

# Análisis de Series Temporales: Modelos SARIMA

Sergio García Prado

[sergio.garcia.prado@alumnos.uva.es](mailto:sergio.garcia.prado@alumnos.uva.es)

5 de enero de 2019

## Resumen

El objetivo de este trabajo es el análisis de una serie temporal univariante. En este caso, se utilizan modelos *SARIMA*, los cuales se refieren a la extensión con estacionalidad de modelos *ARIMA*. Estos permiten la modelización de series temporales a partir de combinaciones lineales de la componente determinista (parte autoregresiva) y la componente estocástica (parte de media móvil) de la serie. Para el análisis se utilizará la serie `weightloss`.

- **Archivo:** `weight-loss.csv`
- **Serie:** Frecuencia de búsquedas para la palabra clave “Weight loss” a través del buscador *Google* por meses, desde *Enero de 2004* hasta *Diciembre de 2018*. Los valores han sido estandarizados en el rango [0, 100].

## 1. Etapa de identificación

### 1.1. Contexto

Tal y como se indica al comienzo de este documento, se va a analizar la serie temporal referida a la frecuencia de búsqueda de la palabra clave “Weight loss” (a nivel mundial) a través del buscador *Google*. Se ha escogido esta serie para el trabajo por su componente estacional claramente marcada. Se cree que esta está estrechamente relacionada con un índice sobre la preocupación de la población por su peso a lo largo del tiempo.

Para evitar problemas de privacidad, los datos se proporcionan estandarizados en el rango [0, 100], lo cual elimina la escala de los mismos y únicamente permite estudiar la estructura de la serie. Esto no es un problema para el análisis que se realizará en este trabajo, dado que precisamente el objetivo del mismo es el de analizar la estructura de una serie temporal, siguiendo la metodología de *Box-Jenkins*.

En cuanto al particionamiento de los datos, estos se proporcionan en agrupaciones mensuales. Dado que se tiene información desde *Enero de 2004* hasta *Diciembre de 2018*, es decir, un total de *15 años*, lo cual suma  $15 * 12 = 180$  observaciones en total. Con esta cantidad de observaciones, se cree que se podrá construir un modelo *SARIMA* (*ARIMA* con estacionalidad) de manera adecuada.

Una vez introducido el contexto de los datos pertenecientes a la serie que se analizará, lo siguiente es empezar a describir la misma a nivel de su estructura estocástica. Tras describir la misma, se procederá a realizar las diferenciaciones pertinentes hasta conseguir que esta sea estacionaria. Una vez se haya conseguido transformar la serie en estacionaria, se tratarán de identificar los parámetros de la parte autoregresiva y de la parte de media móvil, tanto de la dependencia entre observaciones a nivel serial (cada observación con las anteriores), como de la dependencia estacional (cada observación con las anteriores dentro de su periodo estacional). Tras dicha descripción, se propondrán un conjunto de modelos *SARIMA*. En la Sección 2 se prodecerá al ajuste de dichos modelos a los datos. Posteriormente, en la Sección 3 serán descartados aquellos modelos que no puedan validarse por su excesiva falta de ajuste, sobre ajuste, parámetros no significativos, etc. De entre los modelos válidos, se seleccionará aquel cuyo ajuste sea el más próximo a los datos, lo cual se comprobará mediante distintas técnicas. Finalmente, en la Sección 4 se realizará una predicción para el próximo año (2019) sobre los valores esperados por el modelo seleccionado.

La metodología que se ha expuesto en el párrafo anterior se corresponde con la propuesta por *Box-Jenkins* para series temporales basada en ajuste de modelos *ARIMA*. En el documento, se sigue un enfoque en paralelo en lugar de iterativo para la búsqueda del mejor modelo para facilitar la interpretación y la organización del mismo. Esta es la única modificación que se ha llevado a cabo respecto de la metodología original.

## 1.2. Análisis Descriptivo

Tras la descripción de la metodología, se va a comenzar con la descripción de la serie temporal. Para ello, nos vamos a apoyar en los gráficos de la Figura 1, a partir de los cuales se puede tener una perspectiva completa acerca de la serie. A través de ella se puede ver el gráfico de la serie, el correlograma, el correlograma parcial, el periodograma y el diagrama de dispersión *rango-media*.

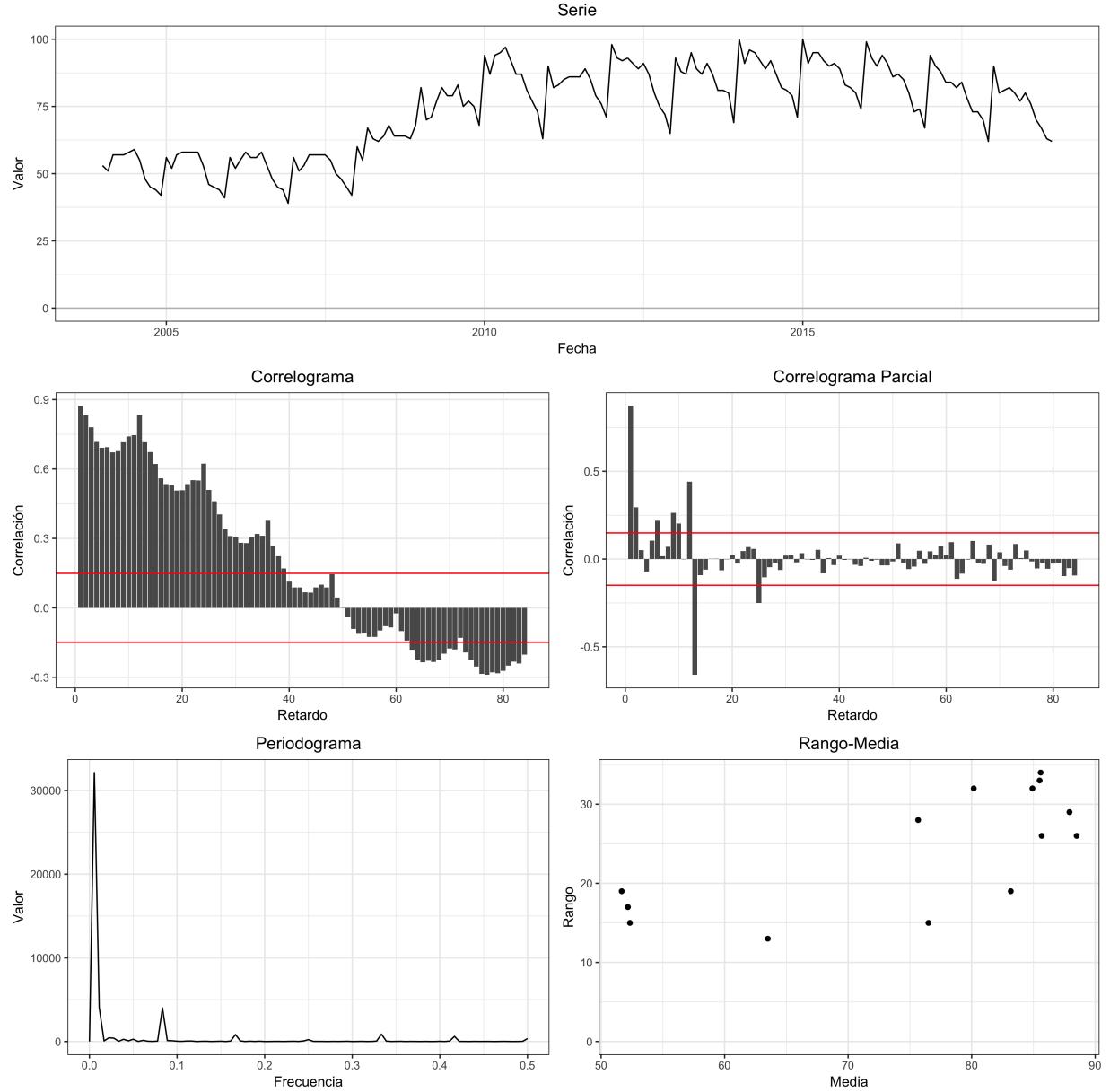


Figura 1: Gráfico de la serie, Correlograma, Correlograma Parcial, Periodograma y diagrama de dispersión *rango-media* para la serie **weightloss**.

En el gráfico de la serie de la Figura 1 se puede apreciar la evolución temporal de los valores a lo largo de los 15 años, separados en observaciones mensuales. En dicha representación destacan dos características de la serie temporal sobre el resto. Estas son: (1) la marcada estructura estacional de periodo 12 (anual) de la serie, que sigue la misma forma en todas las estacionalidades (años). Esto es un fuerte crecimiento durante el primer mes (Enero), posteriormente se produce un suave decrecimiento hasta el tercer cuarto del año, para producirse un fuerte decrecimiento en torno a los meses de Septiembre - Octubre. Esta estructura estacional es coherente con el fenómeno conocido como *Operación Bikini*, que consiste en la preocupación por estar en buena forma física durante los meses de verano. Esta preocupación comienza en torno a principios de año y se mantiene hasta los meses de verano. Puesto que una vez pasados dichos

meses, la forma física deja de estar comprometida, la preocupación por la pérdida de peso de la población también disminuye. (2) el cambio de nivel que se produce entre el año 2008 y el año 2009. Durante este periodo se produce un cambio drástico en el nivel de la serie. Parece que durante el final del año 2008 no se produjo el fenómeno esperado de un fuerte decrecimiento que sí se produce durante el resto de años. Este cambio en el nivel de la serie pudo deberse a distintos factores. Uno de ellos pudo ser la crisis económica cuyas evidencias comonezaron en torno a dicho año. Si se confirmase que las razones del aumento del nivel en las búsquedas del término *Weight loss* fueron debidas a la crisis económica, una interpretación para la misma podría ser la siguiente: Con el aumento del riesgo en la estabilidad financiera, la población aumentó también su preocupación en su apariencia física, lo cual se ha mantenido hasta la actualidad. Otro de los factores de peso pudo haber sido el acercamiento de la tecnología y redes sociales a la mayoría de la población, que hasta entonces se había mantenido alejada de ella. Dado que muchas redes sociales permiten la posibilidad de incluir imágenes personales que otras personas pueden juzgar, el índice de preocupación por el nivel de forma física puede haberse incrementado en los últimos años debido a dicha razón. Sin embargo, estas interpretaciones queda fuera del objetivo de este trabajo.

En cuanto al correlograma de la serie que se muestra en la Figura 1, se puede apreciar la componente estacional de periodo 12 en la estructura de correlaciones. Destacan sobre el resto los retardos de la forma  $i \bmod 12 = 0$ , estos son los retardos 12, 24, 36, .... Estos se relacionan entre si presentando un decrecimiento lineal, por lo que son indicativo de que la serie no es estacionaria. Por lo tanto, tendremos que llevar a cabo al menos una diferenciación estacional para conseguir estacionarizar la serie. También llama la atención la gran cantidad de correlaciones con valores significativos, por lo que la tendencia podría estar ocultando algún otro comportamiento no visible a simple vista. Por lo tanto, la realización de una diferenciación regular también podría ser una buena estrategia para tratar de comprender en mayor medida la estructura de correlaciones de la serie. Como se verá en el siguiente párrafo, estas interpretaciones se ven reflejadas en el correlograma parcial.

En el correlograma parcial de la serie de la Figura 1 se representan las correlaciones entre observaciones de la serie, tratando de eliminar de estas la relación procedente de otros retardos. Es decir, el correlograma parcial trata de representar de manera aislada la correlación entre una observación y la del  $k$ -ésimo retardo posterior, eliminando la influencia del resto. En este caso destacan sobre el resto los retardos 1, 12 y 13. Se cree que el retardo 1 destaca sobre el resto debido a la tendencia de la serie mientras que el retardo 12 se debe a la estacionalidad de la misma. También se piensa que el retardo 13 es un reflejo del 1, en la estacionalidad anterior, de ahí la razón de que destaque de tal manera.

En cuanto al periodograma de la serie que se muestra en la Figura 1, se puede apreciar que el primer armónico recoge gran parte de la variabilidad de la serie, lo cual de nuevo vuelve a indicar que la tendencia de la misma puede estar ocultando información sobre la estructura estocástica de la serie. Por lo tanto, se cree que una diferenciación permitiría visualizar en mayor medida el comportamiento de la misma. También destacan (aunque de una manera mucho menos pronunciada) los armónicos de la forma  $i + 1/12$  con  $i \in 1, 2, \dots, 6$ , esto es  $1/12, 2/12, \dots, 6/12$ , lo cual es otro argumento de peso en favor de la componente estacional de periodo 12 (anual) de la serie.

En cuanto al diagrama de dispersión *rango-media* de la Figura 1, parece que existe una leve relación entre el nivel de la serie y la dispersión del mismo. Este fenómeno podría requerir de alguna transformación de estabilización de varianza, como las de la forma *Box-Cox*. La razón de ello es que los modelos que se ajustarán a la serie temporal requieren de estacionaridad en la serie (entre otros requisitos, la varianza debe ser constante a lo largo de la serie). Sin embargo, tal y como se verá posteriormente, las diferenciaciones eliminarán la relación entre nivel y dispersión, por lo que es algo de lo que no nos preocuparemos en gran medida.

### 1.3. Diferenciaciones

Anteriormente, se ha realizado un análisis descriptivo acerca de la estructura de la serie **weightloss**, tras el cual se llegó a la conclusión de que la serie no presentaba la propiedad de estacionaridad (nivel y estructura de correlaciones constante para todas las observaciones). Por tanto, se indicó que para llegar a una serie estacionaria, la solución sería encontrar una diferenciación o varias diferenciaciones que transformen la serie en estacionaria.

Tal y como se indicó anteriormente, para conseguir la propiedad de estacionaridad dichas diferenciaciones podrían ser de dos tipos: (1) diferenciación regular para reducir la componente de tendencia de la serie, y (2) diferenciación estacional de periodo 12 para reducir las correlaciones entre observaciones de un mismo índice estacionalidad. Entonces, la estrategia a seguir será la de probar entre distintas combinaciones de diferenciaciones (de orden reducido, ya que de no ser así entonces se producirían efectos tales como el aumento de la varianza de la serie).

Por tanto, para elegir el orden de diferenciación se han probado distintas alternativas. En la Tabla 1 se incluye una relación entre la varianza resultante de las observaciones de la serie y el orden de diferenciación de la misma. Como se puede apreciar, la serie original tiene una varianza de 258,67. Sin embargo, la de la serie tras aplicar una diferenciación regular y otra estacional es únicamente de 14,07.

Serie	Varianza
$X_t$	258.67
$\nabla X_t$	62.92
$\nabla_{12}X_t$	49.09
$\nabla^2_{12}X_t$	86.28
$\nabla\nabla_{12}X_t$	14.07
$\nabla\nabla^2_{12}X_t$	42.11
$\nabla^2\nabla_{12}X_t$	34
$\nabla^2\nabla^2_{12}X_t$	99.44

Tabla 1: Relación entre distintos órdenes de diferenciación y varianza de la serie resultante.

Tanto por la reducida varianza, como por la propiedad de estacionaridad, finalmente escogeremos los grados de diferenciación 1 para la componente regular y 1 para la componente estacional. Sin embargo, a continuación se van a estudiar un poco más en detalle los resultados de la serie tras la aplicación de distintas diferenciaciones: En la Subsección 1.3.1 se analiza la serie diferenciada regularmente y en la Subsección 1.3.2 se analiza la serie diferenciada estacionalmente. Por último, en la Subsección 1.3.3 se muestra la serie diferenciada tanto regular como estacionalmente. Tal y como se ha indicado anteriormente, esta será la serie con la que se continuará el análisis y posterior ajuste.

### 1.3.1. Diferenciación regular

La diferenciación regular consiste en la aplicación del operador  $\nabla$  una vez a la serie objetivo. Esto es equivalente a multiplicar la misma por el término  $(1 - B)$ , donde  $B$  consiste en el operador *backward*, que genera como resultado la observación anterior. En la Ecuación 1 se muestra la representación de la diferenciación regular. Esto consiste en substraer el valor de la observación actual a la anterior. Tras aplicar esta operación a todas las observaciones de la serie, se consigue la eliminación de una tendencia de carácter lineal. Nótese que para hacer esto, se están perdiendo 1 observación.

$$\begin{aligned}\nabla X_t &= (1 - B) \cdot X_t \\ &= X_t - X_{t-1}\end{aligned}\tag{1}$$

En la Figura 2 se muestran el gráfico de la serie, el correlograma, el correlograma Parcial, el periodograma y el diagrama de dispersión *rango-media* para la serie tras la realización de una diferenciación regular.

Tal y como se puede apreciar en el gráfico de la serie, se ha eliminado la tendencia de la misma. Ahora la media está en torno al valor 0, sobre el cual oscila. Sin embargo, se puede apreciar que la serie sigue teniendo una estructura estacional, que la diferenciación no ha podido eliminar. En el correlograma se puede apreciar que dicha estructura de estacionalidad (con las correlaciones múltiples de 12 claramente marcadas) que estas presentan un decrecimiento de carácter lineal, lo cual indica que la serie no es estacionaria. En el correlograma parcial se puede apreciar la reducción de la correlación del primer retardo, lo cual era algo obvio tras la diferenciación regular. En cuanto al periodograma, ahora se pueden apreciar de una manera muy marcada los armónicos significativos, los cuales siguen la forma  $i/12$ . Este es otro de los argumentos

en favor de la diferenciación estacional, que se realizará en la Subsubsección 1.3.2. En cuanto al gráfico de dispersión *rango-media*, se puede comprobar que ahora se ha conseguido controlar en mayor medida la relación entre el nivel y la dispersión, es decir, se ha estabilizado en mayor medida la varianza. Sin embargo, esto se conseguirá en mayor media en las diferenciaciones posteriores.

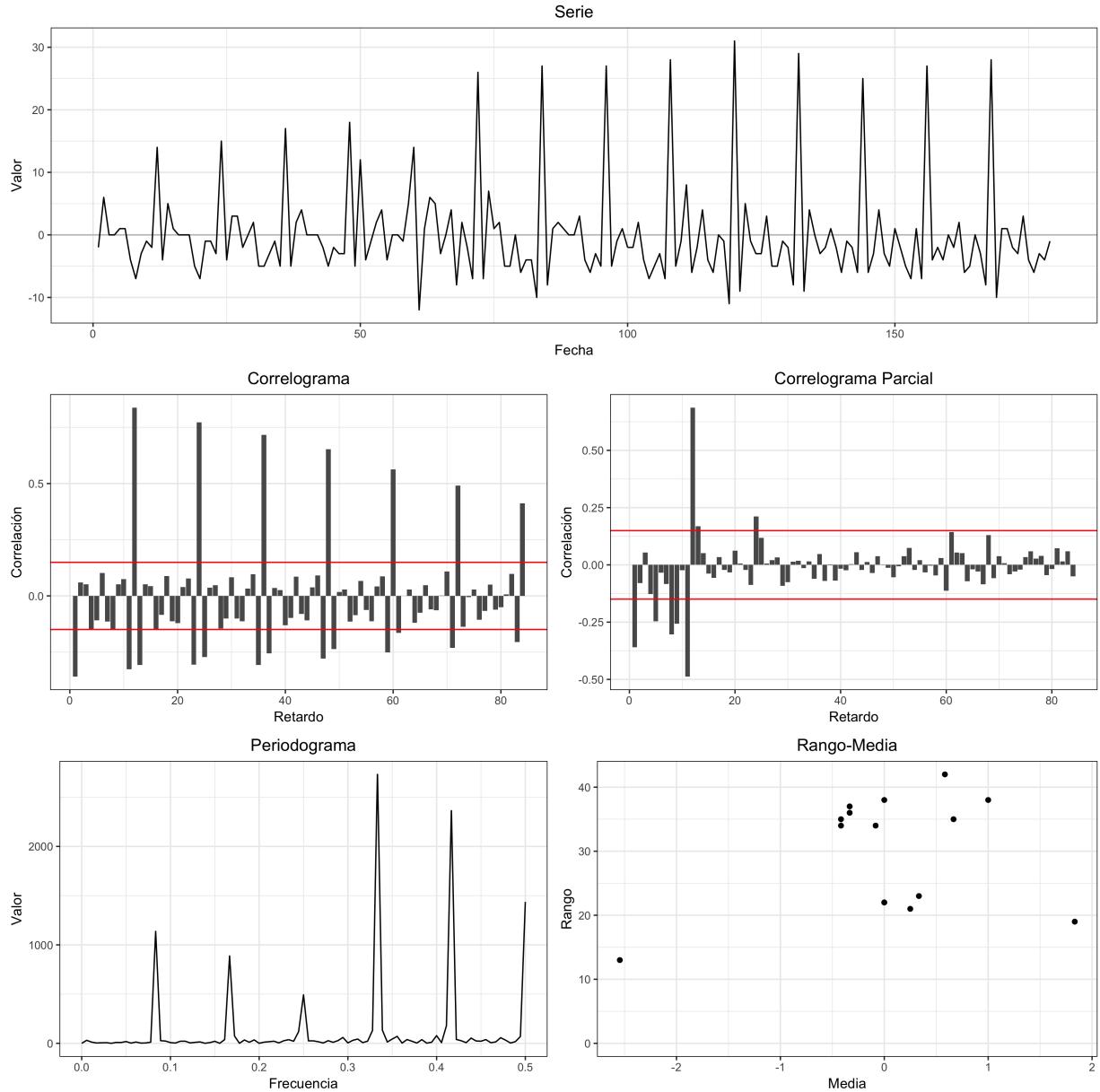


Figura 2: Gráfico de la serie, Correlograma, Correlograma Parcial, Periodograma y diagrama de dispersión *rango-media* para la serie **weightloss** tras la realización de una diferenciación regular.

La aplicación de la una diferenciación regular a la serie **weightloss** ha repercutido de manera positiva sobre la misma desde el punto de vista de la estacionariedad (ahora se encuentra más cerca de cumplir dicha propiedad). En la Tabla 1 se indica que el valor de la varianza es 62,92, lo cual es una drástica reducción en la dispersión de la serie. Sin embargo, todavía no puede ser considerada estacionaria. A continuación probaremos la alternativa estacional y estudiaremos los resultados.

### 1.3.2. Diferenciación estacional

La diferenciación estacional se basa en las mismas ideas que la diferenciación regular descrita anteriormente. Sin embargo, en este caso se utiliza una notación de subíndices para indicar el retardo en que aplicar la diferenciación. Esto es, en lugar de substraer de la observación actual la inmediatamente anterior, se

substraer la anterior respecto referida al mismo índice estacional. En nuestro caso, esto se refiere a substrair el valor del mismo mes del año anterior, esto es la observación 12 posiciones hacia atrás. En la Ecuación 2 se muestra la expresión de la diferenciación que se acaba de describir. Tal y como se puede apreciar, esta corresponde con la substracción de la observación 12 posiciones hacia atrás. Si se aplica esta operación a todas las observaciones de la serie, entonces se consigue la serie diferenciada estacionalmente. Nótese que para hacer esto, se están perdiendo 12 observaciones.

$$\begin{aligned}\nabla_{12} X_t &= (1 - B^{12}) \cdot X_t \\ &= X_t - X_{t-12}\end{aligned}\quad (2)$$

En la Figura 3 se muestran el gráfico de la serie, el correlograma, el correlograma parcial, el periodograma y el diagrama de dispersión *rango-media* para la serie tras la realización de una diferenciación estacional (12 retardos).

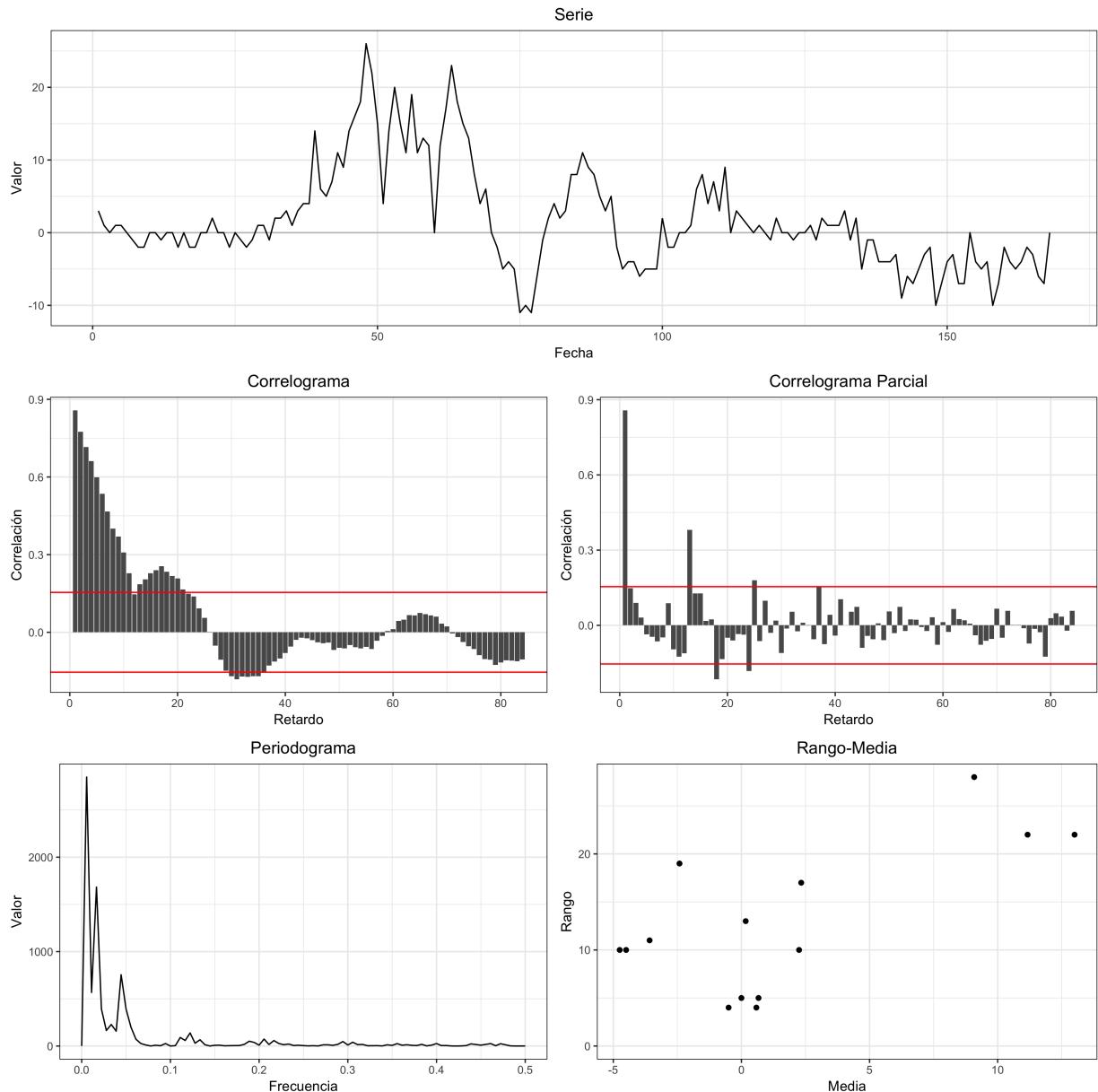


Figura 3: Gráfico de la serie, Correlograma, Correlograma Parcial, Periodograma y diagrama de dispersión *rango-media* para la serie `weightloss` tras la realización de una diferenciación estacional (12 retardos).

En el gráfico de la serie se puede apreciar la eliminación de la estructura estacional tras la serie, pero no la tendencia de la misma. Ahora la serie oscila en torno a valores próximos a cero, pero la tendencia hace que este comportamiento no pueda ser considerado estacionario. En el correlograma sucede algo similar: en este caso se ha conseguido eliminar la estructura lineal de decrecimientos en las correlaciones de la estacionalidad, pudiéndose considerar ahora decrecimiento exponencial. Sin embargo, la componente regular no tiene una estructura de decrecimientos exponenciales, lo cual hacer rechazar la hipótesis de que la serie sea estacionaria. En cuanto al correlograma parcial, se puede apreciar que se han reducido drásticamente las correlaciones referidas a los múltiplos de la estacionalidad, pero siguen destacando en gran medida las correlaciones en las posiciones 1 y 13. Se cree que la primera correlación se debe a la componente de tendencia de la serie, siendo la del 13-ésimo retardo un reflejo estacional de esta. En cuanto al periodograma, se puede apreciar que el primer armónico destaca en gran medida sobre el resto, lo cual se cree que es debido a la tendencia de la serie. En cuanto al diagrama de dispersión *rango-media*, se puede ver que la relación entre nivel y dispersión ha aumentado, lo cual es otro reflejo de que la serie resultante no es estacionaria.

La diferenciación estacional ha conseguido reducir en gran medida la componente estacional de la serie. En la Tabla 1 se indica que la varianza es 49,09, lo cual es una drástica reducción en la dispersión de la serie. Sin embargo ha dejado inalterada la tendencia de la misma, por lo que no se ha conseguido llegar a una serie estacionaria, que tomar de partida para el ajuste de un modelo autoregresivo con media móvil. Como se verá a continuación, la serie se transforma en estacionaria tras aplicar una combinación de las dos diferenciaciones realizadas previamente. Esto es, la diferenciación regular y la diferenciación regular, que aisla tanto la tendencia de la serie, como la estructura estacional de la misma.

### 1.3.3. Diferenciación regular y estacional

Tras analizar los resultados obtenidos al aplicar la diferenciación regular y la diferenciación estacional, para la cual se producen distintas modificaciones de la serie original que la dejan más cerca de la propiedad de estacionariedad, el siguiente paso es combinar ambas estrategias. Dado que la operación de diferenciación es conmutativa, no importa el orden de aplicación. En este caso, se ha desarrollado la expresión conjunta en la Ecuación 3.

$$\begin{aligned}\nabla \nabla_{12} X_t &= (1 - B) \cdot (1 - B^{12}) \cdot X_t \\ &= (1 - B) \cdot (X_t - X_{t-12}) \\ &= X_t - X_{t-1} - X_{t-12} + X_{t-13}\end{aligned}\tag{3}$$

En la Figura 4 se muestran distintos resúmenes visuales de la serie diferenciada regular y estacionalmente. En el gráfico de la serie se observa como la tendencia de la serie ha sido eliminada, al igual que la estacionalidad, lo cual hace que el comportamiento de la serie sea mucho más similar a un ruido blanco, aunque sin llegar a serlo, tal y como se puede apreciar a partir del resto de gráficos. Llama la atención el aumento de la dispersión en torno al primer cuarto de la serie, lo cual asumimos que podría darse al cambio de nivel durante finales del año 2008 en la serie original. En cuanto al correlograma, se puede apreciar un decrecimiento exponencial en las primeras correlaciones. De la misma manera, también se puede apreciar un decrecimiento exponencial en las correlaciones estacionales. Estas son las de la forma  $12i$ , es decir, las correlaciones múltiplo de 12. Dichos decrecimientos exponenciales son un indicador de que la serie resultante será estacionaria. Sobre el correlograma parcial, en este caso también se muestran decrecimientos exponenciales. Sin embargo, tanto este fenómeno como un análisis más detallado del correlograma, serán llevados a cabo en la Subsección 1.4 cuando se propongan distintos modelos *SARIMA*. En el periodograma se puede apreciar como destacan de una serie de armónicos sobre el resto. Sin embargo, en este caso no lo hacen de manera determinista, por lo que no pueden ser considerados una única estacionalidad que poder aislar de manera sencilla. Por lo tanto, lo trataremos como una componente aleatoria. Por último, a través del diagrama de dispersión *rango-media* se puede comprobar que la relación entre nivel y dispersión ha desaparecido. Entonces, ya podemos considerar que la serie resultante cumple la propiedad de *homocedastecidad*.

Dado que la serie cumple la propiedad de media y varianza constantes a lo largo del tiempo y además presenta decrecimientos exponenciales en el correlograma, consideraremos que la serie resultante es estacionaria. Además, la varianza de 14,07 es la menor que se ha obtenido, tal y como se puede apreciar en la

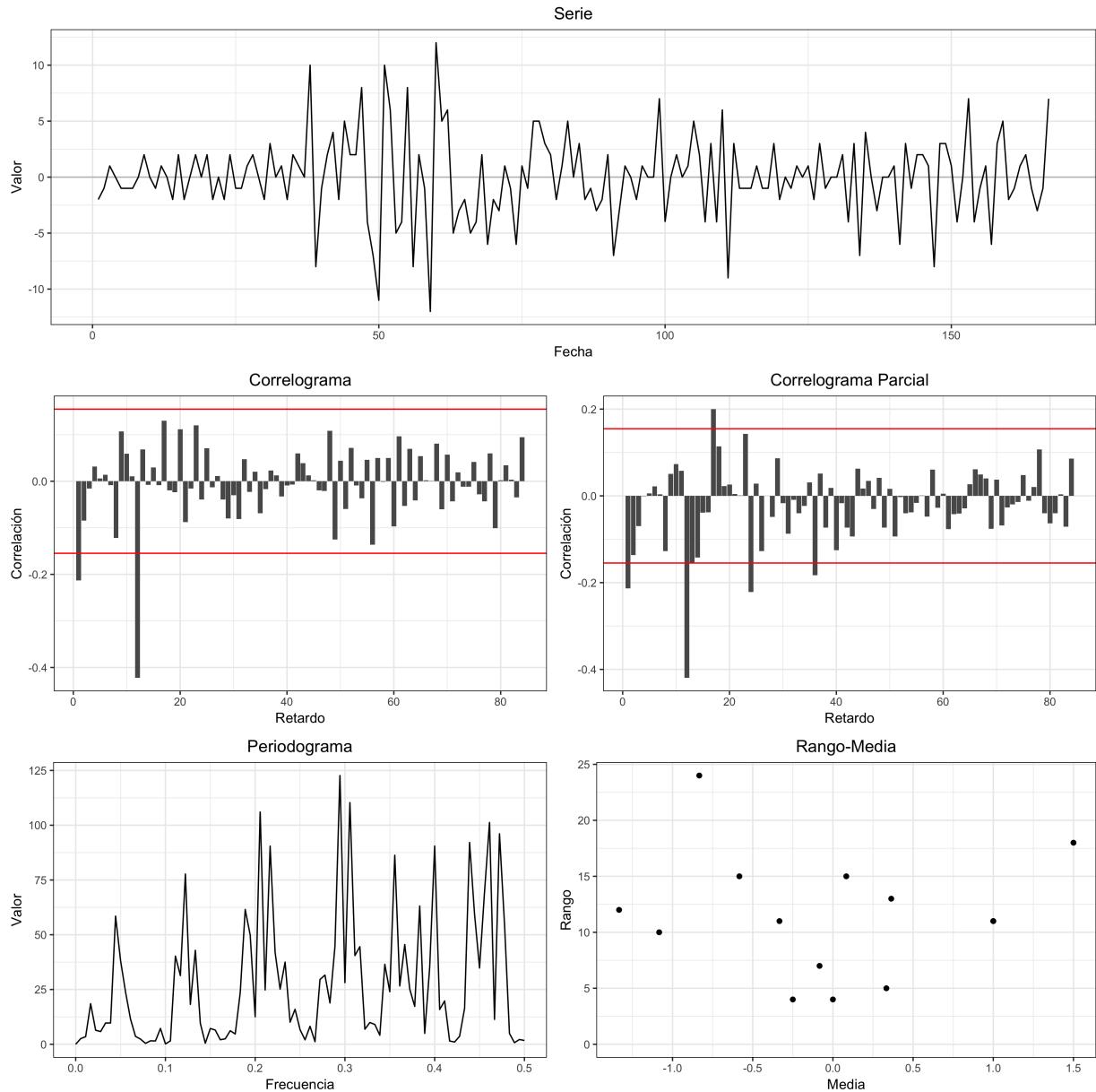


Figura 4: Gráfico de la serie, Correlograma, Correlograma Parcial, Periodograma y diagrama de dispersión *rango-media* para la serie `weightloss` tras la realización de una diferenciación regular y otra diferenciación estacional (12 retardos).

Tabla 1. Por estas razones, diremos que una diferenciación regular y otra estacional son suficientes para conseguir la propiedad de estacionaridad, necesaria para el ajuste de un modelo *SARIMA*.

Por último, se desea comprobar si la media de la serie puede ser considerada significativamente distinta de 0, en cuyo caso deberíamos ajustar un modelo con término independiente. Para ello, se ha llevado a cabo un test *t de student*, tras el cual se ha obtenido un valor  $t_{obs} = -0,061898$  lo cual indica el  $pvalor = 0,9507$ , no teniendo evidencias para el rechazo. Dicho test se incluye en el Código Fuente 3. Además, a continuación se incluye el resultado de dicho test:

Una vez realizado el análisis descriptivo de la serie, así como la búsqueda por los órdenes de diferenciación necesarios para conseguir estacionarizar nuestra serie, ya estamos en condiciones de analizar la serie desde otro punto de vista: la proposición de modelos, lo cual se expone a continuación.

```

One Sample t-test
t = -0.061898, df = 166, p-value = 0.9507
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
-0.5909606 0.5550325
sample estimates:
mean of x
-0.01796407

```

Tabla 2: Resultados del test *t de student* sobre la media nula ( $H_0 : \mu = 0$ ) para la serie *weightloss* diferenciada regular y estacionalmente.

#### 1.4. Modelos propuestos

Tal y como se ha dicho posteriormente, en este apartado se van a proponer distintos modelos para el ajuste de la serie. Estos se proponen teniendo en cuenta únicamente el análisis descriptivo de la serie original, así como de las diferenciadas. Puesto que ya se ha decidido el grado de diferenciación de la serie en la Subsubsección 1.3.2 este será un valor fijo en la proposición de modelos. Esto se debe a que el ajuste de este tipo de modelos requiere de la estacionarización de la serie original.

Para indicar los modelos se utilizará la notación que se describe en la Ecuación 4, donde la primera tupla se refiere a los parámetros regulares, mientras que las subsiguientes se refieren al periodo estacional  $s_i$ -ésimo en cada caso. El parámetro  $p$  indica el orden de la componente autoregresiva en cada tupla, el parámetro  $d$  se refiere al grado de diferenciación y por último, el parámetro  $q$  al orden de la componente de media móvil.

$$\text{SARIMA}(p, d, q)(P_1, D_1, Q_1)_{s_1}(P_2, D_2, Q_2)_{s_2} \dots (P_q, D_q, Q_q)_{s_q} \quad (4)$$

Puesto que anteriormente se indicó que era necesaria una diferenciación regular y otra estacional (de longitud 12) en la Subsubsección 1.3.3, sabemos que los modelos propuestos tendrán que ser de la forma  $\text{SARIMA}(p, 1, q)(P, 1, Q)_{12} \dots (P_q, D_q, Q_q)_{s_q}$ . Para proponer modelos, nos vamos a fijar en la Figura 4, que recoge distintos gráficos sobre la serie diferenciada regular y estacionalmente. En concreto, nos fijaremos sobre todo tanto en el correlograma como en el correlograma parcial, el cual sirve de guía a la hora de elegir órdenes de las componentes autorregresivas y de medias móviles.

El primer modelo que se propone es el indicado en la Ecuación 5, el cual se basa en la utilización de 1 parámetro de media móvil en la parte regular y otro en la parte estacional de periodicidad 12. La justificación detrás de este modelo es la siguiente: se pueden apreciar decrecimientos exponenciales en el diagrama de correlación parcial que comienzan en los retardos 1 y 12. Además, se dan valores elevados en el correlograma en las correlaciones 1 y 12, lo cual es implica que un modelo de esta forma podría ajustarse de manera coherente a la serie.

$$\text{SARIMA}(0, 1, 1)(0, 1, 1)_{12} \quad (5)$$

Se intuye que el modelo de la Ecuación 5 será el más adecuado para la serie *weightloss*. Sin embargo, a continuación se proponen otros modelos para los que se cree que existen distintas razones por las cuales se deberían comportar de manera adecuada.

El modelo de la Ecuación 6 consiste en una ampliación respecto del anterior. En este se ha añadido 1 parámetro de media móvil más, en el retardo 17. La razón por la cual se propone este modelo es el decrecimiento exponencial en el correlograma parcial de la Figura 4, que comienza en dicho retardo.

$$\text{SARIMA}(0, 1, 1)(0, 1, 1)_{12}(0, 0, 1)_{17} \quad (6)$$

Para completar la proposición de modelos que podrían ajustarse de manera adecuada a la serie, se añaden los indicados en la Ecuación 7 y en la Ecuación 8, los cuales son ampliaciones de los indicados en la Ecuación 5 y en la Ecuación 6 respectivamente. Estos añaden 1 parámetro autoregresivo a la componente regular de la serie. La justificación sobre este parámetro se debe al decrecimiento exponencial a partir del segundo retardo en el correlograma (ya que el valor del primer retardo proviene de la componente de media móvil). Si estos modelos finalmente son considerados válidos, entonces también habrá que considerar la primera correlación parcial como procedente de la componente autoregresiva regular. Sin embargo, se intuye que estos modelos finalmente serán invalidados.

$$\text{SARIMA}(1, 1, 1)(0, 1, 1)_{12} \quad (7)$$

$$\text{SARIMA}(1, 1, 1)(0, 1, 1)_{12}(0, 0, 1)_{17} \quad (8)$$

Nótese que no es adecuado proponer modelos de la forma  $\text{SARIMA}(1, 0, q)(1, 0, Q_{12})_{12} \dots (P_q, D_q, Q_q)_{s_q}$ , es decir, substituyendo las diferenciaciones por 1 parámetros autoregresivos. Lo que sucedería sería que se estaría incumpliendo la condición de estacionaridad. Además, los parámetros autoregresivos serían próximos a 1, ya que estarían tratando de hacer dicho papel. Una consecuencia de esto es que el error estándar de estos sería extremadamente elevado. Por tanto, se concluye que estos modelos no deberían ser propuestos

El siguiente paso es proceder al ajuste de los modelos propuestos, prestando especial atención a la significancia de los tests para los parámetros en cada modelo, así como a los residuos obtenidos, los cuales permitirán validar (e invalidar) aquellos modelos que no se ajusten de manera satisfactoria a los datos.

## 2. Etapa de estimación y validación

Una vez propuestos distintos modelos que podrían ajustar de manera adecuada a la serie, a continuación se procederá al ajuste de los mismos, para ello se utilizará el lenguaje *SAS*. Para ello, lo primero es cargar el conjunto de datos, lo cual se ha llevado a cabo mediante las sentencias del Código Fuente 4.

Una vez hecho esto, lo siguiente es ajustar los modelos propuestos en la Subsección 1.4. Además, se procederá al estudio de la validez de los mismos, descartando aquellos que no cumplan una serie de requisitos mínimos tales como significancia de sus parámetros, independencia y normalidad. Por lo tanto, procederemos a analizar (y validar si se cumplen las condiciones necesarias) los modelos propuestos.

### 2.1. Estimación y Validación de SARIMA(0, 1, 1)(0, 1, 1)\_{12}

Para el ajuste de este modelo en la etapa de validación, se ha utilizado el Código Fuente 5. Los parámetros ajustados se recogen en la Tabla 3. Estos son significativos tanto para el caso del parámetro de media móvil en el primer retardo (regular), como para el caso de duodécimo retardo (estacional). Los valores estimados son 0,33162 y 0,622239 respectivamente, con errores de 0,07282 y 0,06576.

Maximum Likelihood Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr >  t	Lag
MA1,1	0.33162	0.07282	4.55	<.0001	1
MA2,1	0.622239	0.06576	9.46	<.0001	12

Tabla 3: Estimación de los parámetros por el método de *Máxima Verosimilitud* para el modelo SARIMA(0, 1, 1)(0, 1, 1)\_{12}

En cuanto a la matriz de correlaciones entre los parámetros estimados, esta se muestra en la Tabla 4. Tal y como se puede apreciar la correlación entre estos es extremadamente baja ( $0,037 \approx 0$ ). Por tanto, se puede considerar que son incorrelados entre sí, lo cual elimina problemas de colinealidad (varios parámetros que indican lo mismo).

Correlations of Parameter Estimates		
Parameter	MA1,1	MA2,1
<b>MA1,1</b>	1.000	0.037
<b>MA2,1</b>	0.037	1.000

Tabla 4: Correlación entre los parámetros del modelo SARIMA(0, 1, 1)(0, 1, 1)<sub>12</sub>.

En la Tabla 5 se recogen distintos estadísticos acerca de la calidad de ajuste del modelo, útiles para realizar comparaciones entre distintos modelos. Puesto que el objetivo de esta sección es el ajuste y validación, se comentarán dichos resultados posteriormente.

<b>Variance Estimate</b>	9.253677
<b>Std Error Estimate</b>	3.041986
<b>AIC</b>	853.4882
<b>SBC</b>	859.7242
<b>Number of Residuals</b>	167

Tabla 5: Estadísticos de ajuste del modelo SARIMA(0, 1, 1)(0, 1, 1)<sub>12</sub>.

En la Figura 5 (generada a partir del Código Fuente 6) se resume la serie generada por los residuales del ajuste del modelo. A partir del gráfico de la serie se aprecia el comportamiento aleatorio de los residuales. Sin embargo, sigue presente el aumento de la varianza en torno al comienzo del primero cuarto de la serie (correspondiente al aumento de nivel durante finales del año 2008). Tanto en el correlograma, como en el correlograma parcial se aprecia un leve estrucutra de ondas, pero que se cree que puede considererarse no significativa por el reducido valor de las caorrelaciones (en todos los casos  $|< 0,2 |$ ). A pesar de ello, en el correlograma parcial destaca sobre el resto el retardo 17 (lo cual también sucedía en la serie de partida y por lo que se propuso el modelo de la Ecuación 6). En cuanto al periodograma, no se aprecia una estructura bien definida con armónicos que destaque sobre el resto. Finalmente, el diagrama de dispersión *rango-media* indica que no hay relación entre el nivel y la dispersión.

Por lo tanto, se cree que la serie de los residuales podría ser considerada un ruido blanco a partir de los gráficos de la Figura 5. Sin embargo, para confirmar dicha hipótesis se ha decidido realizar distintos contrastes.

```
Box-Ljung test
X-squared = 0.059792, lag = 1, p-value = 0.8068
```

Tabla 6: Resultados del test de *Ljung-Box* de dependencia serial en los residuales ajustados por el modelo SARIMA(0, 1, 1)(0, 1, 1)<sub>12</sub>

En la Tabla 6 y en la Tabla 7 se incluyen los tests de *Ljung-Box* acerca de la independencia serial de la serie de residuales, respecto del primer y duodécimo retardos respectivamente. En ambos casos aceptamos la hipótesis de independencia por no tener indicios suficientes para el rechazo (*pvalor* > 0,8 en ambos casos). Las sentencias necesarias para la reproducción de estos tests se incluyen en el Código Fuente 6

```
Box-Ljung test
X-squared = 5.9799, lag = 12, p-value = 0.9171
```

Tabla 7: Resultados del test de *Ljung-Box* de dependencia estacional en los residuales ajustados por el modelo SARIMA(0, 1, 1)(0, 1, 1)<sub>12</sub>

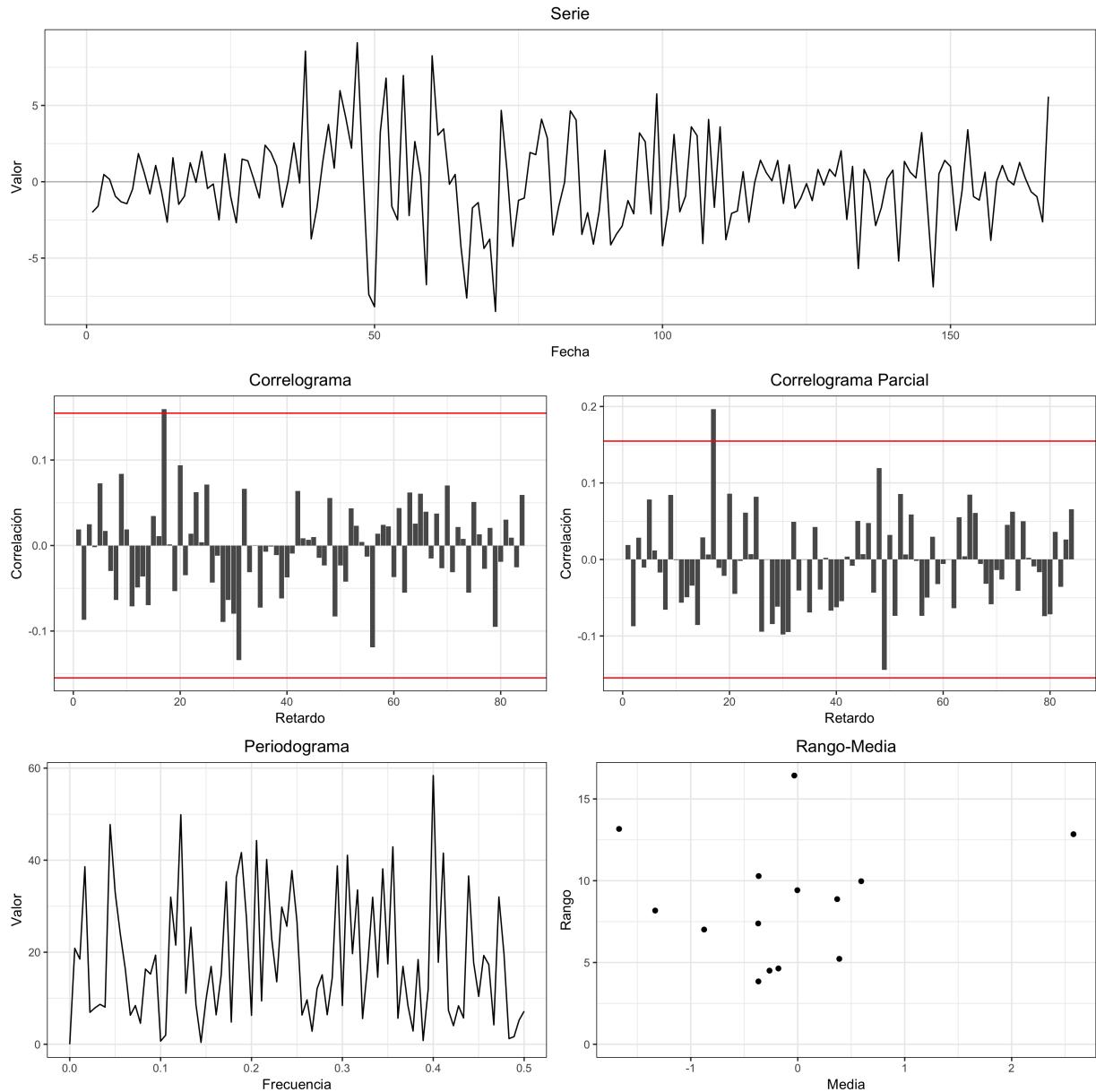


Figura 5: Gráfico de la serie, Correlograma, Correlograma Parcial, Periodograma y diagrama de dispersión *rango-media* para la serie de residuales ajustado por el modelo SARIMA(0, 1, 1)(0, 1, 1)<sub>12</sub>.

Otra de las hipótesis que deben cumplir los residuales de un modelo ajustado es la de normalidad. En la Figura 6 se muestra un gráfico *cuantil-cuantil* de normalidad. A partir de este, podemos apreciar un muy buen ajuste en la parte central de la distribución. Sin embargo, en los extremos se puede ver que el ajuste no es tan adecuado (aunque si que es simétrico respecto de cada lado). Se puede comprobar que dichas desviaciones se deben a que la distribución de los residuales es más apuntada que la normal (lo cual puede ser considerado un hecho positivo).

Como alternativa al gráfico *cuantil-cuantil* de la Figura 6, en la Tabla 8 y en la Tabla 9 referidas a los resultados de los tests de *Lilliefors* y *Shapiro-Francia* respectivamente. Las sentencias necesarias para la reproducción de estos tests se incluyen en el Código Fuente 6. Para el caso de *Lilliefors* el pvalor es 0,109, por lo que no rechazamos la hipótesis de normalidad (por poco). Sin embargo, para el caso de *Shapiro-Francia*, nos vemos obligados a rechazar dicha hipótesis. Este último es un test mucho menos conservador que el primero, por tanto, los resultados son coherentes.

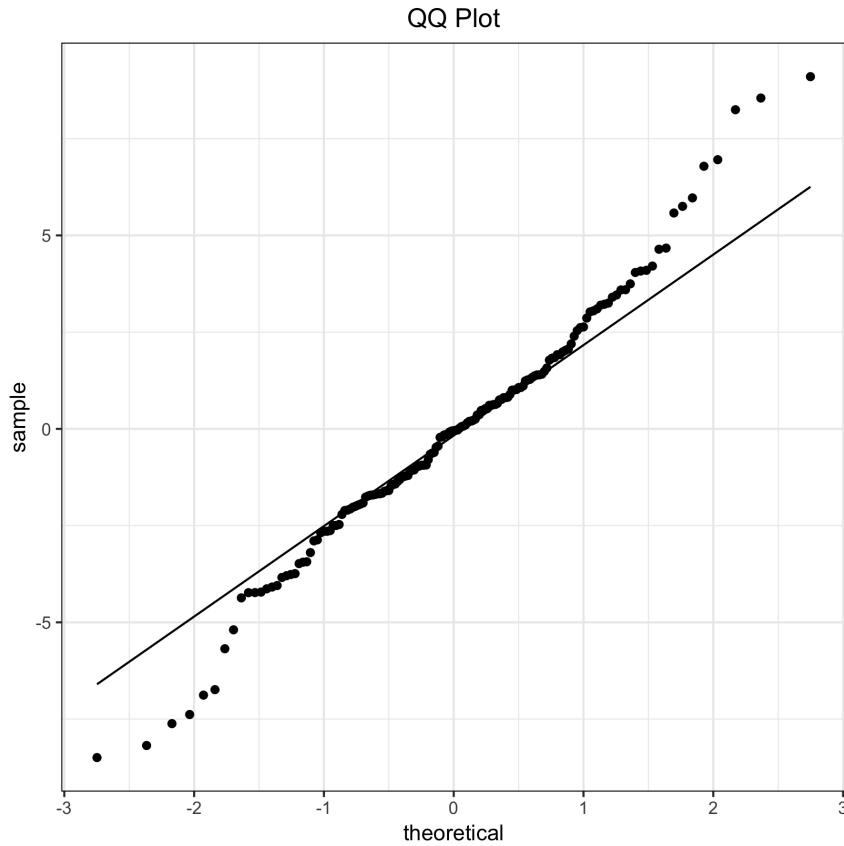


Figura 6: Gráfico de normalidad *cuantil-cuantil* para la serie de residuales ajustado por el modelo SARIMA(0, 1, 1)(0, 1, 1)<sub>12</sub>.

La razón por la cual estos valores son tan bajos, se cree que se debe al excesivo apuntamiento de la serie de residuales que provoca la falta de normalidad. Sin embargo, el ajuste se considera adecuado, por la independencia entre observaciones, así como la simetría de su distribución.

```
Lilliefors (Kolmogorov-Smirnov) normality test
D = 0.062944, p-value = 0.109
```

Tabla 8: Resultados del test de *Lilliefors* de normalidad en los residuales ajustados por el modelo SARIMA(0, 1, 1)(0, 1, 1)<sub>12</sub>

```
Shapiro-Francia normality test
W = 0.98217, p-value = 0.03111
```

Tabla 9: Resultados del test de *Shapiro-Francia* de normalidad en los residuales ajustados por el modelo SARIMA(0, 1, 1)(0, 1, 1)<sub>12</sub>

Puesto que el modelo cumple todas las características para ser considerado un modelo válido, en la Sección 3 será comparado con el resto de modelos válidos en la fase de selección de modelos. A pesar de no haberse podido confirmar la hipótesis de normalidad en los residuales (por el excesivo apuntamiento de la distribución), se considera un modelo válido.

## 2.2. Estimación y Validación de SARIMA(0, 1, 1)(0, 1, 1)<sub>12</sub>(0, 0, 1)<sub>17</sub>

Para el ajuste de este modelo en la etapa de validación, se ha utilizado el Código Fuente 7. Los parámetros ajustados se recogen en la Tabla 10. Estos son significativos para el caso del parámetro de media móvil en

el primer retardo (regular), para el caso de duodécimo retardo (estacional) y para el parámetro referido al décimo septimo retardo. Los valores estimados son 0,33629, 0,62106 y -0,15874 respectivamente, con errores de 0,07278, 0,06586 y 0,07817.

Maximum Likelihood Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr >  t	Lag
<b>MA1,1</b>	0.33629	0.07278	4.62	<.0001	1
<b>MA2,1</b>	0.62106	0.06586	9.43	<.0001	12
<b>MA3,1</b>	-0.15874	0.07817	-2.03	0.0423	17

Tabla 10: Estimación de los parámetros por el método de *Máxima Verosimilitud* para el modelo SARIMA(0, 1, 1)(0, 1, 1)<sub>12</sub>(0, 0, 1)<sub>17</sub>.

En cuanto a la matriz de correlaciones entre los parámetros estimados, esta se muestra en la Tabla 11. Tal y como se puede apreciar la correlación entre estos es extremadamente baja. Por tanto, se puede considerar que son incorrelados entre sí, lo cual elimina problemas de colinealidad (varios parámetros que indican lo mismo).

Correlations of Parameter Estimates			
Parameter	MA1,1	MA2,1	MA3,1
<b>MA1,1</b>	1.000	0.027	-0.025
<b>MA2,1</b>	0.027	1.000	-0.021
<b>MA3,1</b>	-0.025	-0.021	1.000

Tabla 11: Correlación entre los parámetros del modelo SARIMA(0, 1, 1)(0, 1, 1)<sub>12</sub>(0, 0, 1)<sub>17</sub>.

En la Tabla 12 se recogen distintos estadísticos acerca de la calidad de ajuste del modelo, útiles para realizar comparaciones entre distintos modelos. Puesto que el objetivo de esta sección es el ajuste y validación, se comentarán dichos resultados posteriormente.

<b>Variance Estimate</b>	9.047408
<b>Std Error Estimate</b>	3.007891
<b>AIC</b>	851.1135
<b>SBC</b>	860.4674
<b>Number of Residuals</b>	167

Tabla 12: Estadísticos de ajuste del modelo SARIMA(0, 1, 1)(0, 1, 1)<sub>12</sub>(0, 0, 1)<sub>17</sub>.

En la Figura 7 (generada a partir del Código Fuente 8) se resume la serie generada por los residuales del ajuste del modelo. A partir del gráfico de la serie se aprecia el comportamiento aleatorio de los residuales. Sin embargo, sigue presente el aumento de la varianza en torno al comienzo del primero cuarto de la serie (correspondiente al aumento de nivel durante finales del año 2008). Tanto en el correlograma, como en el correlograma parcial se aprecia un leve estrucutra de ondas, pero que se cree que puede considererarse no significativa por el reducido valor de las caorrelaciones (en todos los casos menor que 0,15 en valor absoluto). En cuanto al periodograma, no se aprecia una estructura bien definida con armónicos que destaque sobre el resto. Finalmente, el diagrama de dispersión *rango-media* indica que no hay relación entre el nivel y la dispersión.

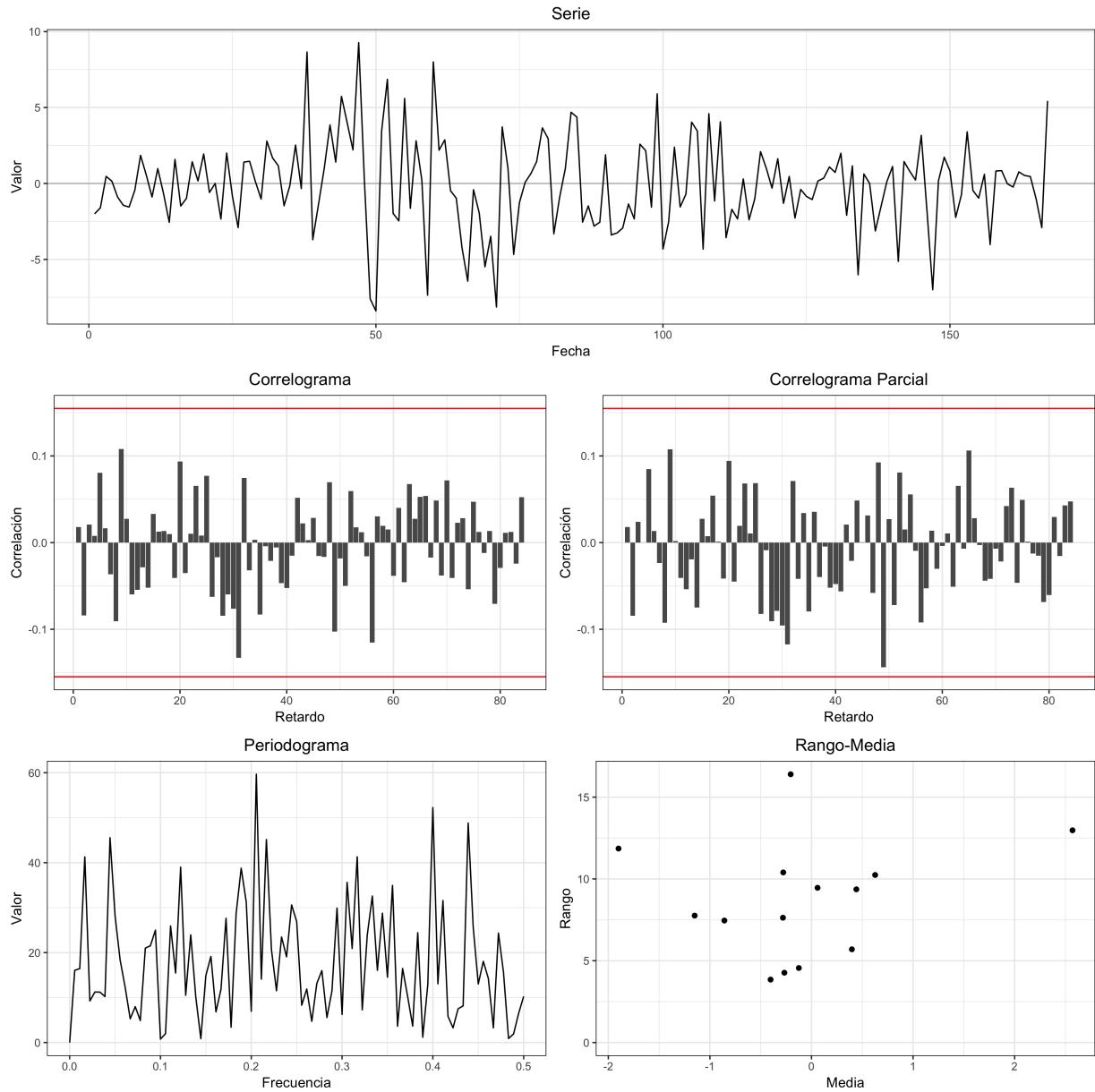


Figura 7: Gráfico de la serie, Correlograma, Correlograma Parcial, Periodograma y diagrama de dispersión *rango-media* para la serie de residuales ajustado por el modelo SARIMA(0, 1, 1)(0, 1, 1)<sub>12</sub>(0, 0, 1)<sub>17</sub>.

Por lo tanto, se cree que la serie de los residuales podría ser considerada un ruido blanco a partir de los gráficos de la Figura 7. Sin embargo, para confirmar dicha hipótesis se ha decidido realizar distintos contrastes.

```
Box-Ljung test
X-squared = 0.055242, lag = 1, p-value = 0.8142
```

Tabla 13: Resultados del test de *Ljung-Box* de dependencia serial en los residuales ajustados por el modelo SARIMA(0, 1, 1)(0, 1, 1)<sub>12</sub>(0, 0, 1)<sub>17</sub>

En la Tabla 13 y en la Tabla 14 se incluyen los tests de *Ljung-Box* acerca de la independencia serial de la serie de residuales, respecto del primer y duodécimo retardos respectivamente. En ambos casos aceptamos la hipótesis de independencia por no tener indicios suficientes para el rechazo (*pvalor* > 0,8 en ambos casos). Las sentencias necesarias para la reproducción de estos tests se incluyen en el Código Fuente 8

```
Box-Ljung test
X-squared = 7.6285, lag = 12, p-value = 0.8134
```

Tabla 14: Resultados del test de *Ljung-Box* de dependencia estacional en los residuales ajustados por el modelo SARIMA(0, 1, 1)(0, 1, 1)<sub>12</sub>(0, 0, 1)<sub>17</sub>

Otra de las hipótesis que deben cumplir los residuales de un modelo ajustado es la de normalidad. En la Figura 8 se muestra un gráfico *cuantil-cuantil* de normalidad. A partir de este, podemos apreciar un muy buen ajuste en la parte central de la distribución. Sin embargo, en los extremos se puede ver que el ajuste no es tan adecuado (aunque si que es simétrico respecto de cada lado). Se puede comprobar que dichas desviaciones se deben a que la distribución de los residuales es más apuntada que la normal (lo cual puede ser considerado un hecho positivo).

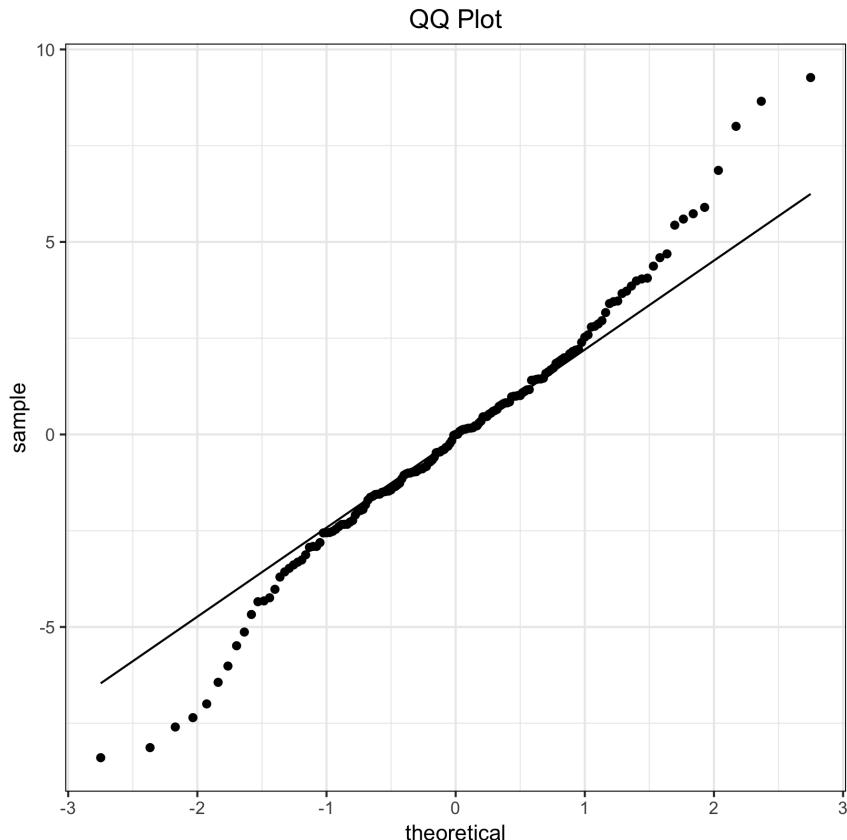


Figura 8: Gráfico de normalidad *cuantil-cuantil* para la serie de residuales ajustado por el modelo SARIMA(0, 1, 1)(0, 1, 1)<sub>12</sub>(0, 0, 1)<sub>17</sub>.

Al igual que en la anterior sección, como alternativa al gráfico *cuantil-cuantil* de la Figura 8, en la Tabla 15 y en la Tabla 16 referidas a los resultados de los tests de *Lilliefors* y *Shapiro-Francia* respectivamente. Las sentencias necesarias para la reproducción de estos tests se incluyen en el Código Fuente 8. Para el caso de *Lilliefors* el pvalor es 0,2067, por lo que no rechazamos la hipótesis de normalidad (por poco). Sin embargo, para el caso de *Shapiro-Francia*, nos vemos obligados a rechazar dicha hipótesis. Este último es un test mucho menos conservador que el primero, por tanto, los resultados son coherentes.

La razón por la cual estos pvalores son tan bajos, se cree que se debe al excesivo apuntamiento de la serie de residuales que provoca la falta de normalidad. Sin embargo, el ajuste se considera adecuado, por la independencia entre observaciones, así como la simetría de su distribución.

Puesto que el modelo cumple todas las características para ser considerado un modelo válido, en la Sección 3 será comparado con el resto de modelos válidos en la fase de selección de modelos. A pesar de

Lilliefors (Kolmogorov-Smirnov) normality test  
 $D = 0.056905$ , p-value = 0.2067

Tabla 15: Resultados del test de *Lilliefors* de normalidad en los residuales ajustados por el modelo SARIMA(0, 1, 1)(0, 1, 1)<sub>12</sub>(0, 0, 1)<sub>17</sub>

Shapiro-Francia normality test  
 $W = 0.98119$ , p-value = 0.02422

Tabla 16: Resultados del test de *Shapiro-Francia* de normalidad en los residuales ajustados por el modelo SARIMA(0, 1, 1)(0, 1, 1)<sub>12</sub>(0, 0, 1)<sub>17</sub>

no haberse podido confirmar la hipótesis de normalidad en los residuales (por el excesivo apuntamiento de la distribución), se considera un modelo válido.

### 2.3. Estimación y Validación de SARIMA(1, 1, 1)(0, 1, 1)<sub>12</sub>

Para el ajuste de este modelo en la etapa de validación, se ha utilizado el Código Fuente 9. Los parámetros ajustados se recogen en la Tabla 17. Estos son significativos para el caso del parámetro de media móvil en el primer retardo (regular), para el caso de duodécimo retardo (estacional) y para el parámetro referido al décimo séptimo retardo. Los valores estimados son 0,46545 y 0,62276 respectivamente, con errores de 0,20516 y 0,06616.

Maximum Likelihood Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr >  t	Lag
<b>MA1,1</b>	0.46545	0.20516	2.27	0.0233	1
<b>MA2,1</b>	0.62276	0.06616	9.41	<.0001	12
<b>AR1,1</b>	0.15694	0.22980	0.68	0.4947	1

Tabla 17: Estimación de los parámetros por el método de *Máxima Verosimilitud* para el modelo SARIMA(1, 1, 1)(0, 1, 1)<sub>12</sub> sobre la serie *weightloss*.

En cuanto al parámetro autoregresivo, este no es significativamente distinto de cero (con un pvalor de 0,4947). Por esta razón, no podemos validar el modelo, ya que siempre podremos encontrar un modelo más sencillo que nos permita obtener los mismos resultados. Además, tal y como se puede ver en la Tabla 18, el parámetro autoregresivo presenta una elevada correlación con el parámetro de media móvil en su mismo retardo, lo cual no es una buena señal.

Correlations of Parameter Estimates			
Parameter	MA1,1	MA2,1	AR1,1
<b>MA1,1</b>	1.000	-0.058	0.943
<b>MA2,1</b>	-0.058	1.000	-0.072
<b>AR1,1</b>	0.943	-0.072	1.000

Tabla 18: Correlación entre los parámetros del modelo SARIMA(1, 1, 1)(0, 1, 1)<sub>12</sub>.

Por las razones descritas, nos vemos obligados a invalidar el modelo, por lo que no será utilizado en la fase de comparación de la Sección 3.

## 2.4. Estimación y Validación de SARIMA(1, 1, 1)(0, 1, 1)<sub>12</sub>(0, 0, 1)<sub>17</sub>

Para el ajuste de este modelo en la etapa de validación, se ha utilizado el Código Fuente 10. Los parámetros ajustados se recogen en la Tabla 19. Estos son significativos para el caso del parámetro de media móvil en el primer retardo (regular), para el caso de duodécimo retardo (estacional) y para el parámetro referido al décimo séptimo retardo. Los valores estimados son 0,46122, 0,62117 y -0,15758 respectivamente, con errores de 0,20376, 0,06629 y 0,07845.

Maximum Likelihood Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr >  t	Lag
<b>MA1,1</b>	0.46122	0.20376	2.26	0.0236	1
<b>MA2,1</b>	0.62117	0.06629	9.37	<.0001	12
<b>MA3,1</b>	-0.15758	0.07845	-2.01	0.0446	17
<b>AR1,1</b>	0.14754	0.22799	0.65	0.5175	1

Tabla 19: Estimación de los parámetros por el método de *Máxima Verosimilitud* para el modelo SARIMA(1, 1, 1)(0, 1, 1)<sub>12</sub>(0, 0, 1)<sub>17</sub> sobre la serie *weightloss*.

En cuanto al parámetro autoregresivo, este no es significativamente distinto de cero (con un pvalor de 0,51175). Por esta razón, no podemos validar el modelo, ya que siempre podremos encontrar un modelo más sencillo que nos permita obtener los mismos resultados. Además, tal y como se puede ver en la Tabla 20, el parámetro autoregresivo presenta una elevada correlación con el parámetro de media móvil en su mismo retardo, lo cual no es una buena señal.

Correlations of Parameter Estimates				
Parameter	MA1,1	MA2,1	MA3,1	AR1,1
<b>MA1,1</b>	1.000	-0.067	0.017	0.941
<b>MA2,1</b>	-0.067	1.000	-0.024	-0.078
<b>MA3,1</b>	0.017	-0.024	1.000	0.025
<b>AR1,1</b>	0.941	-0.078	0.025	1.000

Tabla 20: Correlación entre los parámetros del modelo SARIMA(1, 1, 1)(0, 1, 1)<sub>12</sub>(0, 0, 1)<sub>17</sub>.

Por las razones descritas, nos vemos obligados a invalidar el modelo, por lo que no será utilizado en la fase de comparación de la Sección 3.

## 3. Comparación de modelos

Tras la fase de validación de los modelos, la cual superaron dos de los cuatro propuestos, en esta sección se procede a la comparación de estos para tratar de seleccionar aquel que represente en mejor el comportamiento de los datos observados de la serie. Los dos modelos que serán comparados son los siguientes:

- SARIMA(0, 1, 1)(0, 1, 1)<sub>12</sub>
- SARIMA(0, 1, 1)(0, 1, 1)<sub>12</sub>(0, 0, 1)<sub>17</sub>

Para la comparación de los modelos existen distintas alternativas. En este caso, se compararán utilizando las tablas de estadísticos de ajustes del modelo, generadas en la fase de ajuste y validación (Sección 2). Además, se analizará la calidad de predicción de los modelos, utilizando la *suma de cuadrados del error de predicción*. Posteriormente, se estudiará la amplitud de los intervalos de predicción de los modelos.

Tal y como se indicó anteriormente, en la Tabla 5 y en la Tabla 12 se muestran distintos estadísticos de resumen sobre la calidad del ajuste de la serie. En la Tabla 21 se resumen los más característicos. Tal y como podemos apreciar, las varianzas residuales estimadas en ambos modelos son muy similares y, por ser modelos anidados, la varianza del error de SARIMA(0, 1, 1)(0, 1, 1)<sub>12</sub>(0, 0, 1)<sub>17</sub> es menor. En cuanto al criterio de *Akaike*, este también es menor en dicho modelo. Sin embargo, al igual que para el caso de las varianzas, estas diferencias son muy reducidas (se recomienda no tener en cuenta el criterio de *Akaike* cuando las diferencias son inferiores a 3.). Por último, el criterio Bayesiano de *Schwarz* otorga ventaja al modelo SARIMA(0, 1, 1)(0, 1, 1)<sub>12</sub>, aunque nuevamente de manera muy poco significativa. La razón de que el criterio Bayesiano sea opuesta al de *Akaike* es la mayor penalización en el número de parámetros que se lleva a cabo en este caso.

Modelo	$\hat{\sigma}^2$	AIC	BIC
SARIMA(0, 1, 1)(0, 1, 1) <sub>12</sub>	9,25	853,5882	859,7242
SARIMA(0, 1, 1)(0, 1, 1) <sub>12</sub> (0, 0, 1) <sub>17</sub>	9,04	851,1135	860,4674

Tabla 21: Estadísticos de ajuste de los modelos validados

Puesto que los estadísticos de ajuste no han aportado argumentos en favor o en contra para ninguno de los modelos, se va a proceder a realizar la comparación desde el punto de vista de las predicciones. Para ello, tal y como se ha indicado anteriormente, se utilizará la *suma de cuadrados del error de predicción*, junto con la amplitud de los intervalos de predicción.

Para calcular el  $SSE_p$ , primero explicaremos resumidamente qué es así como en qué está basado. El  $SSE_p$  es se corresponde con el cálculo de una medida para comparar la capacidad de predicción de un modelo de serie temporal. Se trata de predecir observaciones de una serie estacional de periodo  $s$ , para medir la capacidad de predicción de un modelo ajustado a dicha serie.

Si se dispone de  $n$  observaciones en total,  $x_1, x_2, \dots, x_n$  se reservan las últimas  $k$  observaciones, donde  $k$  es un múltiplo de  $s$ . Para el ajuste se utilizan  $m$  observaciones ( $m = n - k$ ) y la medida se obtiene sumando los cuadrados de los residuales  $\{1, 2, \dots, k\}$  pasos hacia adelante. Esto se define en la Ecuación 9.

$$\begin{aligned}
 SSE_p &= \sum_{j=1}^k (x_{m+j} - x_m(j))^2 \\
 &= \sum_{j=1}^k e_m(j)^2 \\
 &= e_m(1)^2 + e_m(2)^2 + \dots + e_m(k)^2
 \end{aligned} \tag{9}$$

Para el cálculo de la suma de errores de predicción primero se han realizado primeramente las predicciones para ambos modelos, reservando las 12 últimas observaciones (que además son las 12 que se han predicho), lo cual se ha llevado a cabo a partir de el Código Fuente 11. A partir de dichas sentencias se han obtenido las predicciones, las cuales se muestran en la Tabla 22 y en la Tabla 22 para cada modelo respectivamente (se han omitido las primeras observaciones). Entonces, para calcular la *suma de cuadrados del error de predicción*, basta con llevar a cabo la suma de los cuadrados de las columnas **Residuo** en cada caso. Esto se ha llevado a cabo utilizando el Código Fuente 12.

En la Tabla 24 se muestran los errores de predicción de cada modelo. Como vemos, el del modelo SARIMA(0, 1, 1)(0, 1, 1)<sub>12</sub>(0, 0, 1)<sub>17</sub> es menor, por lo tanto intuimos que las predicciones de este tenderán a ser mejores.

Otro de los puntos a tener en cuenta es la amplitud de los intervalos de predicción obtenidos en cada modelo, dado que en un gran número de ocasiones es más interesante conocer en torno a qué valores se encontrará una nueva observación que el valor exacto que tendrá la misma. Es por ello que en la Figura 9 se muestra una comparativa entre los dos modelos de los intervalos de predicción para las 12 observaciones predichas. En esta representación se han centrado los datos respecto de los datos observados

<b>Fecha</b>	<b>Obs.</b>	<b>Predicho</b>	<b>Error Est.</b>	<b>L. Inf. (95 %)</b>	<b>L. Sup. (95 %)</b>	<b>Residuo</b>
01JAN18	90	89.372687134	3.0419856496	83.410504819	95.334869448	0.6273128662
01FEB18	80	83.420542837	3.6588976238	76.249235272	90.591850403	-3.420542837
01MAR18	81	83.128602332	4.1858555816	74.924476148	91.332728517	-2.128602332
01APR18	82	83.065259999	4.6535193218	73.944529727	92.185990271	-1.065259999
01MAY18	80	81.319891171	5.0782966837	71.366612568	91.273169774	-1.319891171
01JUN18	77	78.557906823	5.4701875961	67.836536146	89.2792775	-1.557906823
01JUL18	80	80.223581053	5.8358210618	68.785581952	91.661580154	-0.223581053
01AUG18	76	76.419844596	6.1798594316	64.307542681	88.532146511	-0.419844596
01SEP18	70	71.150057444	6.5057296072	58.399061721	83.901053167	-1.150057444
01OCT18	67	68.902734377	6.8160379145	55.543545547	82.261923207	-1.902734377
01NOV18	63	67.209298059	7.1128213797	53.268424327	81.150171792	-4.209298059
01DEC18	62	59.758412249	7.3977079632	45.259171073	74.257653425	2.2415877512

Tabla 22: Predicción 12 observaciones hacia delante, reservando las 12 últimas para calcular el  $SSE_p$  de la serie `weightloss` utilizando el modelo SARIMA(0, 1, 1)(0, 1, 1)<sub>12</sub>.

<b>Fecha</b>	<b>Obs.</b>	<b>Predicho</b>	<b>Error Est.</b>	<b>L. Inf. (95 %)</b>	<b>L. Sup. (95 %)</b>	<b>Residuo</b>
01JAN18	90	89.398500026	3.0078909592	83.503142077	95.293857976	0.6014999735
01FEB18	80	83.622205013	3.6101038056	76.546531573	90.697878452	-3.622205013
01MAR18	81	82.447992938	4.1253231331	74.362508172	90.533477703	-1.447992938
01APR18	82	82.879699663	4.582982917	73.897218204	91.862181122	-0.879699663
01MAY18	80	81.185353534	4.998917271	71.387655721	90.983051347	-1.185353534
01JUN18	77	78.415309456	5.3828073854	67.865200844	88.965418067	-1.415309456
01JUL18	80	80.572456961	5.7410849857	69.320137157	91.824776766	-0.572456961
01AUG18	76	76.304267818	6.0782808653	64.391056234	88.217479402	-0.304267818
01SEP18	70	70.022420276	6.3977292646	57.483101335	82.561739218	-0.022420276
01OCT18	67	68.178616345	6.7019684577	55.042999542	81.314233148	-1.178616345
01NOV18	63	66.749557117	6.9929838176	53.04356069	80.455553544	-3.749557117
01DEC18	62	59.335448002	7.2723630368	45.081878368	73.589017637	2.6645519979

Tabla 23: Predicción 12 observaciones hacia delante, reservando las 12 últimas para calcular el  $SSE_p$  de la serie `weightloss` utilizando el modelo SARIMA(0, 1, 1)(0, 1, 1)<sub>12</sub>(0, 1, 1)<sub>17</sub>.

Modelo	$SSE_p$
SARIMA(0, 1, 1)(0, 1, 1) <sub>12</sub>	49.84
SARIMA(0, 1, 1)(0, 1, 1) <sub>12</sub> (0, 0, 1) <sub>17</sub>	42.73

Tabla 24: Suma de Cuadrados del Error de Predicción acumulada para 12 observaciones hacia delante para los modelos SARIMA(0, 1, 1)(0, 1, 1)<sub>12</sub> y SARIMA(0, 1, 1)(0, 1, 1)<sub>12</sub>(0, 0, 1)<sub>17</sub>.

para tratar de eliminar la forma de la serie en la misma y fijarse únicamente en los residuales. Tal y como se puede ver, la predicción del modelo SARIMA(0, 1, 1)(0, 1, 1)<sub>12</sub>(0, 0, 1)<sub>17</sub> se encuentra más cercana a los datos, lo cual sirve como confirmación para los resultados de la *suma de cuadrados del error de predicción* de la Tabla 24. Sin embargo, esta representación no proporciona demasiada información acerca de la amplitud de los intervalos, ya que parecen muy similares para ambos modelos.

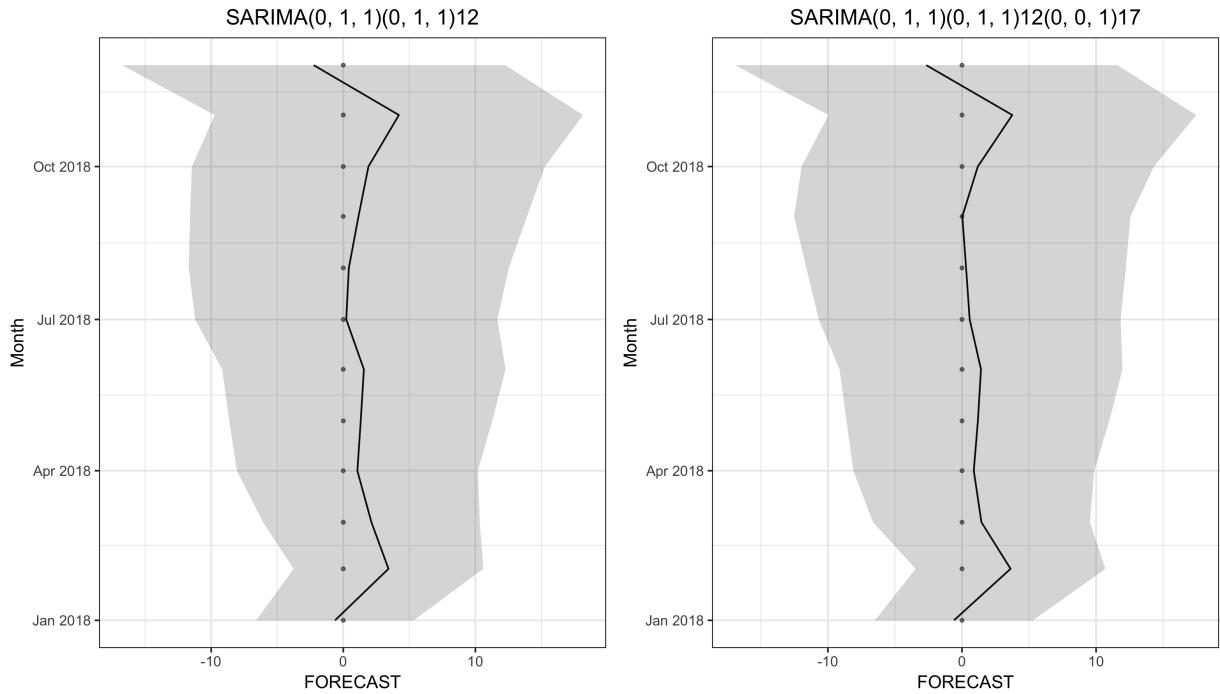


Figura 9: Comparación de las amplitudes de los intervalos de predicción para los modelos SARIMA(0, 1, 1)(0, 1, 1)<sub>12</sub> y SARIMA(0, 1, 1)(0, 1, 1)<sub>12</sub>(0, 0, 1)<sub>17</sub>.

Para contrastar la amplitud de los intervalos de predicción de una manera más rigurosa, se ha llevado a cabo el test *t de student* descrito en la Ecuación 10. Dicho test se basa en la comparación de la amplitud de ambos modelos, donde el modelo 1 será SARIMA(0, 1, 1)(0, 1, 1)<sub>12</sub> y el 2 será SARIMA(0, 1, 1)(0, 1, 1)<sub>12</sub>(0, 0, 1)<sub>17</sub>. Este test se llevará a cabo sobre las 12 últimas predicciones, ya que estas son referidas a observaciones desconocidas (en los modelos ajustados para la comparación). A pesar de no serlo, asumiremos que las observaciones son independientes entre si para poder llevar a cabo el test.

$$\begin{aligned} H_0 : U_1 - L_1 &= U_2 - L_2 \\ H_1 : U_1 - L_1 &\neq U_2 - L_2 \end{aligned} \tag{10}$$

El test de la Ecuación 10 se han llevado a cabo utilizando el Código Fuente 12, y el resultado del mismo se muestra en la Tabla 25. El pvalor indica que tenemos que rechazar la hipótesis de igualdad en la amplitud de los intervalos de predicción. En concreto, se indica que la amplitud del intervalo para el modelo SARIMA(0, 1, 1)(0, 1, 1)<sub>12</sub> es al menos 0,36 puntos mayor. Nótese que la hipótesis de independencia de las observaciones no se cumple totalmente, por lo que debemos tomar estos resultados con mucha cautela.

[TODO]

**One Sample t-test**

```
t = 10.377, df = 11, p-value = 5.101e-07
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.2687731 0.4134798
sample estimates:
mean of x
0.3411265
```

Tabla 25: Resultados del test *t de student* sobre la media nula ( $H_0 : \mu = 0$ ) para la diferencia entre amplitudes en los intervalos de predicción para los modelos SARIMA(0, 1, 1)(0, 1, 1)<sub>12</sub> y SARIMA(0, 1, 1)(0, 1, 1)<sub>12</sub>(0, 0, 1)<sub>17</sub>.

## 4. Predicción

Una vez realizada la comparación de modelos (donde finalmente se ha decidido que el modelo que generará resultados más cercanos a la realidad es el **SARIMA**(0, 1, 1)(0, 1, 1)<sub>12</sub>(0, 0, 1)<sub>17</sub> por su mayor grado de ajuste a los datos), en esta sección se va a proceder a realizar la predicción de la serie 2 años hacia delante. La expresión extendida de este modelo se incluye en la Ecuación 11.

$$[\text{TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO}] \quad (11)$$

$$[\text{TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO}] \quad (12)$$

Una vez descrito completamente el modelo elegido para realizar las predicciones, ya estamos en condiciones suficientes para llevar a cabo las mismas. Para ello, es necesario escribir la expresión explícita de la predicción con el modelo escogido, la cual se muestra en la ecuación Ecuación 13.

$$[\text{TODO TODO TODO TODO TODO TODO TODO TODO TODO TODO}] \quad (13)$$

Una vez descrita la expresión de predicción, basta con aplicarla de manera secuencial utilizando como base la última observación de los datos, esto es  $n = 15 \cdot 12 = 180$ . Tras la aplicación secuencial de la ecuación, se consiguen las predicciones que se muestran en la Tabla 26, en la columna **Predicho**. En dicha tabla se han incluido también los datos observados referidos al año 2018 para poder apreciar la predicción de la serie. También se incluye las cotas inferior y superior con una confianza del 95 %, las cuales forman los intervalos de predicción del 90 %. Dicha predicción se ha llevado a cabo utilizando el Código Fuente 13.

Para facilitar la compresión de los resultados, en la Figura 10 se muestra con una línea la predicción del modelo para toda la serie, indicando en las predicciones futuras los intervalos de predicción del 90 %. Además, se han incluido los valores observados en forma de puntos. La Figura 11 se corresponde con una visión ampliada de los años 2018, 2019 y 2020 de las predicciones. Estas imágenes han sido generadas mediante el Código Fuente 14. Tal y como se puede apreciar, las predicciones comienzan a partir de la observación 14. Esto se debe a las diferenciaciones regular y estacional, que en total conllevan la pérdida de 13 observaciones.

En la Figura 11 se puede apreciar como las predicciones del modelo escogido siguen un comportamiento muy próximo a las que sigue la serie durante los años observados, lo cual es el comportamiento esperado. En este caso, es algo natural dado que la serie presenta una estacionalidad de periodo anual claramente marcada, tal y como se indicó en la fase descriptiva. Otra de las características que llaman la atención de las predicciones del modelo es la progresiva reducción del nivel. Esta se puede comprobar de manera más adecuada en la Figura 10, que muestra las predicciones para la serie completa. Una interpretación para este fenómeno podría ser que (si el modelo realmente se ajusta a la realidad) la frecuencia de búsquedas de la palabra clave **Weight Loss** decrecerá, lo cual podría ser un reflejo del indicador de la importancia que la población da a la preocupación por su estado de forma física. Por último, se puede comprobar que los intervalos de predicción de la serie aumentan conforme se alejan de los últimos datos observados de la serie, lo cual es algo coherente dado que la varianza de las predicciones aumenta conforme, estas se alejan de los datos observados.

<b>Fecha</b>	<b>Obs.</b>	<b>Predicho</b>	<b>Error Est.</b>	<b>L. Inf. (95%)</b>	<b>L. Sup. (95%)</b>	<b>Residuo</b>
01JAN18	90	89.398500026	3.0078909592	83.503142077	95.293857976	0.6014999735
01FEB18	80	84.021424508	3.0078909592	78.126066559	89.916782458	-4.021424508
01MAR18	81	80.17816649	3.0078909592	74.282808541	86.07352444	0.8218335099
01APR18	82	81.155329531	3.0078909592	75.259971582	87.050687481	0.8446704688
01MAY18	80	80.021596758	3.0078909592	74.126238808	85.916954707	-0.021596758
01JUN18	77	77.237218769	3.0078909592	71.341860819	83.132576718	-0.237218769
01JUL18	80	79.236922615	3.0078909592	73.341564665	85.132280564	0.7630773852
01AUG18	76	75.475192959	3.0078909592	69.57983501	81.370550909	0.5248070406
01SEP18	70	69.541663309	3.0078909592	63.646305359	75.437021258	0.4583366913
01OCT18	67	68.00206046	3.0078909592	62.10670251	73.897418409	-1.00206046
01NOV18	63	65.907927102	3.0078909592	60.012569153	71.803285052	-2.907927102
01DEC18	62	56.563807609	3.0078909592	50.66844966	62.459165558	5.436192391
01JAN19		87.137067037	3.0078909592	81.241709088	93.032424986	
01FEB19		79.655504231	3.6101038056	72.579830791	86.73117767	
01MAR19		80.427359797	4.1253231331	72.341875031	88.512844562	
01APR19		80.506969453	4.582982917	71.524487994	89.489450912	
01MAY19		78.537346934	4.998917271	68.739649121	88.335044747	
01JUN19		75.831461925	5.3828073854	65.281353314	86.381570536	
01JUL19		77.332910288	5.7410849857	66.080590484	88.585230092	
01AUG19		73.79647995	6.0782808653	61.883268366	85.709691534	
01SEP19		68.341015383	6.3977292646	55.801696442	80.880334325	
01OCT19		65.761030145	6.7019684577	52.625413342	78.896646947	
01NOV19		63.156892018	6.9929838176	49.450895591	76.862888445	
01DEC19		58.283328556	7.2723630368	44.029758921	72.53689819	
01JAN20		84.845896767	7.9197687858	69.323435181	100.36835835	
01FEB20		77.407877612	8.3845641119	60.974433927	93.841321298	
01MAR20		77.636521475	8.8249132047	60.340009427	94.933033523	
01APR20		77.466460277	9.244310185	59.347945252	95.584975301	
01MAY20		76.594845064	9.6454885061	57.690034979	95.499655149	
01JUN20		73.507441435	10.172037866	53.570613568	93.444269302	
01JUL20		75.425285296	10.625143695	54.600386324	96.250184268	
01AUG20		71.674509467	11.059701746	49.997892366	93.351126569	
01SEP20		66.163019501	11.477818908	43.66690782	88.659131181	
01OCT20		63.61316726	11.881231042	40.326382327	86.899952194	
01NOV20		61.031699423	12.271388481	36.98021996	85.083178885	
01DEC20		56.075043039	12.649517754	31.282443819	80.867642259	

Tabla 26: Predicción 24 observaciones hacia delante de la serie `weightloss` utilizando el modelo SARIMA(0, 1, 1)(0, 1, 1)<sub>12</sub>(0, 0, 1)<sub>17</sub>. (Últimas observaciones de la serie).

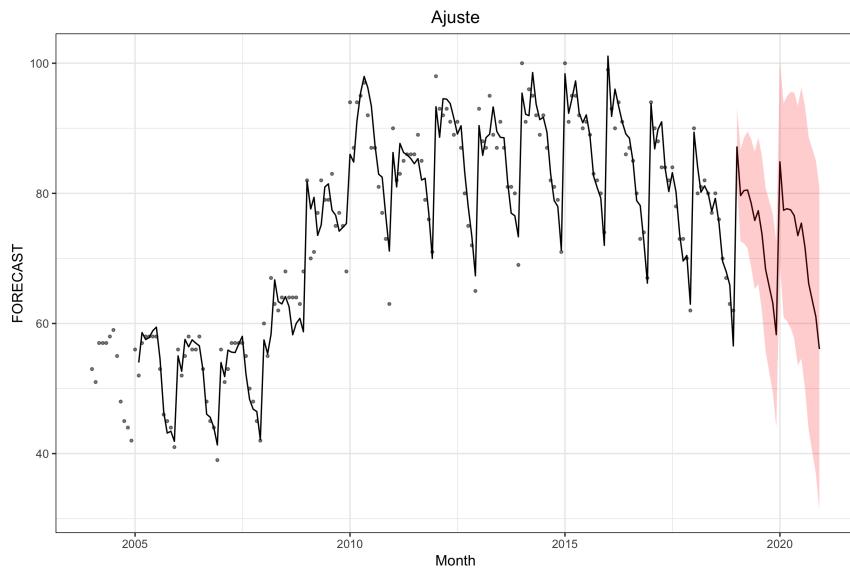


Figura 10: Predicción 24 observaciones hacia delante de la serie `weightloss` utilizando el modelo SARIMA(0, 1, 1)(0, 1, 1)<sub>12</sub>(0, 0, 1)<sub>17</sub>. (Serie completa)

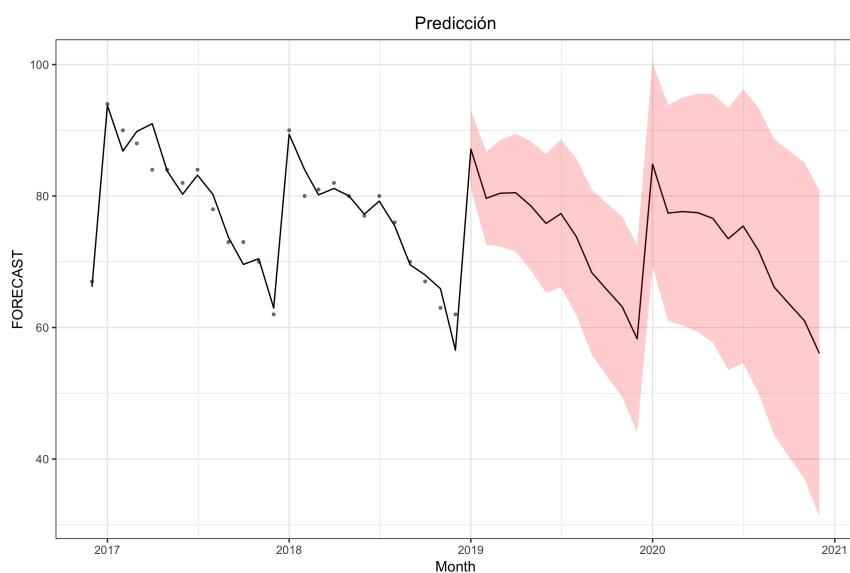


Figura 11: Predicción 24 observaciones hacia delante de la serie `weightloss` utilizando el modelo SARIMA(0, 1, 1)(0, 1, 1)<sub>12</sub>(0, 0, 1)<sub>17</sub>. (Últimas observaciones de la serie)

## A. Código Fuente

```

RangeMean <- function(x, seasonality) {
  n <- length(x)
  seq(1, n, by=seasonality) %>%
    sapply(function(i){
      a <- x[i:(i + seasonality - 1)]
      c(mean=mean(a, na.rm=TRUE), range=diff(range(a, na.rm = TRUE)))
    }) %>%
    t() %>%
    as.data.frame()
}

Correlogram <- function(x, n = length(x) - 1) {
  result <- acf(x, lag.max=n, plot=FALSE)$acf[1:n + 1]
  data.frame(lag = 1:length(result), values = result)
}

PartialCorrelogram <- function(x, n = length(x) - 1) {
  result <- pacf(x, lag.max=n, plot=FALSE)$acf
  data.frame(lag = 1:length(result), values = result)
}

Periodogram <- function(x) {
  result <- TSA:::periodogram(x, plot=FALSE)
  data.frame(freq = c(0, result$freq), spec = c(0, result$spec))
}

PredictionError <- function(df, lags) {
  sum(df[(nrow(df) - lags + 1):nrow(df), 'RESIDUAL'] ^ 2)
}

```

Código Fuente 1: Conjunto de funciones de apoyo necesarias para el análisis descriptivo de series temporales utilizando el lenguaje R.

```

source("res/code/functions.r")

library(ggplot2)

PlotTimeSeries <- function(df, seasonality, armonics = c(), lags = MAX_LAG){
  p.a <- ggplot(df) +
    aes(x = index, y = values) +
    xlab("Fecha") +
    ylab("Valor") +
    geom_hline(yintercept = 0, color = "gray") +
    geom_line() +
    theme_bw() +
    theme(plot.title = element_text(hjust = 0.5),
          panel.border = element_rect(colour = "black", fill=NA)) +
    ggttitle('Serie')

  p.b <- ggplot(RangeMean(df$values, seasonality)) +
    aes(x = mean, y = range) +
    geom_point() +
    xlab("Media") +
    ylab("Rango") +
    expand_limits(y=0) +
    theme_bw() +
    theme(plot.title = element_text(hjust = 0.5),
          panel.border = element_rect(colour = "black", fill=NA)) +
    ggttitle('Rango-Media')

  p.c <- ggplot(Correlogram(df$values, lags)) +
    aes(x = lag, y = values) +
    xlab("Retardo") +
    ylab("Correlación") +
    geom_bar(stat="identity") +
    geom_hline(yintercept = 2/sqrt(nrow(df)), color = "red") +
    geom_hline(yintercept = -2/sqrt(nrow(df)), color = "red") +
    theme_bw() +
    theme(plot.title = element_text(hjust = 0.5),
          panel.border = element_rect(colour = "black", fill=NA)) +
    ggttitle('Correlograma')

  p.partial.correlogram <- ggplot(PartialCorrelogram(df$values, lags)) +
    aes(x = lag, y = values) +
    xlab("Retardo") +
    ylab("Correlación") +
    geom_bar(stat="identity") +
    geom_hline(yintercept = 2/sqrt(nrow(df)), color = "red") +
    geom_hline(yintercept = -2/sqrt(nrow(df)), color = "red") +
    theme_bw() +
    theme(plot.title = element_text(hjust = 0.5),
          panel.border = element_rect(colour = "black", fill=NA)) +
    ggttitle('Correlograma Parcial')

  p.d <- ggplot(Periodogram(df$values)) +
    aes(x = freq, y = spec) +
    xlab("Frecuencia") +
    ylab("Valor") +
    geom_line() +
    theme_bw() +
    theme(plot.title = element_text(hjust = 0.5),
          panel.border = element_rect(colour = "black", fill=NA)) +
    ggttitle('Periodograma')

  for (a in armonics) {
    p.d <- p.d + geom_vline(xintercept = a, color = "red", alpha = 0.4)
  }
  plot_grid(p.a, plot_grid(p.c, p.partial.correlogram, p.d, p.b, ncol = 2),
            ncol = 1, rel_heights = c(1, 2))
}

}

```

Código Fuente 2: Función de generación de representación gráfica de una serie utilizando el lenguaje R.

```

rm(list = ls())

library(magrittr)
library(dplyr)
library(latex2exp)
require(reshape2)
library(forecast)
library(cowplot)
library(lubridate)

source("res/code/functions.r")
source("res/code/plotting.r")

BASE_PATH <- './'
BASE_IMG_PATH <- paste0(BASE_PATH, 'res/img/')
BASE_DATA_PATH <- paste0(BASE_PATH, 'res/data/')

weightloss <- read.csv(paste0(BASE_DATA_PATH, 'weight-loss.csv'))
weightloss$Month <- ymd(weightloss$Month, truncated = 2)
colnames(weightloss) <- c("index", "values")

PlotTimeSeries(weightloss, seasonality = 12, lags = 84) %>%
  { save_plot(paste0(BASE_IMG_PATH, 'weightloss.png')), ,
    base_aspect_ratio = 1, base_height = 12 }

values <- diff(weightloss$values, 1)
df <- data.frame(index = 1:length(values), values=values)
PlotTimeSeries(df, seasonality = 12, lags = 84) %>%
  { save_plot(paste0(BASE_IMG_PATH, 'weightloss-diff-1.png')), ,
    base_aspect_ratio = 1, base_height = 12 }

values <- diff(weightloss$values, 12)
df <- data.frame(index = 1:length(values), values=values)
PlotTimeSeries(df, seasonality = 12, lags = 84) %>%
  { save_plot(paste0(BASE_IMG_PATH, 'weightloss-diff-12.png')), ,
    base_aspect_ratio = 1, base_height = 12 }

values <- diff(weightloss$values, 12, 2)
df <- data.frame(index = 1:length(values), values=values)
PlotTimeSeries(df, seasonality = 12, lags = 84) %>%
  { save_plot(paste0(BASE_IMG_PATH, 'weightloss-diff-12-12.png')), ,
    base_aspect_ratio = 1, base_height = 12 }

values <- diff(diff(weightloss$values, 1), 12)
df <- data.frame(index = 1:length(values), values=values)
PlotTimeSeries(df, seasonality = 12, lags = 84) %>%
  { save_plot(paste0(BASE_IMG_PATH, 'weightloss-diff-1-12.png')), ,
    base_aspect_ratio = 1, base_height = 12 }

t.test(values, mu=0)

values <- diff(diff(weightloss$values, 1), 12, 2)
df <- data.frame(index = 1:length(values), values=values)
PlotTimeSeries(df, seasonality = 12, lags = 84) %>%
  { save_plot(paste0(BASE_IMG_PATH, 'weightloss-diff-1-12-12.png')), ,
    base_aspect_ratio = 1, base_height = 12 }

values <- diff(diff(weightloss$values, 1, 2), 12)
df <- data.frame(index = 1:length(values), values=values)
PlotTimeSeries(df, seasonality = 12, lags = 84) %>%
  { save_plot(paste0(BASE_IMG_PATH, 'weightloss-diff-1-1-12.png')), ,
    base_aspect_ratio = 1, base_height = 12 }

```

Código Fuente 3: Análisis descriptivo utilizando el lenguaje R de la serie weightloss.

```
FILENAME REFFILE '/folders/myshortcuts/sarima-weight-loss/res/data/weight-loss.csv';
PROC IMPORT DATAFILE=REFFILE
DBMS=CSV
OUT=WORK.weightloss;
GETNAMES=YES;
RUN;
```

Código Fuente 4: Importación de datos utilizando el lenguaje SAS del conjunto de datos para la serie weightloss.

```
proc arima data = weightloss;
  identify var=weight_loss(1, 12);
  estimate q=(1)(12) noint method=ml;
  forecast id=month interval=month lead=0 out=arimaout_1;
run;

proc export data=arimaout_1
  outfile='/folders/myshortcuts/sarima-weight-loss/res/data/validation-1.csv'
  dbms=csv replace;
run;

proc univariate data = arimaout_1;
  var residual;
run;
```

Código Fuente 5: Validación utilizando el lenguaje SAS del modelo SARIMA(0, 1, 1)(0, 1, 1)<sub>12</sub> sobre la serie weightloss.

```

rm(list = ls())

library(ggplot2)
library(magrittr)
library(cowplot)
library(nortest)

source("res/code/functions.r")
source("res/code/plotting.r")

BASE_PATH <- './'
BASE_IMG_PATH <- paste0(BASE_PATH, 'res/img/')
BASE_DATA_PATH <- paste0(BASE_PATH, 'res/data/')

validation.1 <- read.csv(paste0(BASE_DATA_PATH, 'validation-1.csv'))

values <- validation.1$RESIDUAL[!is.na(validation.1$RESIDUAL)]
df <- data.frame(index = 1:length(values), values=values)

PlotTimeSeries(df, seasonality = 12, lags = 84) %>%
  { save_plot(paste0(BASE_IMG_PATH, 'validation-1-residuals.png'),.,
    base_aspect_ratio = 1, base_height = 12) }

Box.test(values, lag = 1, type="Ljung-Box")

Box.test(values, lag = 12, type="Ljung-Box")

sf.test(values)

lillie.test(values)

{ggplot(df, aes(sample = values)) +
  geom_qq() +
  geom_qq_line() +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5),
        panel.border = element_rect(colour = "black", fill=NA)) +
  ggtitle('QQ Plot')} %>%
  { save_plot(paste0(BASE_IMG_PATH, 'validation-1-normality.png'),.,
    base_aspect_ratio = 1, base_height = 6) }

```

Código Fuente 6: Validación utilizando el lenguaje R del modelo SARIMA(0,1,1)(0,1,1)<sub>12</sub> sobre la serie weightloss.

```

proc arima data = weightloss;
  identify var=weight_loss(1, 12);
  estimate q=(1)(12)(17) noint method=ml;
  forecast id=month interval=month lead=0 out=arimaout_2;
run;

proc export data=arimaout_2
  outfile='/folders/myshortcuts/sarima-weight-loss/res/data/validation-2.csv'
  dbms=csv replace;
run;

```

Código Fuente 7: Validación utilizando el lenguaje SAS del modelo SARIMA(0,1,1)(0,1,1)<sub>12</sub>(0,0,1)<sub>17</sub> sobre la serie weightloss.

```

rm(list = ls())

library(ggplot2)
library(magrittr)
library(cowplot)
library(nortest)

source("res/code/functions.r")
source("res/code/plotting.r")

BASE_PATH <- './'
BASE_IMG_PATH <- paste0(BASE_PATH, 'res/img/')
BASE_DATA_PATH <- paste0(BASE_PATH, 'res/data/')

validation.2 <- read.csv(paste0(BASE_DATA_PATH, 'validation-2.csv'))

values <- validation.2$RESIDUAL[!is.na(validation.2$RESIDUAL)]
df <- data.frame(index = 1:length(values), values=values)

PlotTimeSeries(df, seasonality = 12, lags = 84) %>%
  { save_plot(paste0(BASE_IMG_PATH, 'validation-2-residuals.png'),.,
    base_aspect_ratio = 1, base_height = 12) }

Box.test(values, lag = 1, type="Ljung-Box")

Box.test(values, lag = 12, type="Ljung-Box")

lillie.test(values)

sf.test(values)

{ggplot(df, aes(sample = values)) +
  geom_qq() +
  geom_qq_line() +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5),
        panel.border = element_rect(colour = "black", fill=NA)) +
  ggtitle('QQ Plot')} %>%
  { save_plot(paste0(BASE_IMG_PATH, 'validation-2-normality.png'),.,
    base_aspect_ratio = 1, base_height = 6) }

```

Código Fuente 8: Validación utilizando el lenguaje R del modelo SARIMA(0, 1, 1)(0, 1, 1)<sub>12</sub>(0, 0, 1)<sub>17</sub> sobre la serie weightloss.

```

proc arima data = weightloss;
  identify var=weight_loss(1, 12);
  estimate p=(1) q=(1)(12) noint method=ml;
  forecast id=month interval=month lead=0 out=arimaout_3;
run;

proc export data=arimaout_3
  outfile='/folders/myshortcuts/sarima-weight-loss/res/data/validation-3.csv'
  dbms=csv replace;
run;

```

Código Fuente 9: Validación utilizando el lenguaje SAS del modelo SARIMA(1, 1, 1)(0, 1, 1)<sub>12</sub> sobre la serie weightloss.

```

proc arima data = weightloss;
  identify var=weight_loss(1, 12);
  estimate p=(1) q=(1)(12)(17) noint method=ml;
  forecast id=month interval=month lead=0 out=arimaout_4;
run;

proc export data=arimaout_4
  outfile='/folders/myshortcuts/sarima-weight-loss/res/data/validation-4.csv'
  dbms=csv replace;
run;

```

Código Fuente 10: Validación utilizando el lenguaje SAS del modelo SARIMA $(1, 1, 1)(0, 1, 1)_{12}(0, 0, 1)_{17}$  sobre la serie **weightloss**.

```

proc arima data = weightloss;
  identify var=weight_loss(1, 12);
  estimate q=(1)(12) noint method=ml;
  forecast id=month interval=month lead=12 out=comparison_1 back = 12;
run;

proc export data=comparison_1
  outfile='/folders/myshortcuts/sarima-weight-loss/res/data/comparison-1.csv'
  dbms=csv replace;
run;

proc arima data = weightloss;
  identify var=weight_loss(1, 12);
  estimate q=(1)(12)(17) noint method=ml;
  forecast id=month interval=month lead=12 out=comparison_2 back = 12;
run;

proc export data=comparison_2
  outfile='/folders/myshortcuts/sarima-weight-loss/res/data/comparison-2.csv'
  dbms=csv replace;
run;

```

Código Fuente 11: Comparación utilizando el lenguaje SAS entre los modelos SARIMA $(0, 1, 1)(0, 1, 1)_{12}$  y SARIMA $(0, 1, 1)(0, 1, 1)_{12}(0, 0, 1)_{17}$  sobre la serie **weightloss**.

```

rm(list = ls())

library(ggplot2)
library(lubridate)
library(magrittr)
library(dplyr)
library(cowplot)

source("res/code/functions.r")

BASE_PATH <- './'
BASE_IMG_PATH <- paste0(BASE_PATH, 'res/img/')
BASE_DATA_PATH <- paste0(BASE_PATH, 'res/data/')

comparison.1 <- read.csv(paste0(BASE_DATA_PATH, 'comparison-1.csv'))
comparison.1$Month <- dmy(comparison.1$Month)

comparison.2 <- read.csv(paste0(BASE_DATA_PATH, 'comparison-2.csv'))
comparison.2$Month <- dmy(comparison.2$Month)

errors <- data.frame()
errors <- rbind(errors, list(name="{$\\text{SARIMA}(0, 1, 1)(0, 1, 1)_{12}$}",
                             error=PredictionError(comparison.1, 12),
                             stringsAsFactors = FALSE)
errors <- rbind(errors, list(name="{$\\text{SARIMA}(0, 1, 1)(0, 1, 1)_{12}(0, 0, 1)_{17}$}",
                             error=PredictionError(comparison.2, 12),
                             stringsAsFactors = FALSE)

errors$error <- round(errors$error, digits = 2)
write.csv(errors, paste0(BASE_DATA_PATH, 'predict-error.csv'), row.names = FALSE, quote = FALSE)

comparison.1.reduced <- comparison.1[(nrow(comparison.1) - 12 + 1):nrow(comparison.1), ]
comparison.2.reduced <- comparison.2[(nrow(comparison.2) - 12 + 1):nrow(comparison.2), ]

plot.1 <- comparison.1.reduced %>%
  mutate(L95 = L95 - Weight_loss,
        U95 = U95 - Weight_loss,
        FORECAST = FORECAST - Weight_loss,
        Weight_loss = 0) %>%
  {ggplot(., aes(y = FORECAST, ymin = L95, ymax = U95, x = Month)) +
    geom_line() +
    geom_point(aes(y = Weight_loss), size = 0.75, alpha = 0.5) +
    geom_ribbon(alpha = 0.2) +
    coord_flip() +
    theme_bw() +
    theme(plot.title = element_text(hjust = 0.5),
          panel.border = element_rect(colour = "black", fill=NA)) +
    ggtitle('SARIMA(0, 1, 1)(0, 1, 1)12')}
```

```

plot.2 <- comparison.2.reduced %>%
  mutate(L95 = L95 - Weight_loss,
        U95 = U95 - Weight_loss,
        FORECAST = FORECAST - Weight_loss,
        Weight_loss = 0) %>%
  {ggplot(., aes(y = FORECAST, ymin = L95, ymax = U95, x = Month)) +
    geom_line() +
    geom_point(aes(y = Weight_loss), size = 0.75, alpha = 0.5) +
    geom_ribbon(alpha = 0.2) +
    coord_flip() +
    theme_bw() +
    theme(plot.title = element_text(hjust = 0.5),
          panel.border = element_rect(colour = "black", fill=NA)) +
    ggtitle('SARIMA(0, 1, 1)(0, 1, 1)12(0, 0, 1)17')}
```

```

save_plot(paste0(BASE_IMG_PATH, 'comparison-ci-amplitude.png'), plot_grid(plot.1, plot.2),
          base_aspect_ratio = 1.75, base_height = 6)

t.test((comparison.1.reduced$U95 - comparison.1.reduced$L95) -
       (comparison.2.reduced$U95 - comparison.2.reduced$L95), mu = 0)

```

Código Fuente 12: Comparación utilizando el lenguaje R entre los modelos SARIMA $(0, 1, 1)(0, 1, 1)_{12}$  y SARIMA $(0, 1, 1)(0, 1, 1)_{12}(0, 0, 1)_{17}$  sobre la serie weightloss.

```

proc arima data = weightloss;
  identify var=weight_loss(1, 12);
  estimate q=(1)(12)(17) noint method=ml;
  forecast id=month interval=month lead=24 out=predict;
run;

proc export data=predict
  outfile='/folders/myshortcuts/sarima-weight-loss/res/data/predict.csv'
  dbms=csv replace;
run;

```

Código Fuente 13: Cálculo de la predicción utilizando el lenguaje SAS de las 24 observaciones siguientes utilizando el modelo SARIMA(0, 1, 1)(0, 1, 1)<sub>12</sub>(0, 0, 1)<sub>17</sub> sobre la serie **weightloss**.

```

rm(list = ls())

library(ggplot2)
library(lubridate)
library(magrittr)
library(cowplot)

BASE_PATH <- './'
BASE_IMG_PATH <- paste0(BASE_PATH, 'res/img/')
BASE_DATA_PATH <- paste0(BASE_PATH, 'res/data/')

predictions <- read.csv(paste0(BASE_DATA_PATH, 'predict.csv'))
predictions$Month <- dmy(predictions$Month)

head(predictions)
predictions[1:180, 'L95'] <- NA
predictions[1:180, 'U95'] <- NA
{ggplot(predictions, aes(y = FORECAST, ymin = L95, ymax = U95, x = Month)) +
  geom_line() +
  geom_point(aes(y = Weight_loss), size = 0.75, alpha = 0.5) +
  geom_ribbon(alpha = 0.2, colour = NA, fill = "Red") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5),
        panel.border = element_rect(colour = "black", fill=NA)) +
  ggtitle('Ajuste')} %>%
  { save_plot(paste0(BASE_IMG_PATH, 'predict-complete.png'),.,
    base_aspect_ratio = 1.5, base_height = 6) }

{ggplot(predictions[156:nrow(predictions), ], aes(y = FORECAST, ymin = L95, ymax = U95, x = Month)) +
  geom_line() +
  geom_point(aes(y = Weight_loss), size = 0.75, alpha = 0.5) +
  geom_ribbon(alpha = 0.2, colour = NA, fill = "Red") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5),
        panel.border = element_rect(colour = "black", fill=NA)) +
  ggtitle('Predicción')} %>%
  { save_plot(paste0(BASE_IMG_PATH, 'predict-ending.png'),.,
    base_aspect_ratio = 1.5, base_height = 6) }

```

Código Fuente 14: Cálculo de la predicción utilizando el lenguaje R de las 24 observaciones siguientes utilizando el modelo SARIMA(0, 1, 1)(0, 1, 1)<sub>12</sub>(0, 0, 1)<sub>17</sub> sobre la serie **weightloss**.