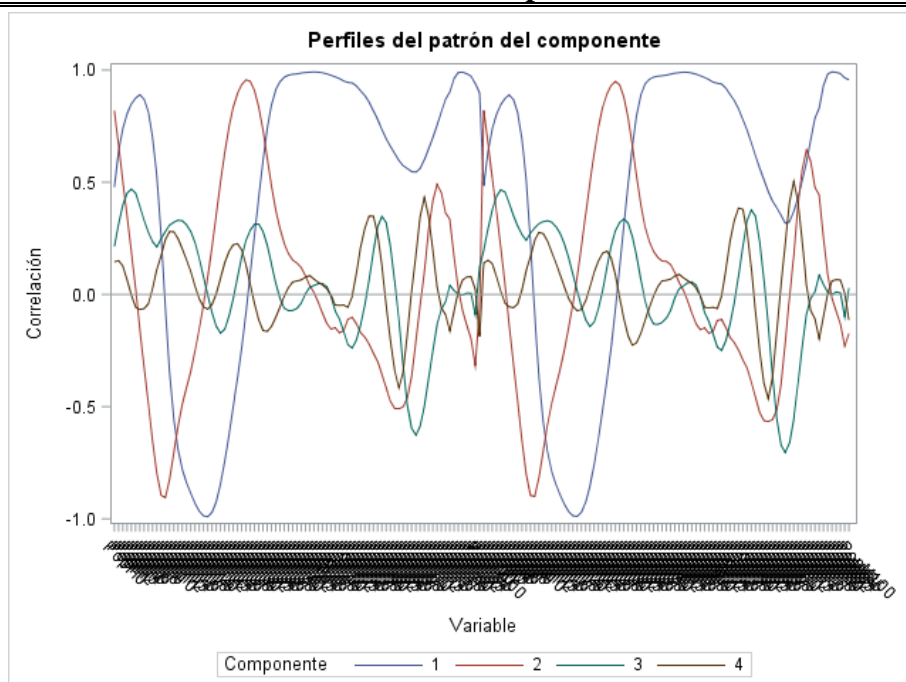


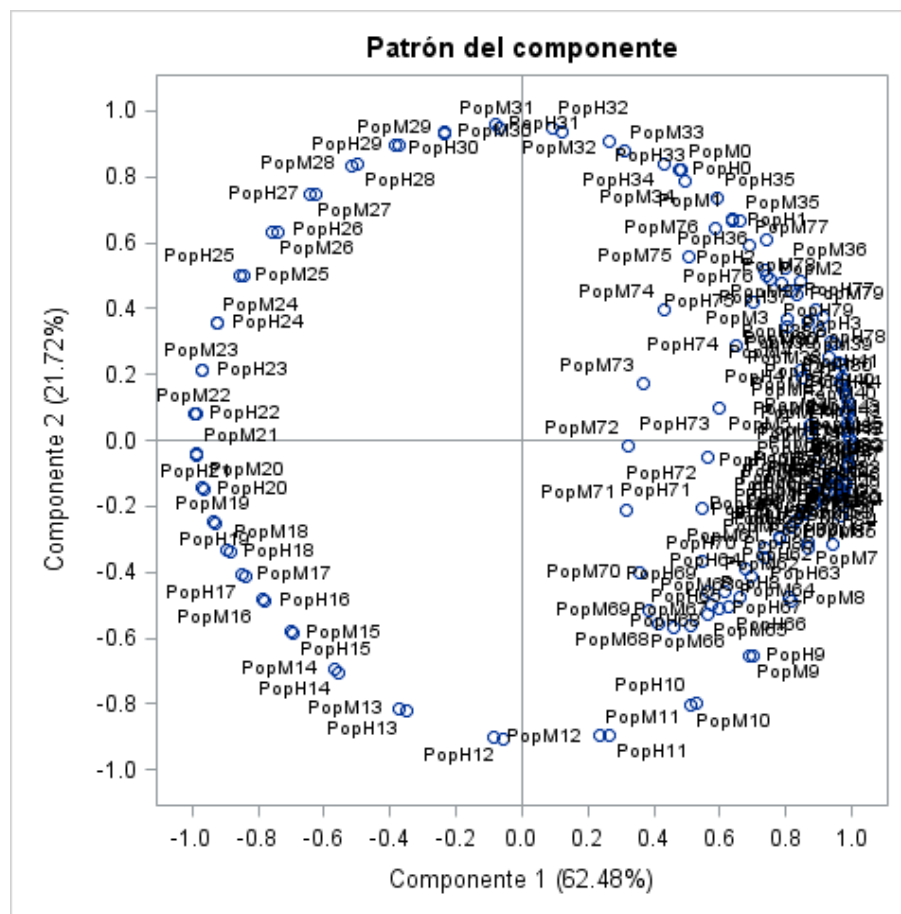
Gráfico 5.17. Patrones Componentes Población



Vemos como en el caso de la Componentes Principales relativas a la serie de población hemos conseguido unas Componentes muy informativas (mucha variabilidad), aunque es difícil de explicar que representa cada una de ellas. La Componente que tal vez sea más sencilla de explicar es la primera, ya que se relaciona de manera negativa con la población más joven y de manera positiva con la población de mayor edad.

A continuación, se muestra el Gráfico 5.18 como ejemplo de la calidad de las componentes que hemos conseguido mediante el ACP. Se observa claramente como las dos primeras componentes del nuevo sistema de coordenadas obtenido consiguen maximizar la información (variabilidad) de nuestros datos.

Gráfico 5.18. Serie Mortalidad - Valores en las Componentes Principales



En resumen, hemos conseguido reducir de 474 variables a solo 11 variables, manteniendo al menos el 90% de la información que teníamos originalmente.

El siguiente paso será determinar nuestro modelo vectorial autorregresivo de medias móviles. Estos modelos necesitan que todas las series sean estacionarias, por lo que en prácticamente todos los casos deberemos transformar y/o diferenciar nuestras series originales regular y/o estacionalmente. Como podemos ver en las tablas 5.19 y 5.20, al aplicar las diferenciaciones pertinentes conseguimos solucionar los problemas de estacionaridad de todas nuestras series.

Tabla 5.19. Pruebas de la raíz unidad de Dickey-Fuller					
Variable	Tipo	Rho	Pr < Rho	Tau	Pr < Tau
PGP	Zero Mean	-581.53	0.0001	-17.11	<.0001
	Single Mean	-581.55	0.0001	-17.08	<.0001
	Trend	-581.73	0.0001	-17.05	<.0001
NAFIL	Zero Mean	-160.93	0.0001	-8.90	<.0001
	Single Mean	-199.04	0.0001	-9.91	<.0001
	Trend	-212.79	0.0001	-10.21	<.0001
PopCP1	Zero Mean	-11.82	0.0162	-2.66	0.0080
	Single Mean	-3.28	0.6211	-2.91	0.0464
	Trend	-45.01	0.0006	-2.00	0.5982
PopCP2	Zero Mean	-1085.8	0.0001	-4.29	<.0001
	Single Mean	-178.30	0.0001	-4.03	0.0016
	Trend	-44.94	0.0006	-4.01	0.0094
PopCP3	Zero Mean	-76.23	<.0001	-7.34	<.0001
	Single Mean	-67.95	0.0015	-7.33	<.0001
	Trend	-67.96	0.0006	-7.15	<.0001
PopCP4	Zero Mean	259.16	0.9999	-7.63	<.0001
	Single Mean	311.81	0.9999	-7.43	<.0001
	Trend	-1153.6	0.0001	-7.49	<.0001
MortCP1	Zero Mean	-764.29	0.0001	-9.96	<.0001
	Single Mean	-2362.6	0.0001	-13.73	<.0001
	Trend	-3081.8	0.0001	-13.66	<.0001
MortCP2	Zero Mean	-493.52	0.0001	-11.27	<.0001
	Single Mean	-475.55	0.0001	-11.35	<.0001
	Trend	-467.56	0.0001	-12.08	<.0001
MortCP3	Zero Mean	-129.18	0.0001	-9.35	<.0001
	Single Mean	-129.11	0.0001	-9.34	<.0001
	Trend	-126.57	0.0001	-12.88	<.0001
InmiCP1	Zero Mean	-24.29	0.0003	-3.47	0.0006
	Single Mean	-24.35	0.0036	-3.47	0.0097
	Trend	-24.36	0.0253	-3.47	0.0454
InmiCP2	Zero Mean	-37.42	<.0001	-3.41	0.0007
	Single Mean	-39.66	0.0015	-3.59	0.0068
	Trend	-44.37	0.0006	-3.97	0.0109

Tabla 5.20. Pruebas estacionales de la raíz unidad de Dickey-Fuller (Nseason=12)					
Variable	Tipo	Rho	Pr < Rho	Tau	Pr < Tau
PGP	Zero Mean	-366.98	0.0001	-25.06	<.0001
	Single Mean	-367.00	0.0001	-25.01	<.0001
NAFIL	Zero Mean	-158.98	0.0001	-10.13	<.0001
	Single Mean	-176.73	0.0001	-10.96	<.0001
PopCP1	Zero Mean	-274.51	0.0001	-22.34	<.0001
	Single Mean	-277.28	0.0001	-21.51	<.0001
PopCP2	Zero Mean	-326.83	0.0001	-28.03	<.0001
	Single Mean	-317.04	0.0001	-26.30	<.0001
PopCP3	Zero Mean	-348.38	0.0001	-40.34	<.0001
	Single Mean	-348.76	0.0001	-40.19	<.0001
PopCP4	Zero Mean	-413.81	0.0001	-25.92	<.0001
	Single Mean	-407.25	0.0001	-25.24	<.0001
MortCP1	Zero Mean	-490.79	0.0001	-52.45	<.0001
	Single Mean	-492.69	0.0001	-53.96	<.0001
MortCP2	Zero Mean	-481.94	0.0001	-52.40	<.0001
	Single Mean	-481.76	0.0001	-52.97	<.0001
MortCP3	Zero Mean	-348.83	0.0001	-39.75	<.0001
	Single Mean	-348.54	0.0001	-39.21	<.0001
InmiCP1	Zero Mean	-283.18	0.0001	-17.20	<.0001
	Single Mean	-283.22	0.0001	-17.17	<.0001
InmiCP2	Zero Mean	-472.44	0.0001	-38.10	<.0001
	Single Mean	-472.24	0.0001	-38.14	<.0001

El siguiente paso será determinar el orden de nuestro modelo VARMA, para ello utilizaremos como regla de decisión el modelo que proporcione un menor estadístico BIC. En este caso usaremos la opción que nos facilita SAS para la estimación automática del mejor modelo.

Tabla 5.21. Criterio de información mínimo basado en SBC									
Retardo	MA 0	MA 1	MA 2	MA 3	MA 4	MA 5	MA 6	MA 7	MA 8
AR 0	14.681905	14.260898	16.693232	19.136672	21.587897	24.023192	26.414263	28.797832	31.199062
AR 1	-44.73095	-44.64565	-42.29815	-39.86485	-37.42203	-35.12859	-32.81033	-30.47602	-28.22642
AR 2	-58.74639	-57.82244	-55.53706	-53.20942	-50.99018	-48.9731	-47.11953	-44.91443	-42.84365
AR 3	-58.31864	-57.41764	-55.2002	-52.89466	-50.65034	-48.74973	-46.8909	-44.89737	-42.83145
AR 4	-57.82119	-57.09906	-55.02307	-52.75657	-50.59121	-48.67519	-46.9632	-45.01607	-42.88832
AR 5	-56.26876	-55.56913	-53.56858	-51.33329	-49.32029	-47.25824	-45.63529	-43.72892	-41.71188
AR 6	-54.8861	-54.5176	-52.39548	-50.27264	-48.16417	-46.3089	-44.62556	-42.76921	-41.01893
AR 7	-53.49021	-53.18385	-51.13183	-49.07915	-47.05889	-45.21019	-43.45217	-41.72904	-40.00695
AR 8	-52.32148	-52.10969	-50.14831	-48.09113	-46.19541	-44.44199	-42.66584	-41.10175	-39.66027
AR 9	-50.465	-50.44184	-48.52291	-46.46784	-44.68239	-43.02391	-41.28175	-39.86458	-38.53834
AR 10	-48.53855	-48.82047	-46.84479	-45.02972	-43.23146	-41.80978	-40.30804	-39.14911	-38.17655
AR 11	-47.09512	-47.66978	-45.82805	-44.1157	-42.51799	-41.06903	-40.13202	-39.12324	-39.40026
AR 12	-46.66515	-47.70869	-45.97136	-44.7427	-43.36691	-42.4597	-42.83032	-42.73389	-45.38457

Elegimos al modelo $VARMA(2,0)$ o lo que es lo mismo un $VAR(2)$.

Para decidir si las variables que estamos utilizando deben entrar en el modelo utilizaremos el test de causalidad de Granger. La hipótesis nula de esta prueba es que las variables de Grupo 1 están influenciadas solo por sí mismas, y no por las variables del grupo 2.

Prueba 1:	Variables Grupo 1:	PGP
	Variables Grupo 2:	NAFIL PopCP1 PopCP2 PopCP3 PopCP4 MortCP1 MortCP2 MortCP3 InmiCP1 InmiCP2

Prueba 2:	Variables Grupo 1:	PGP
	Variables Grupo 2:	NAFIL

Prueba 3:	Variables Grupo 1:	PGP
	Variables Grupo 2:	PopCP1 PopCP2 PopCP3 PopCP4

Prueba 4:	Variables Grupo 1:	PGP
	Variables Grupo 2:	MortCP1 MortCP2 MortCP3

Prueba 5:	Variables Grupo 1:	PopCP1 PopCP2 PopCP3 PopCP4
	Variables Grupo 2:	MortCP1 MortCP2 MortCP3 InmiCP1 InmiCP

Tabla 5.22. Test de Granger-Causality Wald			
Test	DF	Chi-cuadrado	Pr > ChiSq
1	20	36.88	0.0121
2	2	4.62	0.0990
3	8	16.84	0.0318
4	6	5.72	0.4549
5	4	10.19	0.0373
6	40	71.37	0.0017

El resultado muestra que se puede aceptar que todas las variables intervienen directamente o a través de otras variables en la variable PGP y, por lo tanto, las 11 variables deben estar en el modelo.

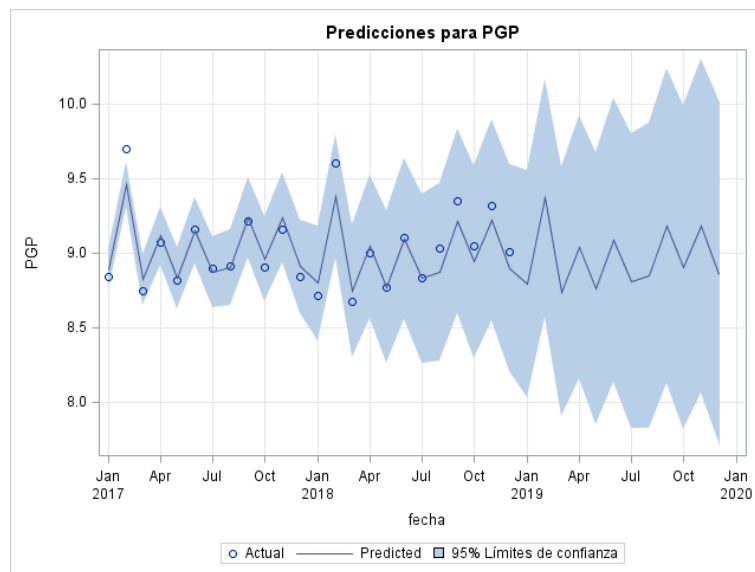
Ahora calculamos los parámetros del modelo $VAR(2)$. El principal problema de este modelo es que tiene una gran cantidad de coeficientes y que muchos de ellos no son estadísticamente significativos. Por eso procedemos a ir eliminando coeficientes de manera iterativa hasta quedarnos con un modelo con menos parámetros, en el que todos los coeficientes sean en general más significativos.

Tabla 5.23. Estimadores del parámetro del modelo $VAR(2)$					
Parámetro	Estimador	Error estándar	Valor t	Pr > t	Variable
CONST1	0.00099	0.00590	0.17	0.8670	1
AR1_1_1	-0.60592	0.05810	-10.43	0.0001	PGP(t-1)
AR1_1_7	-0.14524	0.10466	-1.39	0.1664	MortCP1(t-1)
...	...	...	..	...	...
AR2_11_10	-0.13126	0.02101	-6.25	0.0001	InmiCP1(t-2)
AR2_11_11	-0.82987	0.03338	-24.86	0.0001	InmiCP2(t-2)

Ya solo queda predecir y comprobar los estadísticos de información asociados al modelo.

Tabla 5.24. Criterios de información $VAR(2)$	
AICC	-11533.8
HQC	-13505.2
AIC	-13834.1
SBC	-13014.7
FPEC	1.99E-27

Gráfico 5.19. Predicciones VAR(2)



Debido a la cantidad de variables del modelo, no vamos a comprobar si todos los residuos cumplen con los supuestos de normalidad y autocorrelación. En este caso nos valdremos únicamente de las predicciones para estipular si el modelo es o no válido. Como vemos que las capacidades de predicción del modelo son buenas, pasamos a explicar brevemente algunas de las interacciones entre las variables de nuestro modelo.

Utilizaremos los impulsos para poder explicar de manera visual cómo influyen unas variables sobre otras. Los impulsos representan la respuesta que se produce en una variable cuando en la otra se incrementa una unidad. Si la gráfica se acerca al punto de equilibrio o regresa a la línea de equilibrio original (cero), significa que la respuesta de una variable al impulso de otra variable desaparece, de modo que no tiene un efecto permanente en el tiempo.

Gráfico 5.20. Respuesta a Impulso en NAFIL de PGP

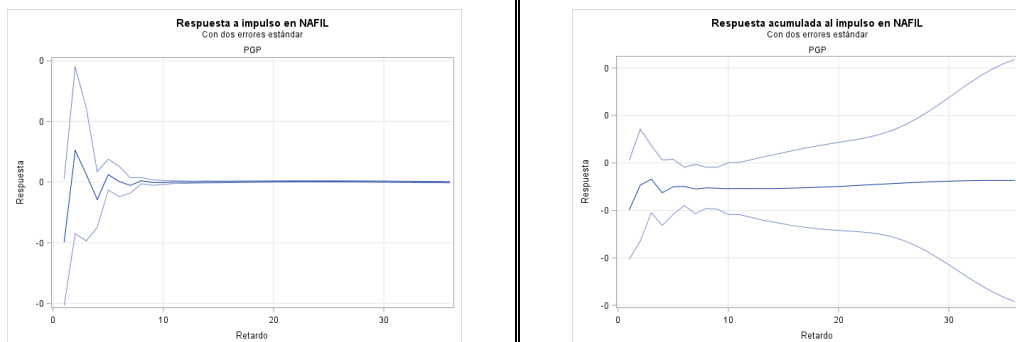


Gráfico 5.21. Respuesta a Impulso en PopCP de PGP

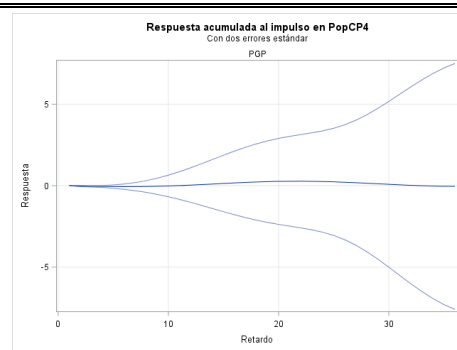
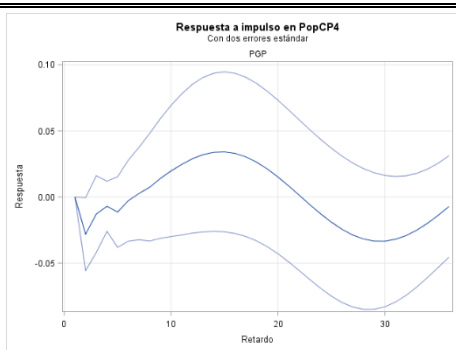
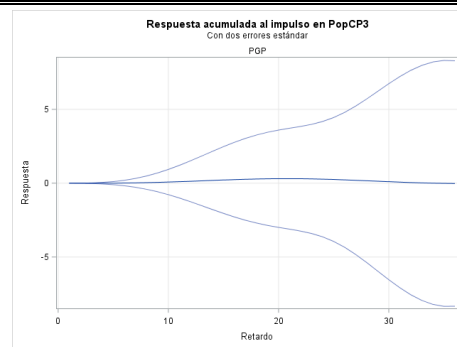
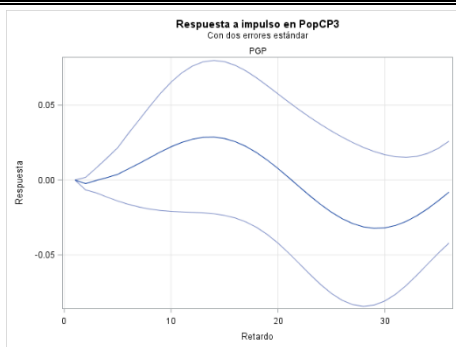
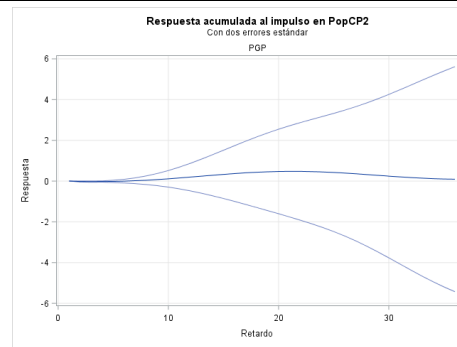
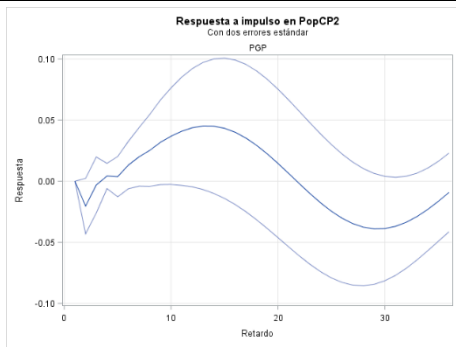
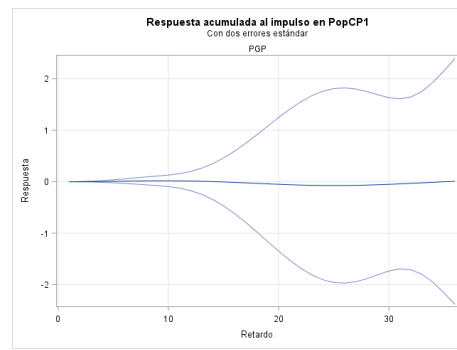
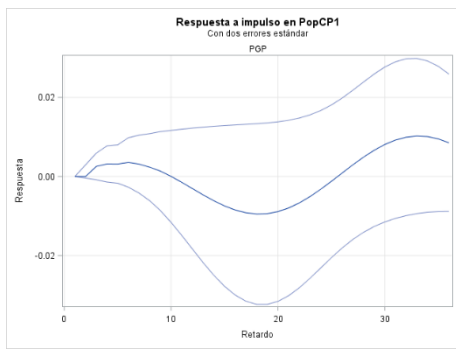


Gráfico 5.22. Respuesta a Impulso en MortCP de PGP

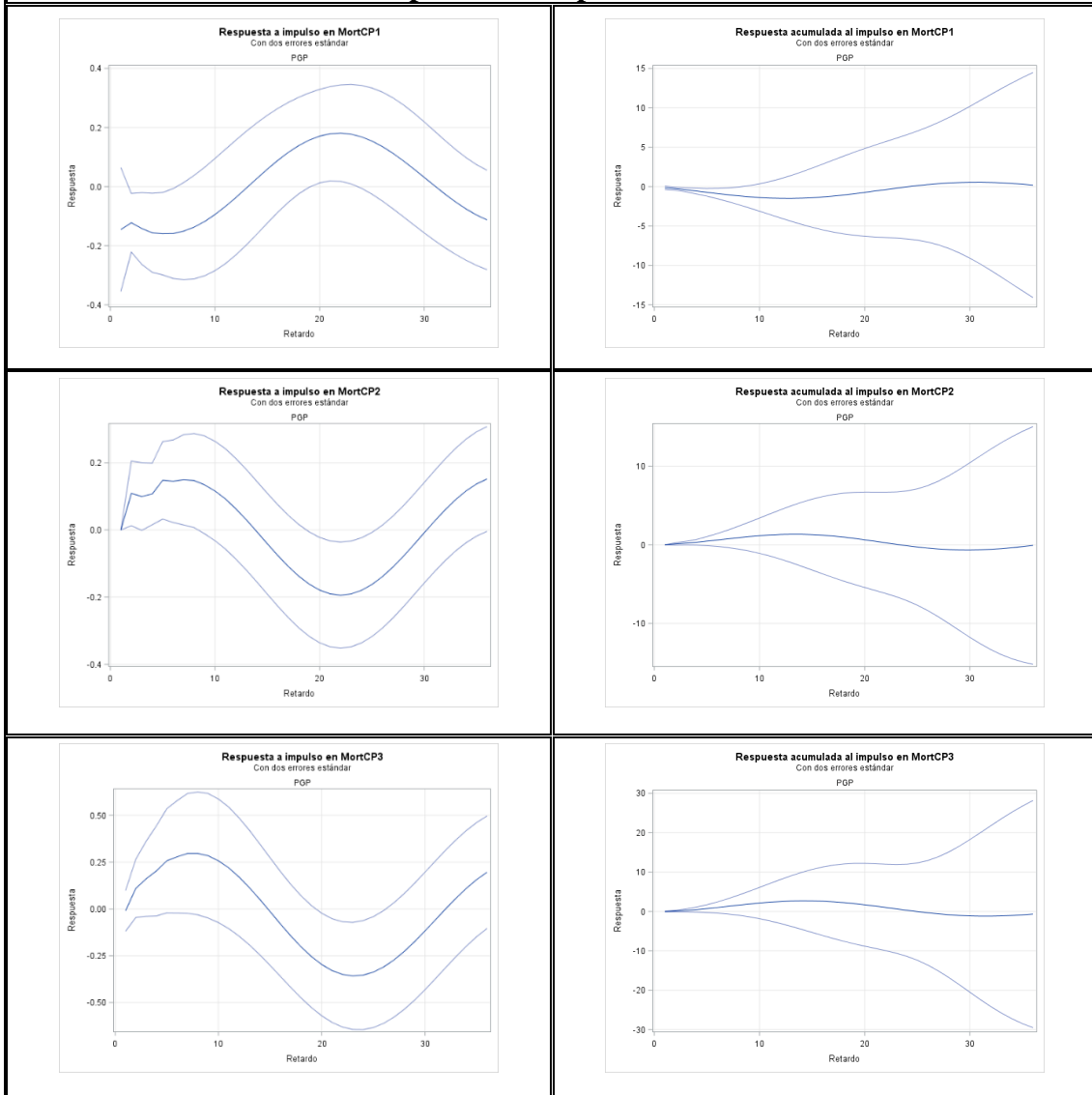


Gráfico 5.23. Respuesta a Impulso en PGP de PGP

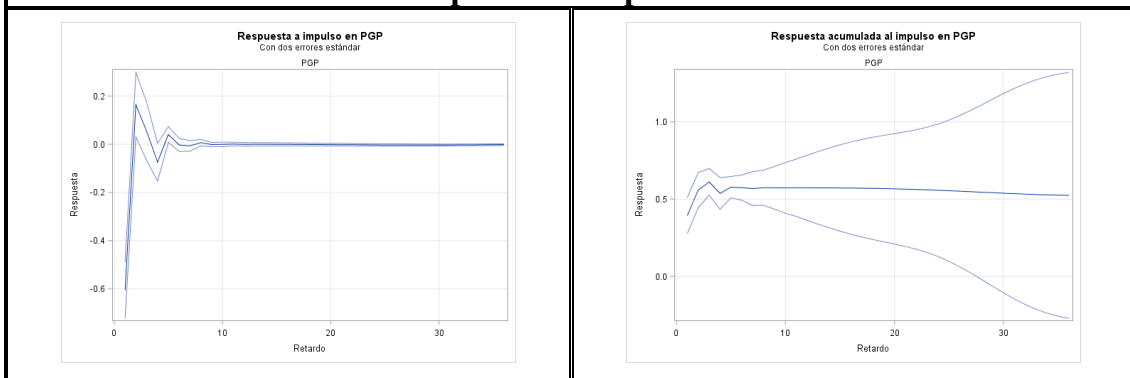
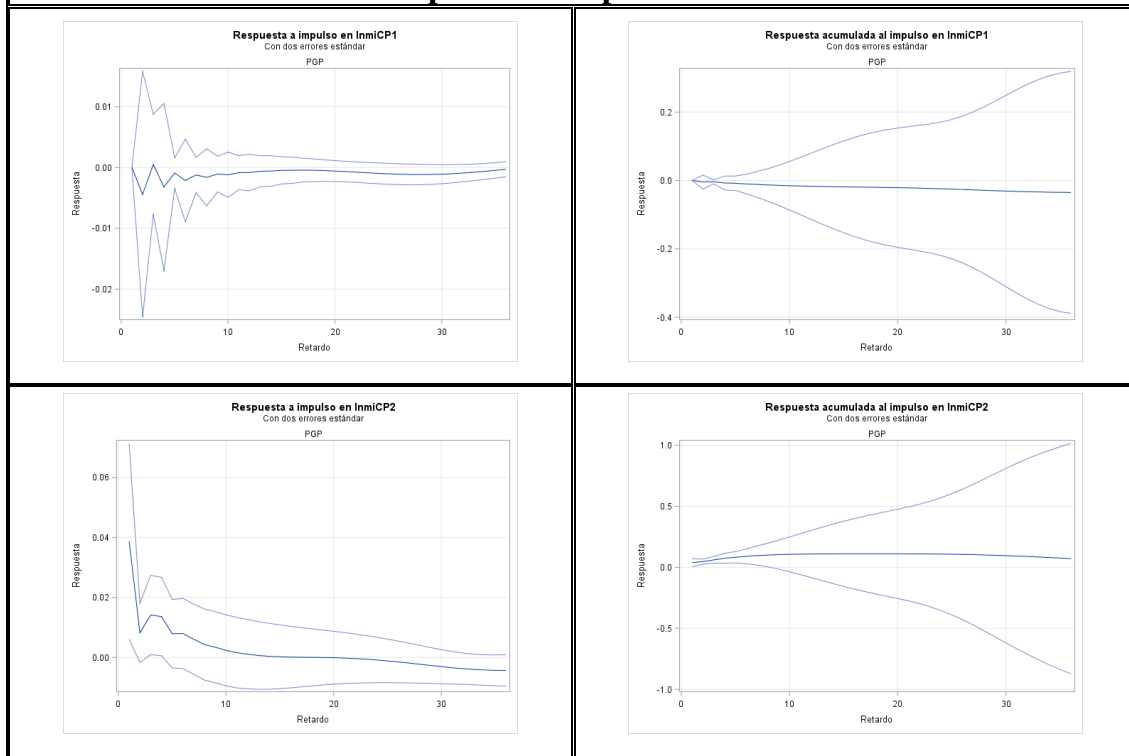


Gráfico 5.24. Respuesta a Impulso en InmiCP de PGP



Un ejemplo concreto de la interpretación de estos resultados es el siguiente: teniendo en cuenta que un valor cercano a 1 en InmiCP1 se asocia a un inmigrante en edad de trabajar, interpretaremos que la llegada de inmigración joven produce una disminución del porcentaje de gasto del PIB en pensiones. Por otro lado, un valor cercano a 1 en CP2 se asocia a un inmigrante mayor de 60 años, luego la llegada de inmigración de más de 60 años produce un aumento del porcentaje de gasto del PIB en pensiones.

4.3. Comparativa y Selección del Mejor Modelo

El principal objetivo de nuestro estudio es pronosticar el comportamiento futuro del porcentaje de gasto del PIB que se dedica a pensiones, luego para determinar qué modelo es el elegido nos guiaremos principalmente por la calidad de las predicciones.

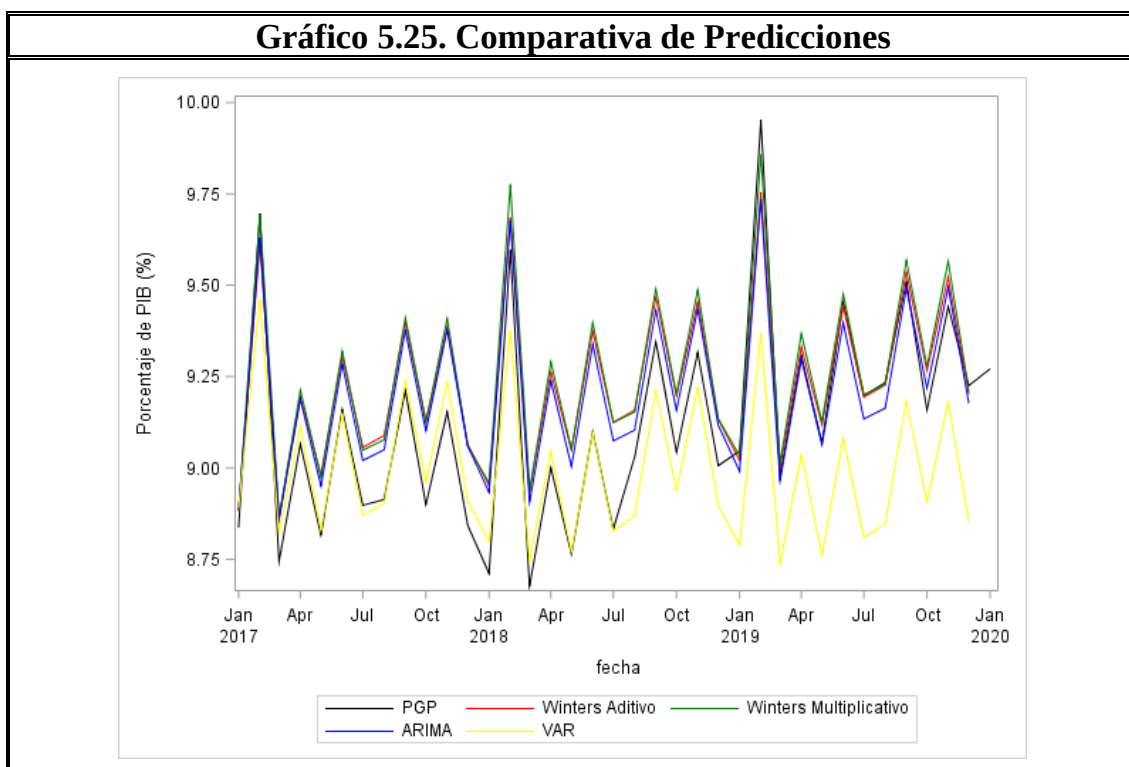
Los estadísticos más comunes para comparar series temporales son el MAE (*Mean Absolute Error* o Error Absoluto Medio) y el RSME (*Root Mean Squared Error* o Error Cuadrático Medio). Los modelos serán mejores cuanto menor sea su MAE o su RMSE. Matemáticamente se definen como:

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

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

Tabla 5.25. Métricas sobre el conjunto Test				
Métrica	H-W Aditivo	H-W Multiplicativo	ARIMA	VAR
MAE	0.0257795	0.0278535	0.0202728	0.0463431
RSME	0.3665130	0.3741546	0.3473594	0.4003582

En el Gráfico 5.25 vemos que para las primeras observaciones el modelo VAR es el que mejor se ajusta, pero a largo plazo pierde su capacidad pronosticadora. Por el contrario, los modelos ARIMA y Holt-Winters son un poco peores en el corto plazo, pero mucho más ajustados a la realidad en el largo plazo. Luego teniendo en cuenta la capacidad pronosticadora a largo plazo escogeremos los modelos de Holt-Winters o el ARIMA.



Otro factor determinante a la hora de decidir el modelo a escoger es que para la variable PGP tenemos datos hasta mediados de 2020, mientras que los datos de flujo de inmigración o población solo están disponibles hasta diciembre de 2018. Por tanto, teniendo en cuenta la disponibilidad de datos más actuales con los que hacer las predicciones escogeremos los modelos de Holt-Winters o el ARIMA.

Por último, hay que tener en cuenta que uno de los objetivos secundarios es poder predecir cual será el impacto de la crisis del COVID-19 en el sistema de pensiones. Esto solo se puede llevar a cabo mediante el uso de variables de intervención. Los modelos de alisado exponencial no están implementados para el uso de variables de intervención, pero los métodos ARIMA y VAR si están implementados para ello.

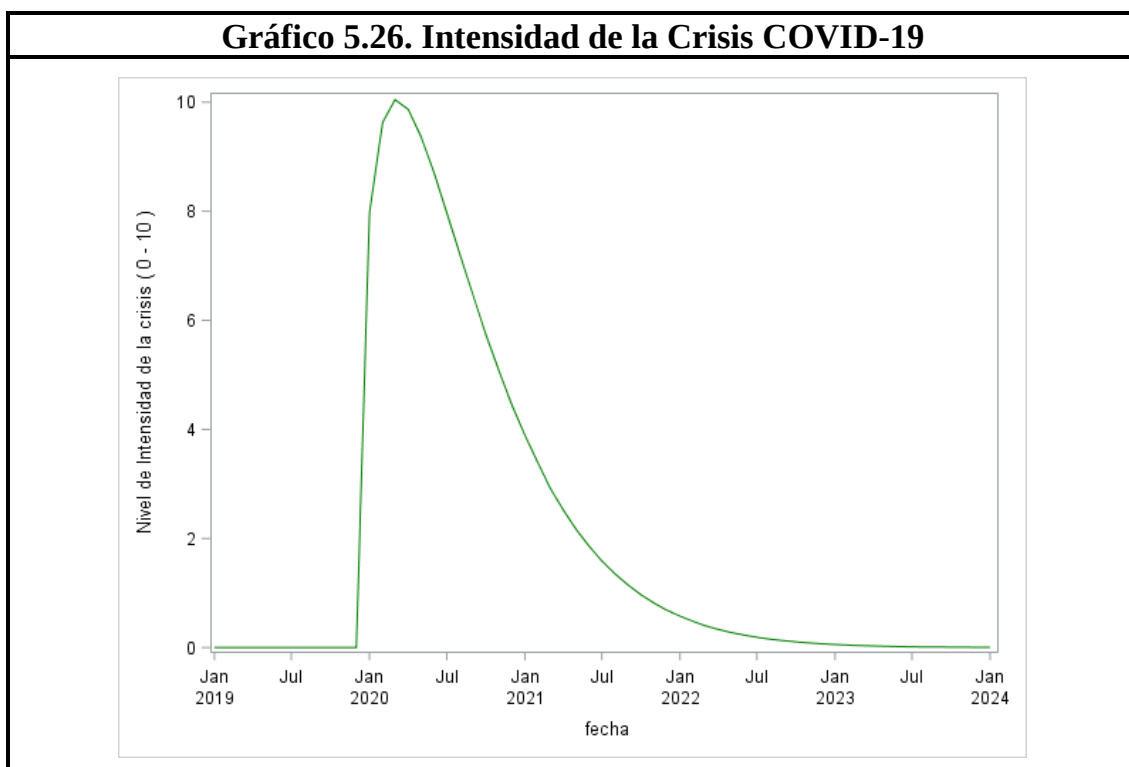
Teniendo en cuenta todos los factores anteriores, determinamos que el mejor modelo para realizar predicciones en nuestro estudio es el modelo ARIMA.

4.4. Predicción

Ahora calcularemos el modelo $ARIMA(7,1,0)(0,1,1)_{12}$ con todos los datos disponibles y variable de intervención para la crisis del COVID-19. Como actualmente desconocemos como va a ser la duración e intensidad de la crisis, vamos a plantear un escenario hipotético. Este escenario se basa en la información publicada en los últimos meses sobre cuál es el avance más probable de la pandemia. Los factores que se han considerado para elaborar este escenario son:

- Hasta el presente, la parte más dura de la crisis fue entre marzo y mayo de 2020.
- La mayoría de los sectores económicos y sociales continúan sufriendo los efectos de la crisis, aunque de manera menos intensa.
- En general los ciudadanos y las empresas están más concienciados y preparados para hacer frente a nuevos repuntes, por lo que de haberlos no tendrán la misma intensidad que la primera oleada.
- El desarrollo de tratamientos y vacunas se prevé que empiece a dar sus frutos a lo largo del año 2021.

Por lo tanto, nuestro escenario determinará de 0 a 10 el nivel de intensidad de la crisis en los próximos años. Como se observa el pico de intensidad de la crisis se produce en los primeros meses de 2020 y va descendiendo de manera progresiva hasta que sus efectos desaparecen a lo largo del 2022.



Para crear este escenario hemos asignado a cada uno de los 120 meses (10 años) que queremos pronosticar un valor x consecutivo del intervalo de 0 a 1, asegurándonos de que ningún mes tomará el valor 0 o 1.

Este proceso se ha conseguido asignando a enero de 2020 el valor de $i = 1$, a febrero de 2020 el valor de $i = 2$ y así sucesivamente. Y luego aplicando la función:

$$x = i/121$$

Finalmente utilizamos el valor obtenido para calcular la intensidad de la crisis utilizando la función de densidad de una $Beta(1.5,20)$, es decir,

$$Nivel\ de\ Intensidad_t = \begin{cases} 0 & si\ x < 0 \\ \frac{\Gamma(1.5 + 20)}{\Gamma(1.5)\Gamma(20)} x^{1.5-1}(1-x)^{20-1} & si\ x \in (0,1) \\ 0 & si\ x > 1 \end{cases}$$

siendo Γ la función gamma de Euler.

A continuación, se muestran los datos relativos a los residuos del modelo con variable de intervención para el COVID-19 y las predicciones obtenidas.

Tabla 5.26. Comprobación de autocorrelación de residuales (Box-Ljung)			
Para retardo	Chi-cuadrado	DF	Pr > ChiSq
6	.	0	.
12	4.50	4	0.3426
18	6.92	10	0.7327
24	10.67	16	0.8296
30	12.69	22	0.9414
36	17.70	28	0.9335
42	23.77	34	0.9049

Gráfico 5.27. Residuos $ARIMA(7, 1, 0)(0, 1, 1)_{12}$ con variable de Intervención

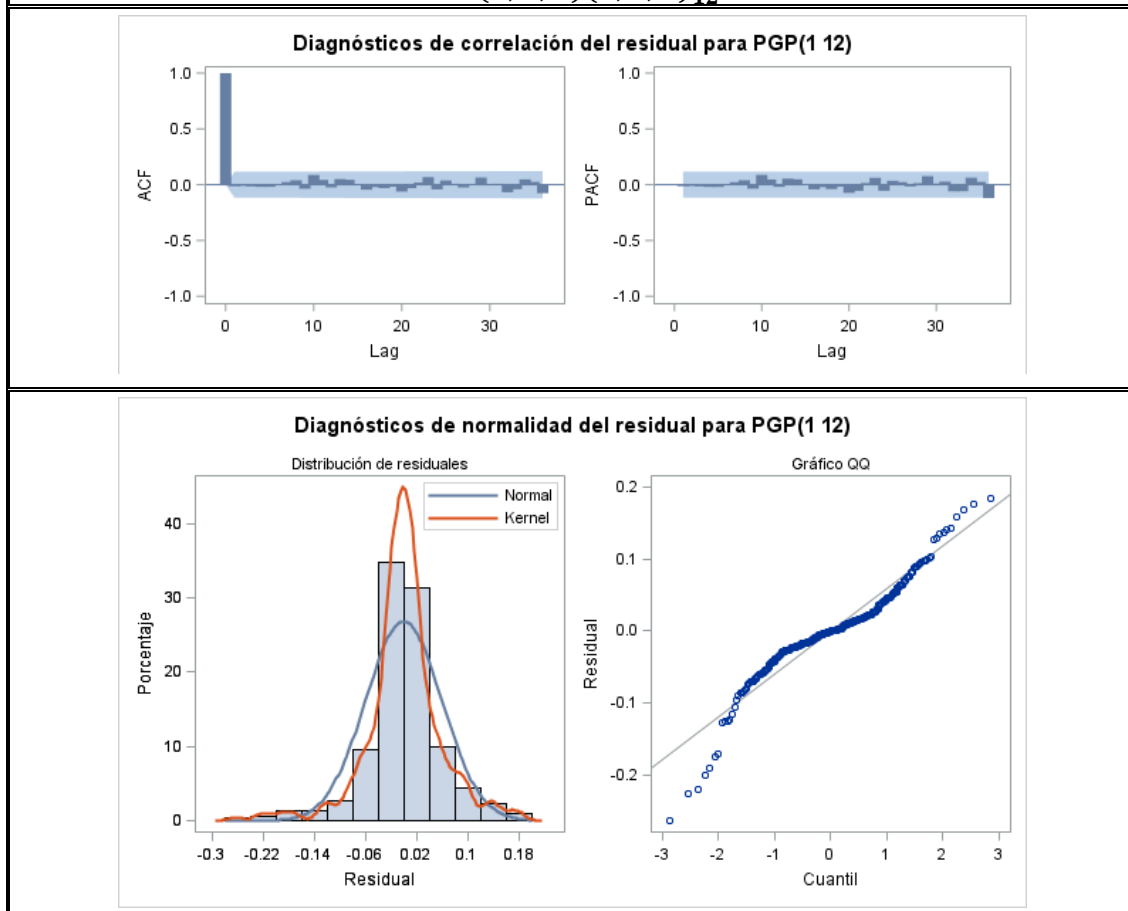
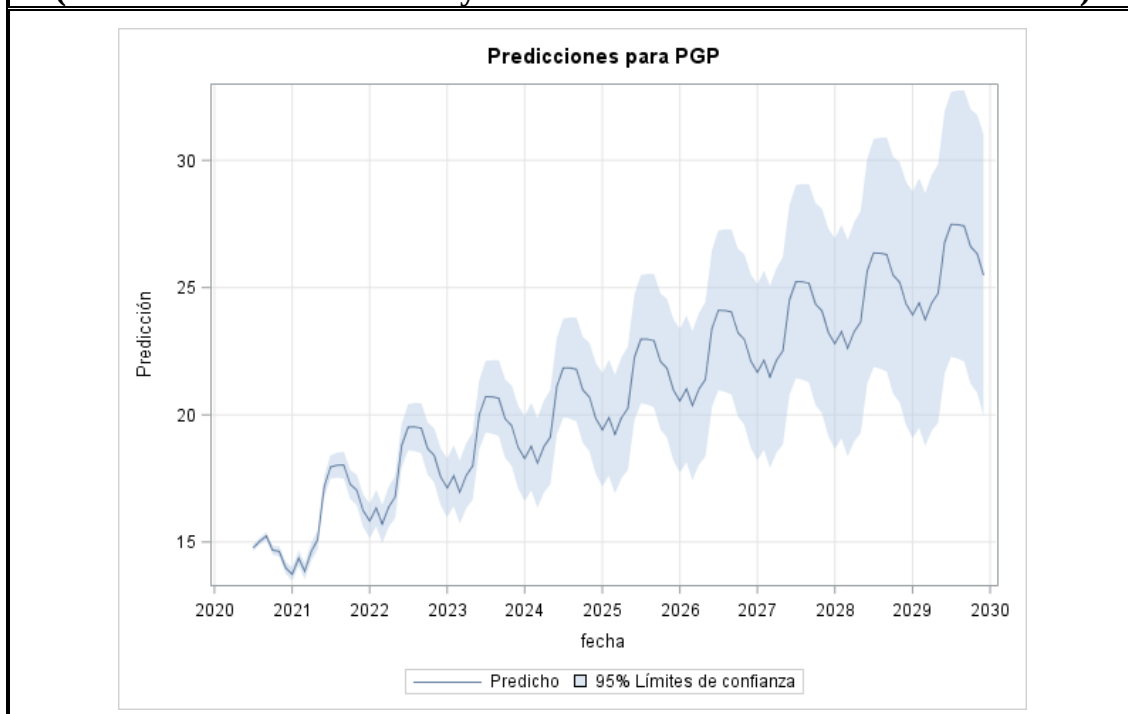


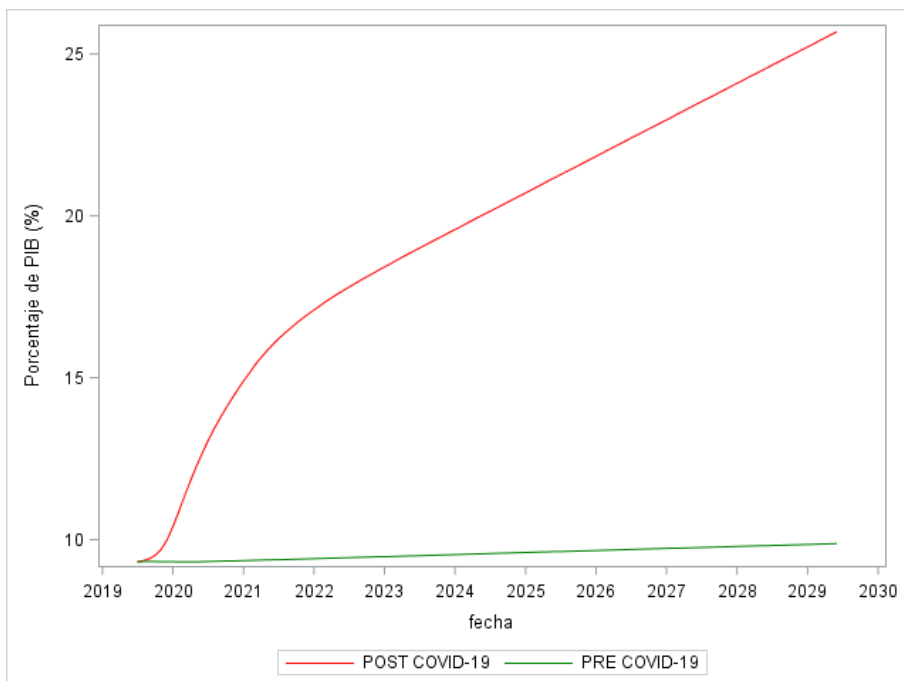
Gráfico 5.28. Predicciones del Gasto en Pensiones 2020-2030 (Datos hasta Junio de 2020 y con variable de Intervención COVID-19)



Vemos que el modelo ARIMA con variable de intervención cumple con los requisitos necesarios para ser un modelo válido. Evidentemente ahora es imposible evaluar la calidad de las predicciones, pues son sucesos que todavía no han ocurrido. Lo que si podemos valorar es que los intervalos de confianza para el modelo son bastante estrechos para los 5 primeros años.

Al ser unas predicciones a largo plazo la estimación puntual no es de gran importancia, ya que un intervalo de tiempo tan amplio contendrá muchos errores. Por eso vamos a quedarnos única y exclusivamente con la tendencia de la serie.

Gráfico 5.29. Comparativa de la Tendencia del Gasto en Pensiones 2020-2030



Nuestro escenario recoge una crecida del porcentaje de PIB destinado pensiones, que se plasma en un aumento en la tendencia muy brusco durante el primer año para luego continuar su subida de forma menos acentuada.

6. DISCURSIÓN Y CONCLUSIONES

A lo largo de este trabajo, hemos desarrollado diferentes métodos para el estudio de series temporales con el objetivo de pronosticar el porcentaje de PIB que se destina al gasto en pensiones en España. Dos de los modelos, Holt-Winters y ARIMA, se basan en las técnicas univariantes, mientras que el modelo VARMA pertenece a la metodología multivariante. De los tres métodos usados, solamente Holt-Winters no se basa en la estacionariedad de los datos.

Para examinar el rendimiento de cada modelo hemos comparado las predicciones que realizan para el periodo entre enero de 2017 y diciembre de 2019. Los tres métodos pudieron pronosticar, con errores razonables, la dirección y magnitud del crecimiento del porcentaje de PIB destinado a cubrir el gasto en pensiones. El mejor método de predicción fue el proporcionado por el modelo ARIMA, aunque los modelos de Holt-Winter también obtuvieron unos resultados muy competitivos.

Aunque en general se suele esperar que los métodos multivariantes generen pronósticos más precisos que los univariantes, los resultados de este estudio muestran que no siempre es así. Una explicación plausible es que la adicción de más parámetros al modelo también introduce más posibilidades de introducir errores en su estimación, y esos errores se van acumulando a medida que aumenta el intervalo de predicción. Otra explicación podría ser la falta de más variables que ayuden a explicar correctamente la estructura de correlaciones y autocorrelaciones.

Esta última explicación concuerda con las limitaciones que hemos sufrido en este trabajo para acceder a series de datos actuales, con muchas variables y extensas en el tiempo. Evidentemente, esto no ha ayudado a poder explicar adecuadamente las complejas relaciones que existen entre la variable objeto de estudio y los factores económicos y demográficos de nuestro país.

Dada la inusual situación de incertidumbre que ha traído el COVID-19 a todos los aspectos de la vida, la realización de proyecciones económicas se ha vuelto todavía más complicada. Por ello nos hemos tenido que valer del uso de una variable de intervención para poder incluir el efecto de la actual crisis en nuestro modelo ARIMA. Para la creación de esta variable auxiliar se ha supuesto que las medidas tomadas irán surtiendo efecto poco a poco y que no habrá una nueva oleada de infecciones tan grande que nos lleve a nuevos confinamientos totales. También se ha asumido que las medidas de distanciamiento social repercutirán en la marcha de la economía hasta que no aparezca una solución definitiva.

Los resultados concluyen que el porcentaje de PIB que se destina a pensiones aumentará de manera brusca durante todo 2020. Esto conllevará que pasemos de un gasto inferior al 10% a finales del 2019, a uno superior al 15% a lo largo del 2021. Después de 2021 se mantiene la tendencia creciente, pero a un menor ritmo, lo que nos situará por encima del 20% en el año 2025.

A comienzos del estudio estipulamos que el sistema estaba en riesgo en la banda del 13%-15% y que por encima del 15% se encontraría en situación de insostenibilidad. Viendo los datos de nuestro modelo, está claro que el COVID-19 ha acelerado de forma evidente

el deterioro de nuestro sistema público de pensiones. Por tanto, para garantizar su continuidad deben tomarse decisiones rápidas que impliquen cambios profundos.

En este estudio no realizamos diferentes escenarios hipotéticos que puedan medir la sensibilidad a cambios en el transcurrir de la crisis, ni tampoco se intentó identificar cuáles son todas las causas que impulsan las fluctuaciones en el porcentaje de PIB que se emplea en pensiones. Esto se debe a que el objetivo era identificar el mejor método de predicción con los datos de los que disponíamos. Sin embargo, sí que se proponen ambas líneas de investigación para futuros trabajos que amplíen la información aquí obtenida.

En resumen, el resultado general de este estudio sustenta los postulados de aquellos que argumentan la necesidad de cambios estructurales en nuestro sistema de pensiones para garantizar su sostenibilidad. También los resultados del modelado y la predicción apoyan la utilización de modelos univariantes como métodos de pronóstico fiables, en especial cuando no se disponen de suficientes datos multivariantes que puedan garantizar el reconocimiento de la estructura de relaciones subyacentes entre los mismos.

7. BIBLIOGRAFÍA

ABOAGYE-SARFO, P., MAI, Q., SANFILIPPO, F., PREEN, D., STEWART, L., FATOVICHE, D., 2015, “*A comparison of multivariate and univariate time series approaches to modelling and forecasting emergency department demand in Western Australia*”, Journal of Biomedical Informatics, Volume 57, 62-73.

ALONSO REVENGA, J.M., 2020, “*Material de la asignatura Técnicas avanzadas de predicción del grado en Estadística Aplicada*”, UCM.

ALONSO REVENGA, J.M., 2018, “*Material de la asignatura Técnicas estadísticas multivariantes I del grado en Estadística Aplicada*”, UCM.

BLAKE, M., 2007, “*Robbing Peter to pay Paul? A Box Jenkins analysis of the future of the UK pensions system*”, University of Lund, M. Sc. Thesis.

BOX, G., JENKINS, G., 1976, “*Time series analysis: forecasting and control*”, Holden Day, San Francisco.

BRIO CARRETERO, J. F., 2015, “*Tres enfoques de análisis de la sostenibilidad del sistema de pensiones español: la tasa interna de retorno, una proyección del balance financiero y un modelo de equilibrio general dinámico*”, UNED, Tesis doctoral.

CONDE-RUIZ, J. I., GÓNZALEZ C. I., 2018, “*European pension system: bismarck or beveridge?*”, FEDEA, Documento de trabajo.

DEVESA, J. E., DEVESA, M., 2009, “*El coste y el desequilibrio financiero-actuarial de los sistemas de reparto. El caso del sistema español*”, IVIE, Serie EC, WP 2009-09.

DOMÉNECH, R., MELGUIZO, A., 2009, “*Projecting pension expenditures in Spain: on uncertainty, communication and transparency*”, BBVA, Economic research department, N° 0911.

GOOIJER, J. G., HYNDMAN, R. J., 2006, “*25 years of time series forecasting*”, International Journal of Forecasting, Volume 22, Issue 3, 443-473.

LESLIE, P. H., 1945, “*On the use of matrices in certain population mathematics*”, Biometrika, Volume 33, Issue 3, 183-212.

PASTOR, A., VILA, M., 2019, “*AIReF model for forecasting pension expenditure in Spain*”, AIReF, WP/2019/1.

PEÑA, D., 2005, “*Análisis de series temporales*”, Alianza Editorial, Madrid.

SIMIONESCU, M., 2013, “*The Use of VARMA Models in Forecasting Macroeconomic Indicators*”, Economics & Sociology, Vol. 6, No 2, 94- 102.

SOLE, M., SOUTO, G., PATXOT, C., 2018, “*Sustainability and adequacy of the spanish pension system after the 2013 reform: a microsimulation análisis*”, IEF, Hacienda Pública Española / Review of Public Economics, 228-(1/2019): 109-150.

TIAN, Y., ZHAO, X., 2016, “*Stochastic forecast of the financial sustainability of basic pension in China*”, Sustainability, 2016, 8, 46.

2018, “*The Swedish pension system and pension projections until 2070*”, Regeringskansliet.