Análisis de Series Temporales: Tarea 3

Sergio García Prado

sergio.garcia.prado@alumnos.uva.es

31 de diciembre de 2018

- **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, en este trabajo se va a trabajar con 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 estructura estacional claramente marcada. Se cree que dicha estructura tiene un alto grado de relación 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 estocástica 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 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*.

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, entre los que destaca la crisis económica que comenzó en dicho año. Si se confirma 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. Otros factores podrían ser el acercamiento de la tecnología y redes sociales al gran público, que hasta entonces se habían mantenido alejadas del mismo.

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. Tal y 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 $i \in 1, 2, ..., 6$, esto es $1/12, 2/12, ..., 6/12$, 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 exista una leve relación entre el nivel de la serie y la dispersión del mismo. Este fenómeno podría requerir de alguna transfomación 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.

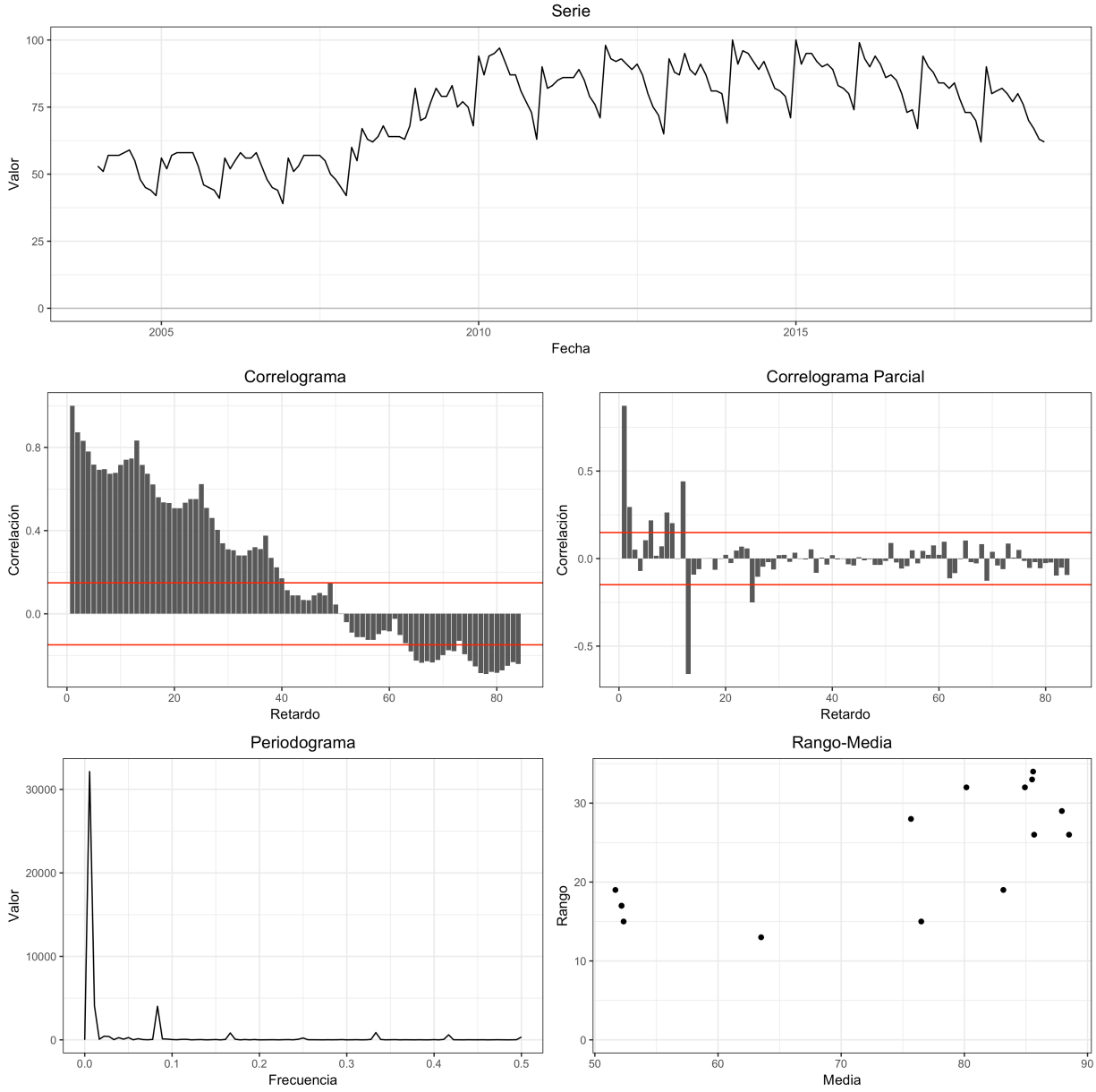

Figura 1: Gráfico de la serie, Correlograma, Correlograma Parcial, Periodograma y diagrama de dispersión *rango-media* para la serie `weightloss`.

1.3. Diferenciaciones

[TODO]

1.3.1. Diferenciación regular

[TODO]

1.3.2. Diferenciación estacional

[TODO]

1.3.3. Diferenciación regular y estacional

[TODO]

[TODO]

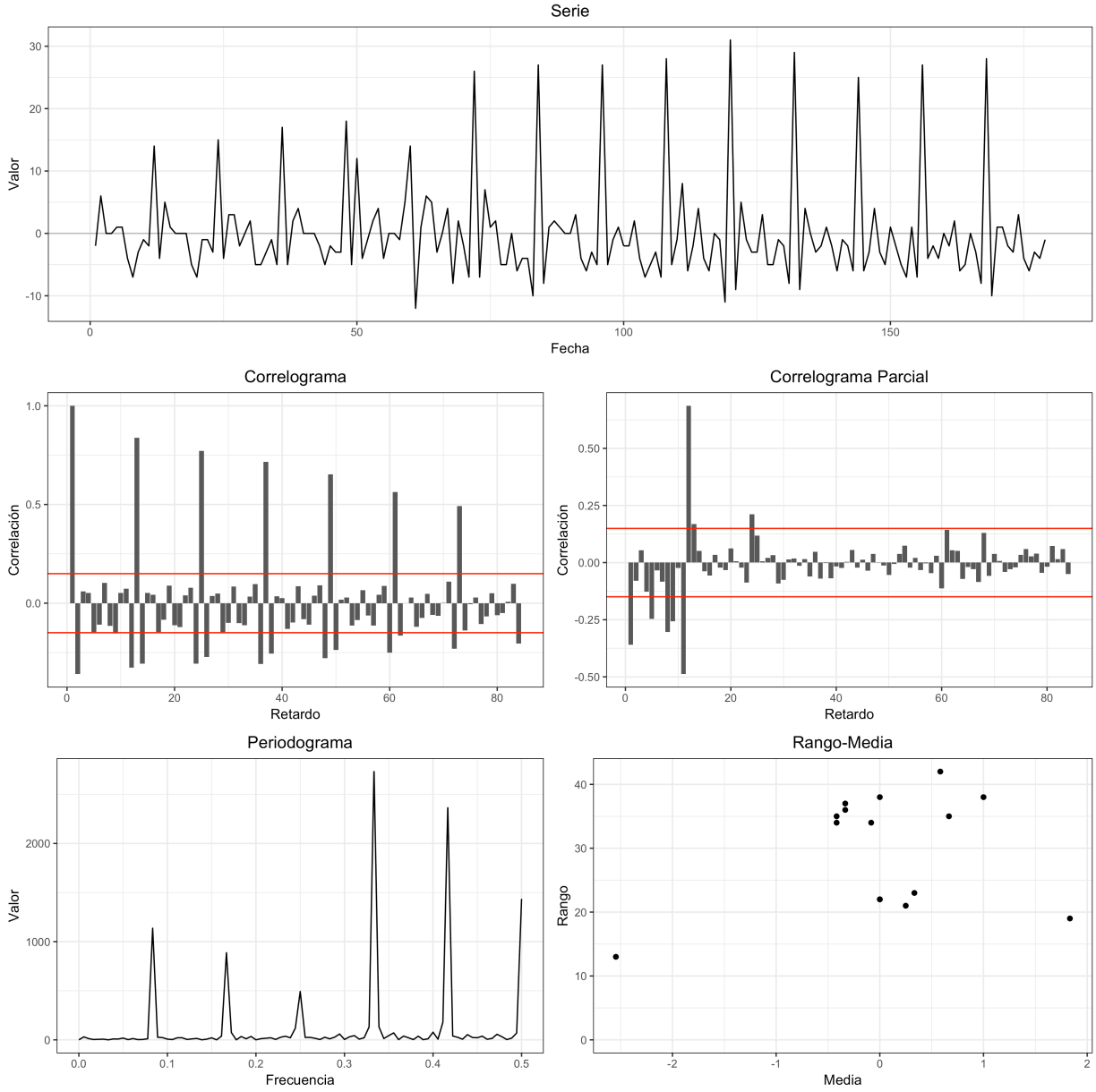

Figura 2: Gráfico de la serie, Correlograma, Correlograma Parcial, Periodograma y diagrama de dispersión *rango-media* para la serie `weightloss` tras la realización de una diferenciación regular.

1.4. Modelos propuestos

[TODO]

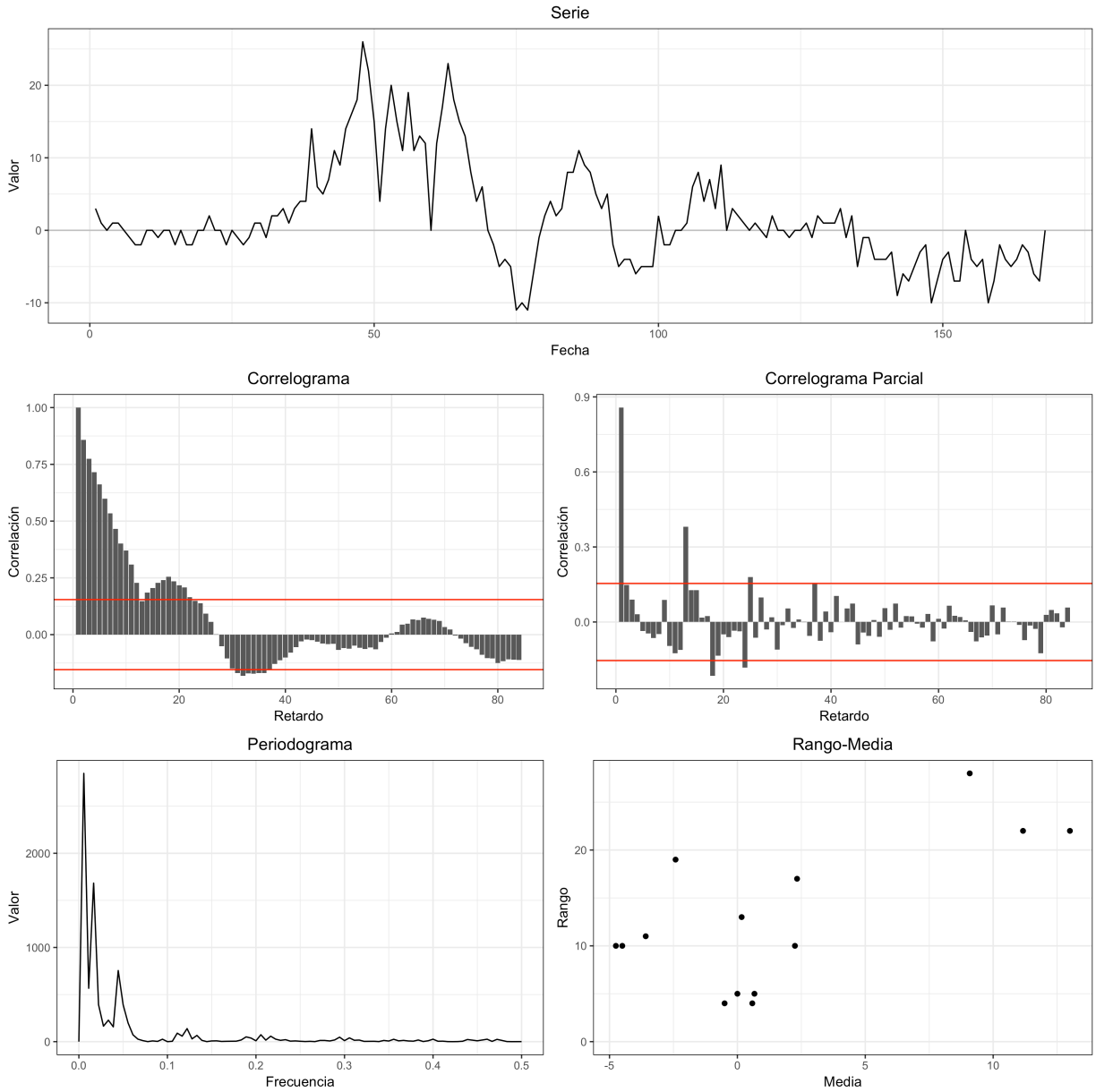

Figura 3: Gráfico de la serie, Correlograma, Correlograma Parcial, Periodograma y diagrama de dispersión *rango-media* para la serie `weightloss` tras la realización de una diferenciación estacional (12 retardos).

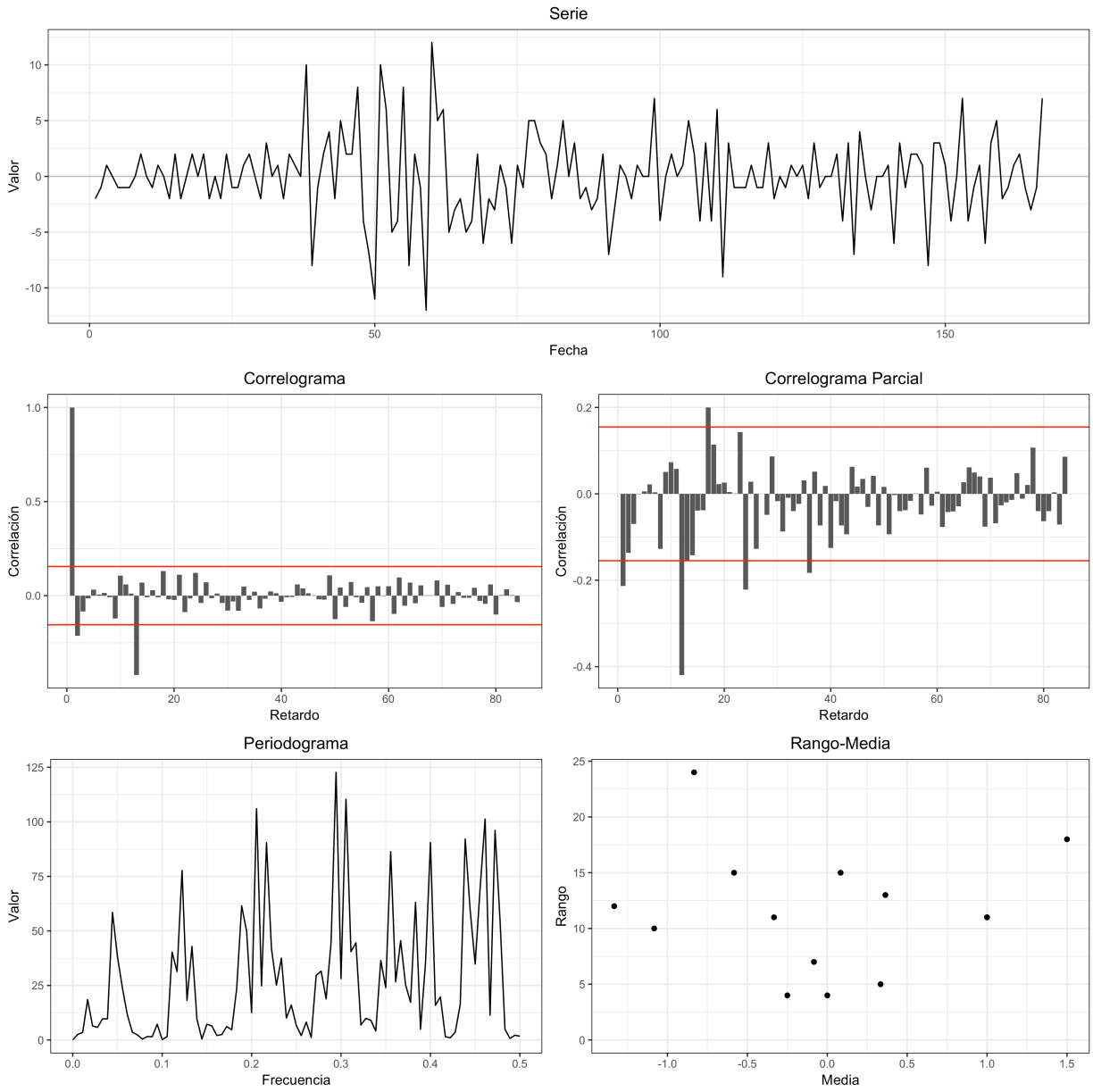

Figura 4: Gráfico de la serie, Correlograma, Correlograma Parcial, Periodograma y diagrama de dispersión *rango-media* para la serie `weightloss` tras la realización de una diferenciación regular y otra diferención estacional (12 retardos).

2. Etapa de estimación y validación

[TODO]

3. Comparación de modelos

[TODO]

4. Predicción

[TODO]

A. Código Fuente

[TODO]

```r
## Author: Sergio García Prado
## Title: Time Series - Weight Loss - EDA

rm(list = ls())

library(magrittr)
library(dplyr)
library(ggplot2)
library(latex2exp)
require(reshape2)
library(forecast)
library(cowplot)
library(lubridate)

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[0:n]
    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))
}

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() +
        # scale_x_continuous(breaks = scales::pretty_breaks(n = 10)) +
        theme(plot.title = element_text(hjust = 0.5),
              panel.border = element_rect(colour = "black", fill=NA)) +
        ggtitle('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)) +
        ggtitle('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)) +
        ggtitle('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") +
```