# Modelo para el histórico desde el año 2000 de la estimación del número de defunciones semanales en Palencia

Eric Jimenez Barril - 1599092
*Universitat Autònoma de Barcelona*
*Series Temporales*

(Dated: December 7, 2023)

## I. ADQUISICIÓN DE LOS DATOS.

En primer lugar, descargamos los datos del histórico desde el año 2000 hasta la semana 44 del año 2023 de la estimación del número de defunciones total semanales de la provincia de Palencia. [1].

### A. Lectura de los datos.

Como podemos ver en la Figura 1, los datos vienen dados por un *data frame* de 6 columnas de las cuales nos interesam solo las dos últimas, ya que son las que indican el periodo y el total de defunciones estimadas.

| | Total <chr> | Nacional <chr> | Provincias <int> | Tipo <chr> | de <chr> | dato <chr> | Periodo <chr> | Total.1 <int> |
|---|---|---|---|---|---|---|---|---|
| 1 | Total | Nacional | 34 | Palencia | Dato | base | 2023SM44 | 27 |
| 2 | Total | Nacional | 34 | Palencia | Dato | base | 2023SM43 | 46 |
| 3 | Total | Nacional | 34 | Palencia | Dato | base | 2023SM42 | 53 |
| 4 | Total | Nacional | 34 | Palencia | Dato | base | 2023SM41 | 33 |
| 5 | Total | Nacional | 34 | Palencia | Dato | base | 2023SM40 | 31 |
| 6 | Total | Nacional | 34 | Palencia | Dato | base | 2023SM39 | 42 |

FIG. 1: Datos dados por el INE para la estimación de defunciones en Palencia por semanas entre la primera semana del 2000 y la semana 44 de 2023.

### B. Limpieza de los datos.

Los datos no contienen valores faltantes, y queremos hacer un modelo de predicción para el número de defunciones, por lo que el único preporcesado a los datos que tenemos que hacer es invertirlos, pues empiezan a final de 2023 y van en orden descendiente de fecha.

Como podemos apreciar en la Tabla I no todos los años tienen las mismas semanas, por lo que definimos una nueva serie `Total` que va del año 2010 al año 2022 y coge las primeras 52 semanas de cada año.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 |
| 52 | 52 | 52 | 52 | 53 | 52 | 52 | 52 |
| 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
| 52 | 53 | 52 | 52 | 52 | 52 | 52 | 53 |
| 2016 | 2017 | 2018 | 2019 | 2020 | 2021 | 2022 | 2023 |
| 52 | 52 | 52 | 52 | 53 | 52 | 52 | 44 |

TABLE I: Semanas por año

## II. VISUALIZACIÓN DE DATOS.

Con tal de visualizar la serie `Total`, la dibujamos en la Figura 2 junto a su media y su media móvil.

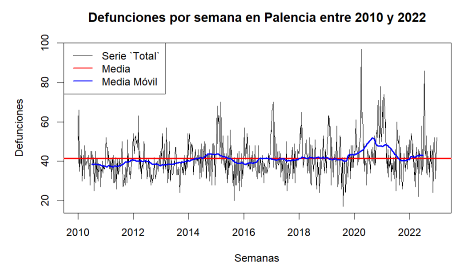


FIG. 2: Número de defunciones por semana en Palencia entre 2010 y 2022 junto a la media y la media móvil.

Podemos observar que la media movil no se ajusta del todo a la media, lo que indica que la serie tiene tendencia y además no es una recta, por lo que tiene componente estacional. Además, es claro que la varianza de la serie no es constante (picos que resaltan sobre la media).

## III. DESCOMPOSICIÓN DE LA SERIE TEMPORAL.

### A. Estudio de la tendencia.

Si la serie tiene tendencia, fijándonos en la Figura 2 aparentemente será creciente. Creando una regresión lineal respecto al tiempo, obtenemos el modelo de Ecuación 1

$$\text{Defunciones en semana t} = 38.26930 + 0.00916\text{t} \qquad (1)$$

donde $t \in \{1, \dots, 676\}$

Observamos que la media de los datos es 41.36982 y el *intercept* de la regresión es 38.26930; en caso de no haber tendencia, la media y el *intercept* coincidirian. Además, obtenemos un $p-valor$ de $6.637 \times 10^{-7}$ para el *test* con hipótesi nula que el coeficiente que acompaña la $t$ es 0, el cual es positivo. Por tanto con un 99.9% de confianza hay una tendencia creciente estadísticamente significativa.

### B. Estudio de la estacionalidad.

Estudiemos ahora si tiene estacionalidad, que en caso de tener, tendría que ser anual, ya que la serie son datos semanales de diferentes años.

Como podemos ver en la Figura 3, los *box* no se ajustan a los mismos rangos de valores, por lo que parece haber una estacionalidad anual.

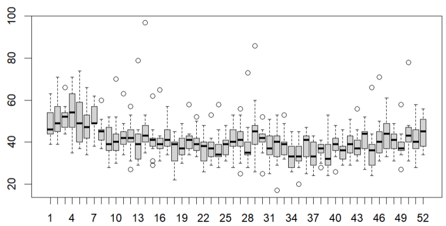

FIG. 3: `box.plot` de la serie `Total` con los datos agrupados por semanas.

Además, haciendo el truco basado el el *likelihood ratio test*, obtenemos un *p-valor* de 0, por lo que rechazamos la hipótesi nula de que el modelo con componente estacional y sin ella sean iguales, es decir, la serie tiene componente estacional anual.

Por último, si observamos en la Figrua 4 la componente estacional de la función `decompose()`, notamos que los valores de la componente estacional tienen un rango entre $-10$ y $\sim 15$, lo que representa aproximadamente un 30% de la varianza global de los datos.

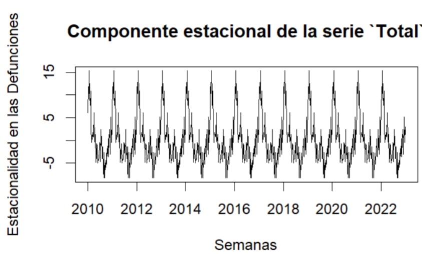


FIG. 4: Componente estacional de la serie `Total` según la función `decompose()`.

### C. Estudio de la serie diferenciada

Con tal de tratar de eliminar la tendencia y la estacionalidad, creamos la serie `Total_diff1_52` que podemos visualizar en la Figura 5, que es la serie `Total` diferenciada por la primera diferencia (para eliminar tendencia) y después por la diferencia 52 (para la estacionalidad).

Realizando el mismo prodecimiento que con la serie `Total`, obtenemos el `box.plot` de la Figura 6 y la gráfica

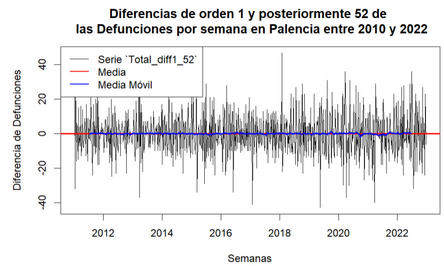


FIG. 5: Serie `Total` diferenciada por la primera diferencia y después por la de orden 52.

de la Figura 7 de la función `decompose()`. Observamos como los boxplots en la Figura 6 están bastante alineados y la variabilidad de la Figura 7 es $\sim 15\%$ de la variabilidad de la serie. Además, haciendo la prueba con el truco basado en el *likelihood ratio test* con las mismas hipótesis que antes, obtenemos un $p-valor$ igual a 1, por lo que no tenemos evidencias para rechazar la hipótesi nula. Por tanto concluimos que la serie `Total_diff1_52` no tiene componente estacional estadísticamente significativa.

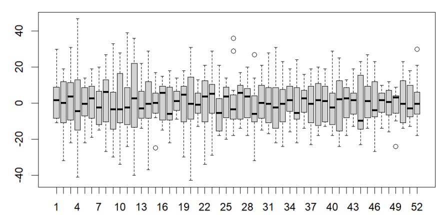

FIG. 6: Box-plot por semanas de la serie `Total_diff1_52`

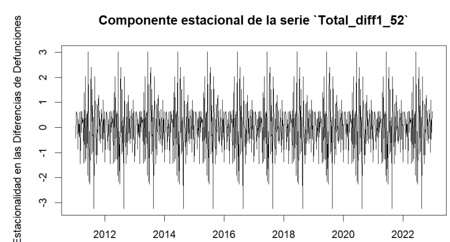


FIG. 7: Componente estacional de la serie `Total_diff1_52` según la función `decompose()`

Ahora, haciendo el *Augmented Dickey-Fuller test* a la serie `Total_diff1_52`, que tiene como hipótesis

$$\begin{cases} H_0\text{: La serie tiene una raíz unitaria y no es estacionaria.} \\ H_1\text{: La serie tiene no tiene raices unitarias y es estacionaria.} \end{cases}$$

obtenemos un *p-valor* menor que 0.01, por lo que podemos rechazar la hipótesi nula y con un 99.9% de confianza, decir que la serie es estacionaria (tiene sentido tanto analíticamente, pues hemos diferenciado la serie para ello, como cualitativamente observando la Figura 5).

### D. Estudio de la variable *COVID*.

Si añadimos una variable *C:=COVID* que represente un vector de unos entre las semanas de marzo de 2021 y todo 21, y el resto 0, al crear un modelo de regresión de la serie `Total` con respecto de la variable $C$, obtenemos el modelo de Ecuación 2

$$\text{Defunciones en semana t} = 40.585 + 5.707C \qquad (2)$$

Haciendo un *test* de coeficientes para el coeficiente que acompaña a $C$, obtenemos un $p - valor$ de $4.323 \times 10^{-8}$, por lo que la variable $C$ es estadísticamente significativa para nuestros datos, es decir, el periodo donde empezó el $C$ y tuvo más impacto, influye en nuestra serie.

## IV. MODELADO Y PREDICCIÓN.

### A. Propuesta de un modelo para la serie diferenciada según su EACF, ACF y PACF

En primer lugar, observando la EACF de la serie `Total_diff1_52` en la Figura 9 diríamos que el modelo para la serie es un `MA(1)`, lo que coincide con lo que observamos en la Figura 8 pues la ACF únicamente tiene el *lag* 1 estadísticamente distito de 0 y la PACF decae a 0 de forma exponencial. Por tanto en caso de tener que escoger un modelo para la serie `Total_diff1_52` basándonos en la EACF, ACF y PACF sería un `MA(1)`.

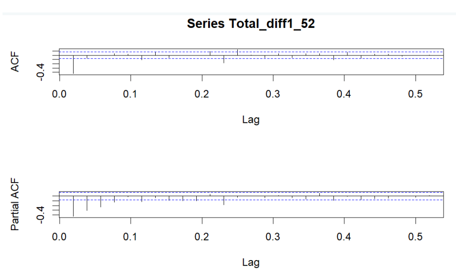


FIG. 8: ACF y PACF de la serie `Total_diff1_52`.

| AR/MA | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | x | o | o | o | o | x | o | o | o | o | x | x | x | o |
| 1 | x | x | o | o | o | x | o | o | o | o | o | x | x | o |
| 2 | x | x | o | o | o | x | o | o | o | o | o | o | x | o |
| 3 | x | x | o | x | o | x | o | o | o | o | o | o | x | o |
| 4 | x | x | x | x | o | o | o | o | o | o | o | o | x | o |
| 5 | x | x | x | x | o | x | o | o | o | o | o | o | x | o |
| 6 | x | x | x | x | o | x | x | o | o | o | o | o | x | o |
| 7 | x | x | x | x | o | x | o | o | o | o | o | x | o | o |

FIG. 9: EACF de la serie `Total_diff1_52`.

### B. Creación del primer modelo para la serie original.

Si creamos un modelo para la serie `Total` con la función `auto.arima()`, nos aconseja un modelo `ARIMA(1,1,2)(0,0,1)[52]` con drift , es decir, nos recomienda un modelo `ARIMA(1,1,2)` con *drift* para la componente no estacional y un modelo `SARIMA(0,0,1)[52]` para la componente estacional, con los coeficientes de la Tabla II, los cuales podemos observar que son todos significativos y un AIC=4787.75 para el modelo.

Haciendo un estudio de los residuos, obtenemos en el *Ljung-box test* un *p-valor* de 0.1076 por lo que no podemos rechazar la hipótesi nula de que los residuos son independientes. No obstante, el el *shapiro-wilks test* obtenemos un *p-valor* de $4.297 \times 10^{-5}$ por tanto descartamos que se distribuyan de forma normal. Haciendo un `checkresiduals`, vemos en la Figura 10 que los residuos no parecen ruido blanco, ya que en la ACF parecen correlacionados y en su *plot* parecen no tener varianza constante. Por tanto según en Ljung-box test son ruido blanco pero no cualitativamente, por tanto no lo consideramos un modelo adecuado (seguramente tenemos que modelar la varianza, ya que no parece constante).

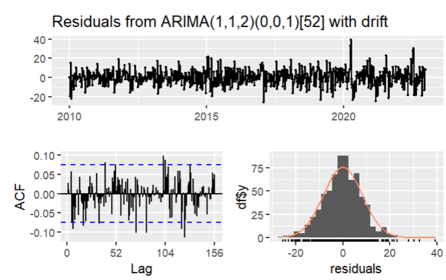


FIG. 10: Estudio de los residuos del modelo dado por `auto.arima()`

### C. Ajuste de un modelo introduciendo la variable $C$ como variable explicativa.

Hagamos ahora lo mismo pero añadiendo el regresor $C$. De esta manera, obtenemos el modelo `ARIMA(2,0,2)(0,0,1)[52]` con *intercept* y el regresor y los coeficientes de la Tabla III. En este caso, obtenemos un AIC=4752.28, pero no todos los coeficientes son significativos. Por lo que ajustamos el modelo pero fijando el coeficiente `ar1=0`. De esta forma, obtenemos el mismo modelo, esta vez con los coeficientes de la Tabla IV que son todos significativos y un AIC= 4749.41. Por lo que a priori, este modelo debería ser mejor que el estimado sin la variable regresora $C$, por lo que la aportación de la variable regresora $C$ es significativa.

### D. Predicción y bondad de ajuste para 2023 con el modelo creado a partir de la serie `Total`

Si hacemos las predicciones para el año 2023 con el modelo `ARIMA(1,1,2)(0,0,1)[52]` con *drift*, obtenemos las estimaciones de la Tabla V. En la Figura 11 podemos ver la serie `Total` junto a las predicciones para 2023 y en la Figura 12 podemos ver los valores predichos para 2023, junto a los valores reales y el intervalo de confianza 95% de predicción. Notamos que el 100% de los valores están dentro del intervalo de confianza 95%, por lo que a pesar de no parecer un buen modelo, predice bien nuestra serie.

Observemos también que si nos fijamos en los datos reales de las muertes en Palencia en 2023 hasta la semana 44, se han estabilizado bastante las medias, debido seguramente a que ya nadie (o prácticamente nadie) muere por $C$ o enfermedades fuera de las habituales (Si hubieramos hecho el modelo entrenado del 2000 al 2019, no hubiera predicho los datos del 2020-21 o incluso 22, donde aún se ve el reflejo de la última ola). [2]

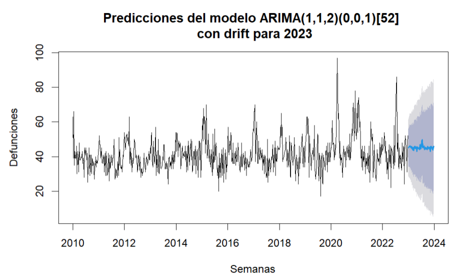


FIG. 11: Serie `Total` junto a las predicciones para 2023 dadas por el modelo `ARIMA(1,1,2)(0,0,1)[52]` con *drift*.

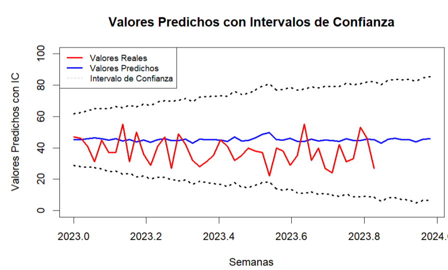


FIG. 12: Predicciones para 2023 junto a intervalo de confianza 95% dados por el modelo `ARIMA(1,1,2)(0,0,1)[52]` con *drift* y junto a los datos reales

## V. ANÁLISI DE LOS RESIDUOS

Estudiemos ahora los residuos y los residuos al cuadrado del modelo. En la Figura 10 ya podemos observar la gráfica y la ACF de los residuos, por tanto es suficiente que estudiemos la PACF de los residuos y los residuos al cuadrado. Como podemos observar tanto en la Figura 10 como en la Figura 13, no se ajustan a ser la ACF y PACF de un ruido blanco. Además, en las Figuras 14 y 15 se ve claramente como la varianza de los residuos al cuarado no es constante y su ACF y PACF no se ajustan a ruido blanco de nuevo. Por tanto, la estructura GARCH en los residuos de estos datos parece razonable.

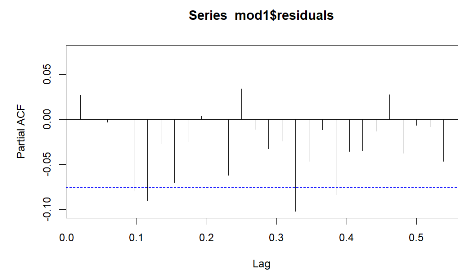


FIG. 13: PACF de los residuos de la serie `Total`.

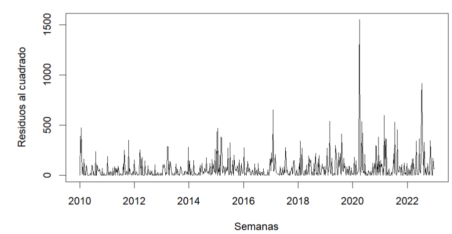


FIG. 14: Cuadrados de los residuos de la serie `Total`.

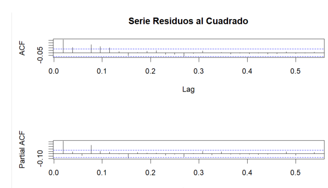


FIG. 15: ACF y PACF del cuadrado de los residuos de la serie `Total`.

## VI. AJUSTE DE UN MODELO GARCH

Por último, ajustamos los modelos `GARCH(1,0)`, `GARCH(1,1)`, `GARCH(2,1)` a la serie `Total`. Obtenemos los AIC de la Tabla VIa, por tanto, nos quedamos con el modelo `GARCH(1,1)` con Ecuación 3.

$$\begin{cases} \text{Defunciones en semana t} = 39.45 + \sigma_t \epsilon_t \\ \sigma_t^2 = 22.6789 + 0.403\epsilon_{t-1}^2 + 0.338\sigma_{t-1}^2 \end{cases} \tag{3}$$

donde $\epsilon_t \sim WN(0,1)$.

## VII. CONCLUSIONES

No obstante, los 3 modelos `GARCH` muestran peores AIC que los dos primeros modelos que hemos estimado. Estudiemos estos dos.

Si hacemos las predicciones con el modelo con la variable $C$, como podemos ver en la Figura 16, en las predicciones para 2023 junto al intervalo de confianza 95% y los datos reales que obtenemos con el modelo, vemos como hay almenos uno de los valores que no pertenece al intervalo de confianza. Si calculamos el $RMSE$ y $R^2$ de ambos modelos, como podemos ver en la Tabla VIb, ambos coeficientes $R^2$ son negativos, lo que quiere decir que la media predice mejor que los modelos. Observando la Figura 16 o 11 notamos que las predicciones van por encima de gran parte de los datos, lo que indica este suceso. En definitiva, de los 5 modelos estudiados, el modelo `ARIMA(2,0,2)(0,0,1)[52]` con *intercept* y variable de regresión $C$, es el mejor modelo para predecir nuestros datos, ya que a pesar de predecir una observación fuera del intervalo de confianza, es el que tiene mejores métricas y mejor ajusta los datos.

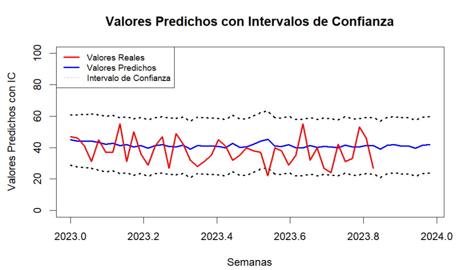


FIG. 16: Predicciones para 2023 junto al intervalo de confianza 95% y los datos reales obtenidos con el modelo `ARIMA(2,0,2)(0,0,1)[52]` con *intercept* y variable de regresión $C$.

## VIII. ANEXOS

### A. Tablas de coeficientes y $p-valores$.

#### 1. Modelo `ARIMA(1,1,2)(0,0,1)[52]` con drift

| Término | ar1 | ma1 |
|---|---|---|
| Coeficiente | -0.9075 | 0.2380 |
| *p-valor* | $< 2.2 \times 10^{-16}$ | $2.47 \times 10^{-5}$ |
| Término | ma2 | sma1 |
| Coeficiente | -0.6547 | 0.1094 |
| *p-valor* | $< 2.2 \times 10^{-16}$ | 0.004 |

TABLE II: Coeficientes del modelo junto a sus *p-valores* obtenidos haciendo el *test* para ver si son distintos de 0.

#### 2. Modelo `ARIMA(2,0,2)(0,0,1)[52]` con intercept y regresor C

| Términi | ar1 | ar2 | ma1 |
|---|---|---|---|
| Coeficiente | -0.086853 | 0.772852 | 0.388345 |
| *p-valor* | 0.170798 | $< 2.2 \times 10^{-16}$ | $4.22 \times 10^{-8}$ |
| Términi | ma2 | sma1 | C |
| Coeficiente | -0.525728 | 0.107998 | 6.570579 |
| *p-valor* | $< 2.2 \times 10^{-16}$ | 0.004209 | 0.004715 |

TABLE III: Coeficientes del modelo junto a los *p-valores*.

| Términi | ar2 | ma1 | ma2 |
|---|---|---|---|
| Coeficiente | 0.717255 | 0.388345 | -0.525728 |
| *p-valor* | $< 2.2 \times 10^{-16}$ | $< 2.2 \times 10^{-16}$ | $< 4.63 \times 10^{-11}$ |
| Términi | sma1 | C | |
| Coeficiente | 0.107998 | 6.570579 | |
| *p-valor* | 0.004864 | 0.004860 | |

TABLE IV: Coeficientes del modelo anterior fijando $ar1$=0, junto a los *p-valores*.

### B. Predicciones para el año 2023 con el modelo `ARIMA(1,1,2)(0,0,1)[52]` con *drift*.

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| S01 | S02 | S03 | S04 | S05 | S06 | S07 | S08 | S09 | S10 | S11 | S12 | S13 |
| 45.2 | 45.2 | 45.7 | 46.2 | 45.7 | 44.9 | 45.8 | 44.4 | 45.4 | 43.9 | 45.0 | 43.5 | 45.2 |
| S14 | S15 | S16 | S17 | S18 | S19 | S20 | S21 | S22 | S23 | S24 | S25 | S26 |
| 45.8 | 44.8 | 44.6 | 45.6 | 43.1 | 45.5 | 45.3 | 45.1 | 44.9 | 44.1 | 46.9 | 44.4 | 44.6 |
| S27 | S28 | S29 | S30 | S31 | S32 | S33 | S34 | S35 | S36 | S37 | S38 | S39 |
| 46.5 | 48.5 | 49.7 | 45.3 | 45.3 | 45.0 | 46.1 | 44.1 | 44.2 | 45.4 | 44.3 | 45.0 | 44.6 |
| S40 | S41 | S42 | S43 | S44 | S45 | S46 | S47 | S48 | S49 | S50 | S51 | S52 |
| 44.1 | 45.8 | 44.5 | 44.7 | 45.4 | 45.3 | 43.0 | 45.6 | 46.0 | 45.2 | 43.7 | 45.6 | 45.9 |

TABLE V: Predicción para las defunciones de 2023.

### C. Tablas comparativas de métricas entre modelos.

| Modelo | AIC |
|---|---|
| `ARCH(1)` | 6962.990 |
| `GARCH(1,1)` | 6961.897 |
| `GARCH(2,1)` | 6987.470 |

(a) AIC para cada uno de los 3 modelos GARCH ajustados

| Métrica | Modelo 1 | Modelo 2 |
|---|---|---|
| AIC | 4787.75 | 4749.41 |
| RMSE | 11.423857 | 9.232847 |
| $R^2$ | -0.9378637 | -0.2658115 |

(b) Comparación de las métricas entre los modelos considerando y sin considerar la variable regresiva $C$.

TABLE VI: Tablas comparativas de métricas entre modelos.

## D. Código

```
#Librer as
library(forecast); library(lmtest); library(TSA)
library(fGarch); library(tseries);
library(tdyverse);library(dplyr)

## LECTURA DE LES DADES
setwd("C:/Users/ERIC/Desktop/5. SERIES TEMPORALES/PR CTICAS/LLIURAMENT 2")
Palencia <- read.table("Palencia.csv", header=T); head(Palencia)

## PRE-PROCESADO
Palencia <- Palencia %>% dplyr::select(Periodo, Total.1)
cat("El total de valores faltantes de Total.1 es ", sum(is.na(Palencia$Total.1)))

## SEMANAS POR A O
semanes_per_any <-c(1, rep(0, 23)) #Ponemos la primera de 2023
a = 1
ano = 2000
for (i in 2:length(Palencia$Periodo)){
  if(substr(Palencia$Periodo[i], 1, 4)==substr(Palencia$Periodo[i-1], 1, 4)){
    semanes_per_any[a]=semanes_per_any[a]+1}
  else{
    a=a+1
    semanes_per_any[a]=semanes_per_any[a]+1}}
for (i in 1:length(semanes_per_any)){
  cat("El a o ", ano+(i-1), "tiene ", semanes_per_any[i], "semanas \n")}

## DEFINICI N SERIE ‘TOTAL‘
Palencia <- as.data.frame(lapply(Palencia, rev))
files_a_excluir = c(261, 522, 835, 1096, 1201:1244)
Total <- Palencia[-files_a_excluir,]
nrow(Total)

Total <- ts(Total$Total.1[521:(sum(semanes_per_any))], frequency=52, start=c(2010,01), end=c(2022,52))

## PLOT SERIE TOTAL
plot(Total, ylab="Defunciones", xlab="Semanas", main="Defunciones por semana en Palencia entre 2010 y 2022")
abline(h=mean(Total), col="red", lwd=2)
lines(ma(Total, order = 52, centre = T),col="blue", lwd=2)
legend('topleft', legend=c('Serie ‘Total‘','Media', 'Media M vil'), col=c('black','red', 'blue'), lty=c(1, 1,1), lwd=c(1,1.5, 1.5), cex=1)
mean(Total)

## REGRESI N ‘TOTAL‘ RESPECTO EL TIEMPO
t <- 1:(length(Total))
total.lm <- lm(Total ~ t)

## TEST COEFICIENTES DE REGRESI N
coeftest(total.lm)

## BOX-PLOT SERIE ‘TOTAL‘
matrix_data_t=t(matrix(data=Total, nrow=52))
boxplot(matrix_data_t)

## PRUEBA ESTACIONALIDAD LIKELIHOOD
fit1 <- ets(as.vector(Total))
fit2 <- ets(as.vector(Total), model="AZN")
deviance <- 2*c(logLik(fit1) - logLik(fit2))
df <- attributes(logLik(fit1))$df - attributes(logLik(fit2))$df
#P value
1-pchisq(deviance,df) H0: No es estacional / H1: es estacional

## ESTACIONALIDAD (DECOMPOSE())
plot(decompose(Total)$seasonal, ylab="Estacionalidad en las Defunciones", xlab="Semanas", main="Componente estacional de la serie ‘Total‘")

## DEFINICI N SERIE ‘TOTAL_DIFF1_52‘
Total_diff1_52 <- diff(diff(Total), lag=52)

## PLOT SERIE ‘TOTAL_DIFF1_52‘
plot.ts(Total_diff1_52, ylab="Diferencia de Defunciones", xlab="Semanas", main="Diferencias de orden 1 y posteriormente 52 de \n las Defunciones por semana en Palencia entre 2010 y 2022")
abline(h=mean(Total_diff1_52), col="red", lwd=2)
lines(ma(Total_diff1_52, order = 52, centre = T),col="blue", lwd=2)
legend('topleft', legend=c('Serie ‘Total_diff1_52‘','Media', 'Media M vil'), col=c('black','red', 'blue'), lty=c(1, 1,1), lwd=c(1,1.5, 1.5), cex=1)

## BOX-PLOT SERIE ‘TOTAL_DIFF1_52‘
matrix_data_t=t(matrix(data=Total_diff1_52, nrow=52))
boxplot(matrix_data_t)

## PRUEBA ESTACIONALIDAD LIKELIHOOD
fit1 <- ets(as.vector(Total_diff1_52))
fit2 <- ets(as.vector(Total_diff1_52), model="AZN")
deviance <- 2*c(logLik(fit1) - logLik(fit2))
df <- attributes(logLik(fit1))$df - attributes(logLik(fit2))$df
#P value
1-pchisq(deviance,df)

## ESTACIONALIDAD (DECOMPOSE())
plot(decompose(Total_diff1_52)$seasonal, ylab="Estacionalidad en las Diferencias de Defunciones", xlab="Semanas", main="Componente estacional de la serie ‘Total_diff1_52‘")

## TEST AUGMENTED DICKEY FULLER A SERIE ‘TOTAL_DIFF1_52‘
adf.test(Total_diff1_52)

## EACF, ACF, PACF ‘TOTAL_DIFF1_52‘
eacf(Total_diff1_52)
par(mfrow=c(2,1))
acf(Total_diff1_52)
pacf(Total_diff1_52, main="")
par(mfrow=c(1,1))

## MODELO auto.arima() PARA LA SERIE ‘TOTAL‘
mod1 <- auto.arima(Total)
```

```
## TEST PARA LOS COEFICIENTES DEL MODELO
coeftest(mod1)

## ESTUDIO DE LOS RESIDUOS PARA EL MODELO
checkresiduals(mod1)

## SHAPIRO WILK TEST PARA LOS RESIDUOS
shapiro.test(mod1$residuals)

## PREDICCIONES 2023 CON EL MODELO
pred1 <- forecast(mod1, h=52)

## DEFINICI N DE LOS DATOS DE DEFUNCI N 2023
real <- ts(tail(Palencia,44)$Total.1, frequency=52, start=c(2023,01), end=c(2023,44))

## VISUALIZAMOS LA PREDICCI N
plot(pred1, xlab="Semanas", ylab="Defunciones", main="Predicciones del modelo ARIMA(1,1,2)(0,0,1)[52] \n con drift para 2023")

plot(pred1$mean, type='l', lwd=2, col='blue', xlab='Semanas', ylab='Valores Predichos con IC', main='Valores Predichos con Intervalos de Confianza', ylim=c(0, 100))
lines(pred1$lower[,2], col='black', lty=3, lwd=2)
lines(pred1$upper[,2], col='black', lty=3, lwd=2)
lines(real, col='red', lwd=2)
legend('topleft', legend=c('Valores Reales','Valores Predichos', 'Intervalo de Confianza'), col=c('red','blue', 'gray'), lty=c(1,1,2), lwd=c(2,2, 1), cex=0.8)

## ESTUDIO DE LOS RESIDUOS (continuaci n)

pacf(mod1$residuals)

plot((mod1$residuals)^2, ylab="Residuos al cuadrado", xlab="Semanas")
par(mfrow=c(2,1))
acf((mod1$residuals)^2, main="Serie Residuos al Cuadrado")
pacf((mod1$residuals)^2, main="")

par(mfrow=c(1,1))

## MODELOS GARCH
mod3 <- garch(Total, order=c(0, 1))
mod4 <- garch(Total, order=c(1, 1))
mod5 <- garch(Total, order=c(1, 2))

## AIC de los modelos
c(AIC(mod3), AIC(mod4), AIC(mod5))

## MEJOR MODELO GARCH
mod4 <- garchFit(~garch(1,1), Total)
coef(mod4)

## AJUSTE A ADIENDO REGRESOR COVID
COVID<-c(rep(0,531), rep(1,93), rep(0,52))

## REGRESI N `TOTAL` RESPECTO A `COVID`
Total.lm2 <- lm(Total~COVID)

## TEST PARA LOS COEFICIENTES DEL MODELO
coeftest(Total.lm2)

## MODELO auto.arima() A ADIENDO EL REGRESOR COVID
mod_cov <- auto.arima(Total, xreg=COVID)

## TEST PARA LOS COEFICIENTES DEL MODELO
coeftest(mod_cov)

## FIJAMOS QUE EL COEFICIENTE ar1 SEA 0
mod_cov <- arima(Total, order=c(2,0,2), seasonal=list(order=c(0,0,1), period=52), xreg=COVID, include.mean=TRUE, fixed=c(0, NA, NA, NA, NA, NA, NA))

## TEST PARA LOS COEFICIENTES DEL MODELO
coeftest(mod_cov)

## ESTUDIAMOS LOS RESIDUOS
checkresiduals(mod_cov)

## PREICCIONES 2023
pred2 <- forecast(mod_cov, h = 52, xreg=numeric(52))

## PLOT BONDAD DE AJUSTE
plot(pred2$mean, type='l', lwd=2, col='blue', xlab='Semanas', ylab='Valores Predichos con IC', main='Valores Predichos con Intervalos de Confianza', ylim=c(0, 100))
lines(pred2$lower[,2], col='black', lty=3, lwd=2)
lines(pred2$upper[,2], col='black', lty=3, lwd=2)
lines(real, col='red', lwd=2)
legend('topleft', legend=c('Valores Reales','Valores Predichos', 'Intervalo de Confianza'), col=c('red','blue', 'gray'), lty=c(1,1,2), lwd=c(2,2, 1), cex=0.8)

## RMSE DE LOS DOS MODELOS ARIMA
rmse1 = sqrt(mean(((pred1$mean)[1:44] - real)^2))
rmse2 = sqrt(mean(((pred2$mean)[1:44] - real)^2))
c(rmse1, rmse2)

## C LCULO DEL R2 DE LOS DOS MODELOS ARIMA
SSR1 = sum(((pred1$mean)[1:44] - real)^2)
SST1 = sum((mean(real)-real)^2)
SSR2 = sum(((pred2$mean)[1:44] - real)^2)

c(1- SSR1/SST1, 1- SSR2/SST1)
```


[1] Estimación del número de defunciones semanales de por provincia (2000–).

[2] La oms decreta el fin de la emergencia internacional por la covid (2023).