Serie de tiempo de visitantes de Canadá

Julian Alejandro Usuga Ortiz - Ivan Santiago Rojas Martinez

2022-11-16

Section 1

Nuestra serie de tiempo

Nuestra serie de tiempo

Nuestra serie consiste en el numero de extranjeros que visitaron a Canadá por mes. En este numero se encuentran personas que viajan por negocios, estudios o turismo.

Nuestra serie de tiempo

Nuestra serie consiste en el numero de extranjeros que visitaron a Canadá por mes. En este numero se encuentran personas que viajan por negocios, estudios o turismo.

Esta serie es importante ya que esta cifra representa mucho para la economía de un país. Poder saber el numero de visitantes es importante y mucho mas después de que el COVID cambiara tanto los hábitos de muchas personas.

Section 2

Datos usados

Obtención de datos

Los datos fueron obtenidos del sitio web **Statistics Canada** que es la oficina nacional de estadística de este país.

Obtención de datos

Los datos fueron obtenidos del sitio web **Statistics Canada** que es la oficina nacional de estadística de este país.

Statistics Canada. Table 24-10-0050-01 Non-resident visitors entering Canada, by country of residence

DOI: https://doi.org/10.25318/2410005001-eng

Lectura

Al descargar los datos encontramos la siguiente estructura

Geography													
Country of residence[2]	January 1990	February 1990	March 1990	April 1990	May 1990	June 1990	July 1990	August 1990	September 1990	October 1990	November 1990	December 1990	January 1991
Non-resident visitors entering Canada	1,852,514	1,721,936	2,216,404	2,465,442	3,244,040	4,278,689	5,635,015	5,773,933	3,652,984	2,761,067	2,242,776	2,145,670	1,686,542

Figure 1: Estructura original de los datos.

	A	B	C	D	E	F	G
1	Non-resident visitors entering Canada, by country of residence 1						
2	Frequency: Monthly						
3	Table: 24-10-0050-01						
4	Release date: 2022-10-24						
5	Geography: Canada, Province or territory						
6							
7							
8							
9	Geography	Canada					
10	Country of residence 2	September 1972	October 1972	November 1972	December 1972	January 1973	February
11		Visitors					
12	Non-resident visitors entering Canada	3628550	2473249	1855704	1730150	1508870	149

Figure 2: Estructura original de los datos.

Limpieza

```
# data tiene 1 fila y 601 columnas
data
```

```
## # A tibble: 1 x 601
##   Country of r~1 Septe~2 Octob~3 Novem~4 Decem~5 Janua~6
##   <chr>          <chr>     <dbl>   <dbl>   <dbl>   <dbl>
## 1 Non-resident ~ 3,628,~ 2473249 1855704 1730150 150887(
## # ... with 592 more variables: 'May 1973' <dbl>, 'June 1
## #   'July 1973' <dbl>, 'August 1973' <dbl>, 'September 1
## #   'October 1973' <dbl>, 'November 1973' <dbl>, 'Decemb
## #   'January 1974' <dbl>, 'February 1974' <dbl>, 'March
## #   'April 1974' <dbl>, 'May 1974' <dbl>, 'June 1974' <d
## #   'August 1974' <dbl>, 'September 1974' <dbl>, 'Octobe
## #   'November 1974' <dbl>, 'December 1974' <dbl>, 'Janua
## # i Use 'colnames()' to see all variable names
```

```r
# Pasar a formato largo
data <- data |> gather(key = "date", value = "visitors")

# Quitar el primer dato
data <- data[-1, ]

# Quitar las comas de los números
data$visitors <- sapply(data$visitors,
                        gsub,
                        pattern = ",",
                        replacement= "")

# Convertir la columna a numérica
data$visitors <- as.integer(data$visitors)

# Convertir de tipo string "mes año" a tipo fecha
data$date <- my(data$date)
```

```
head(data)
```

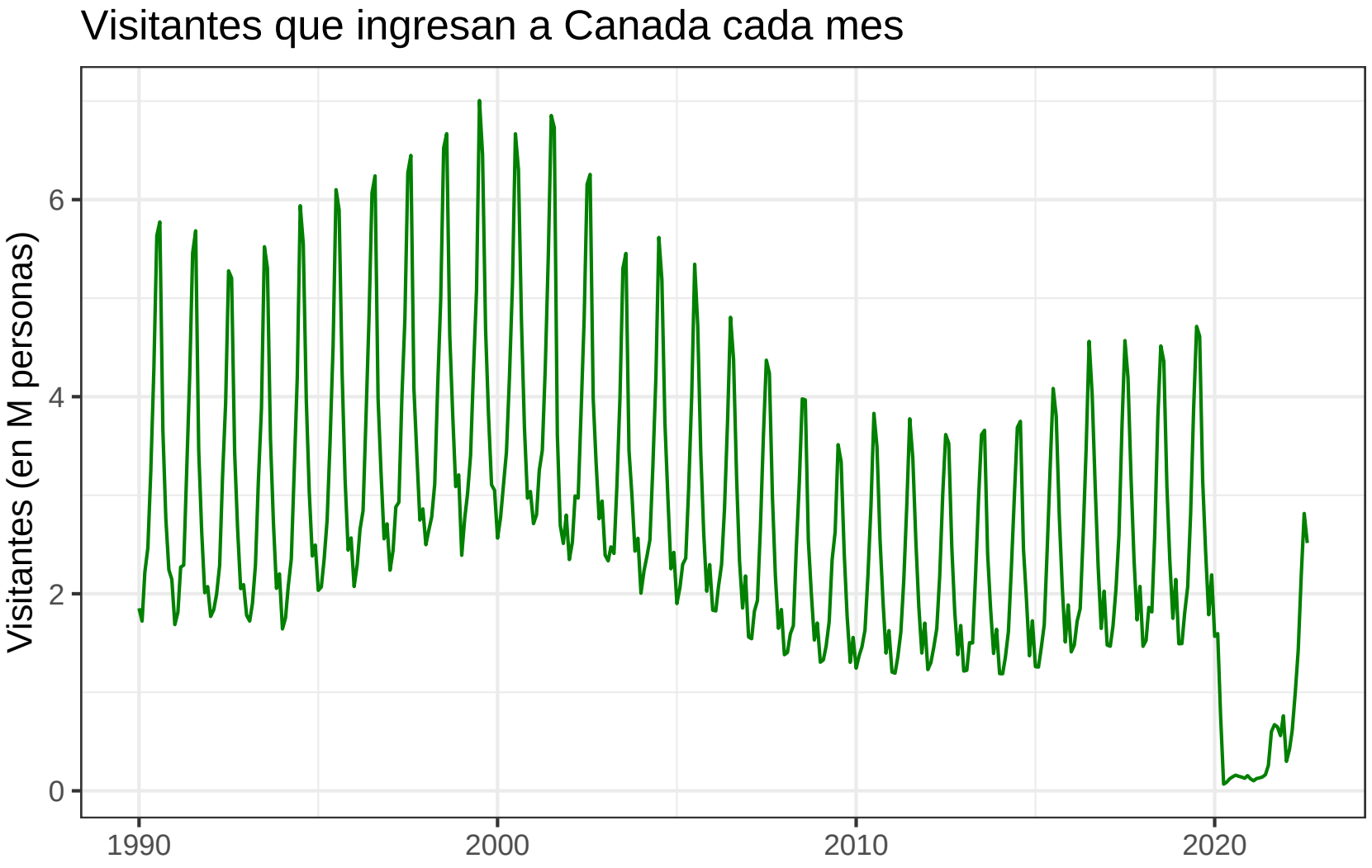
```
## # A tibble: 6 x 2
##   date       visitors
##   <date>        <int>
## 1 1972-09-01  3628550
## 2 1972-10-01  2473249
## 3 1972-11-01  1855704
## 4 1972-12-01  1730150
## 5 1973-01-01  1508870
## 6 1973-02-01  1496932
```

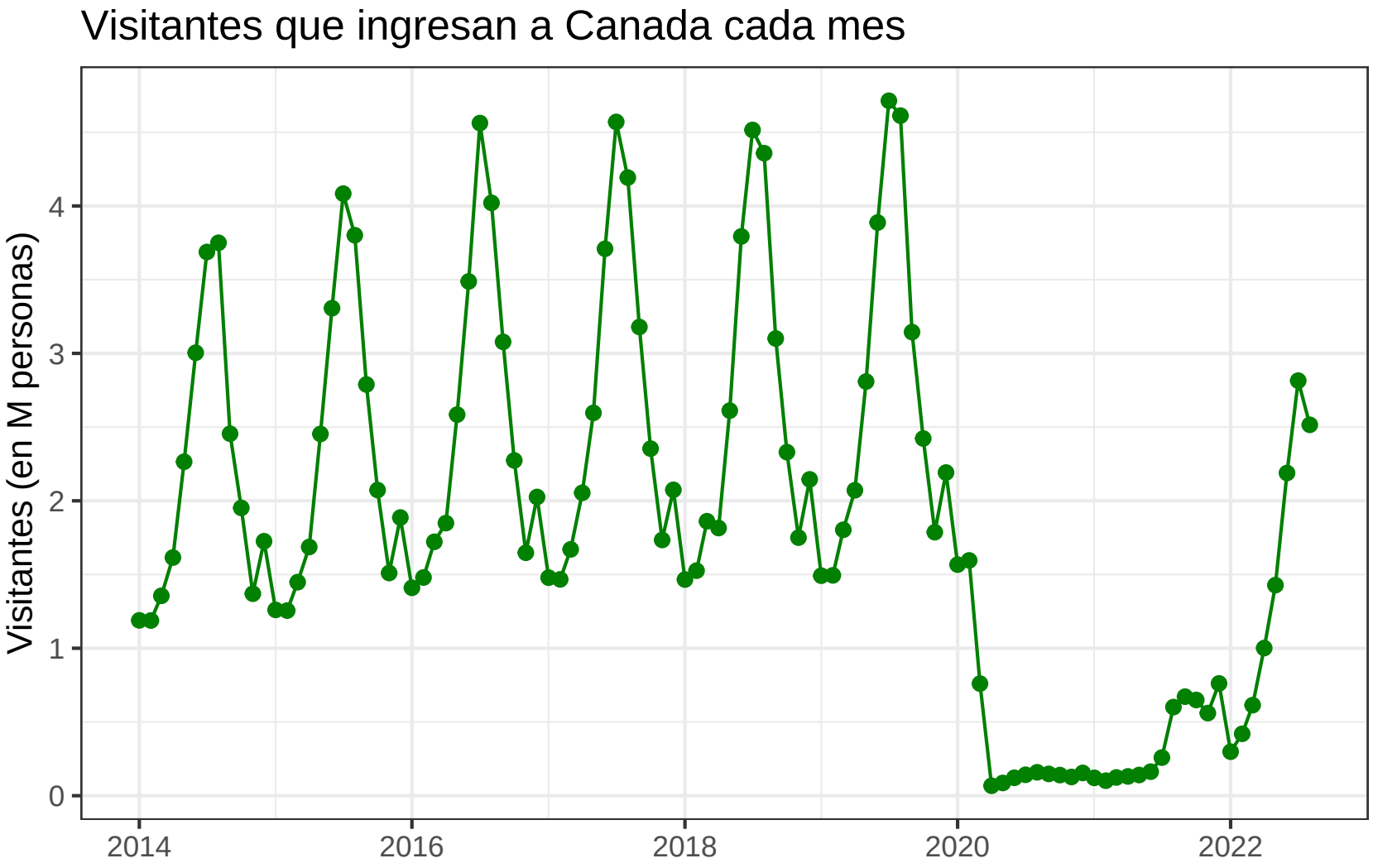
Usaremos los datos desde 1990 y pasar de escala a millones.

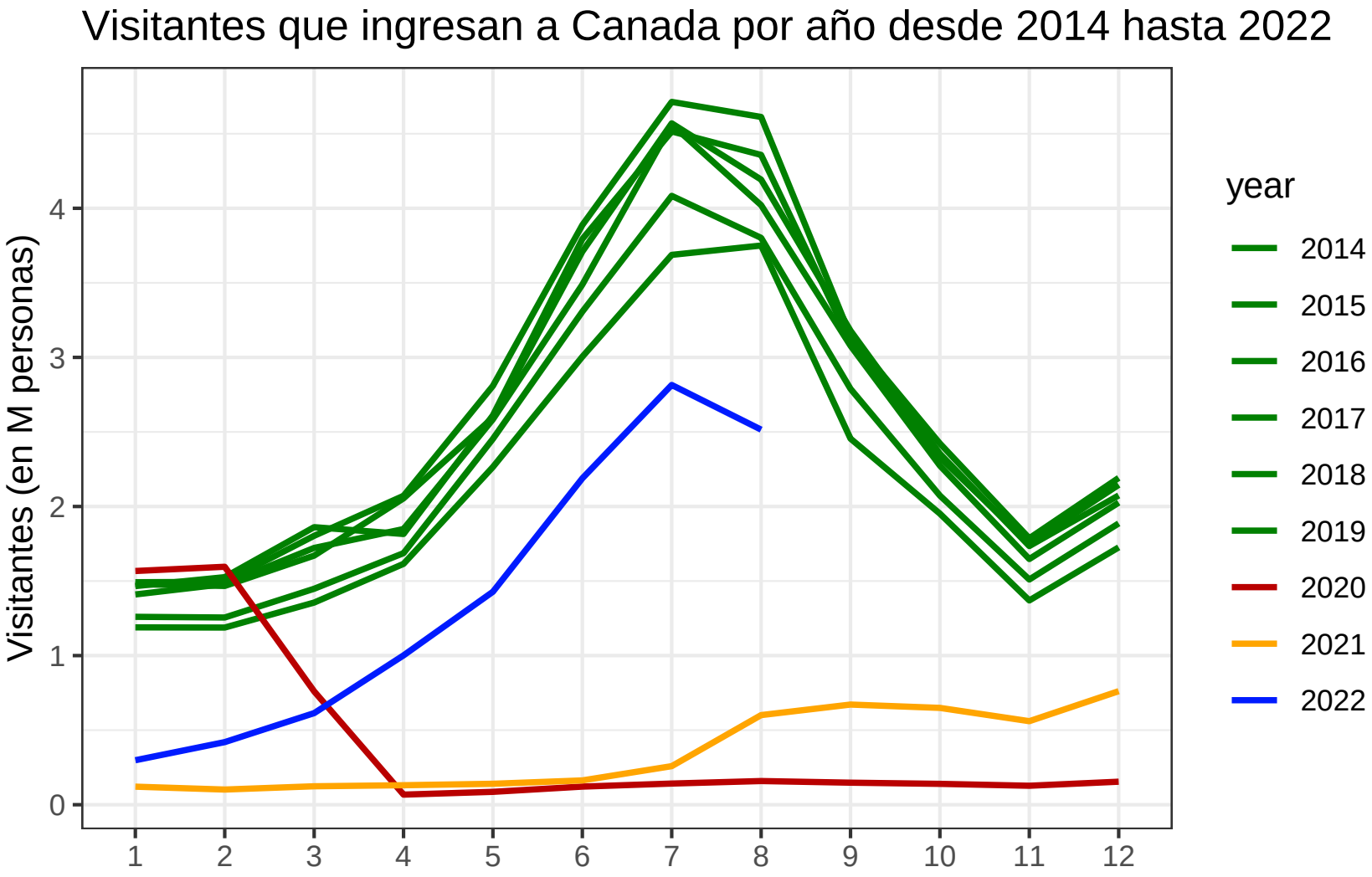
```
data <- data |> filter(date >= as.Date("1990/01/01"))

data$visitors<-data$visitors/1000000

head(data)
```

```
## # A tibble: 6 x 2
##   date       visitors
##   <date>        <dbl>
## 1 1990-01-01     1.85
## 2 1990-02-01     1.72
## 3 1990-03-01     2.22
## 4 1990-04-01     2.47
## 5 1990-05-01     3.24
## 6 1990-06-01     4.28
```

Analisis descriptivo




Visitantes que ingresan a Canada cada mes
Visitantes (en M personas)
0
1
2
3
4
2014
2016
2018
2020
2022


Visitantes que ingresan a Canada por año desde 2014 hasta 2022
Visitantes (en M personas)
0
1
2
3
4
1
2
3
4
5
6
7
8
9
10
11
12
year
2014
2015
2016
2017
2018
2019
2020
2021
2022

Section 3

Modelo a utilizar y alternativas.

Modelo a utilizar y alternativas.

Como se pudo ver nuestra serie no se podria modelar con un SARIMA por su forma y la caida abrupta del turismo que ocurrió en 2020. Por lo que intentamos identificar todos los outliers.

Modelo a utilizar y alternativas.

Como se pudo ver nuestra serie no se podria modelar con un SARIMA por su forma y la caida abrupta del turismo que ocurrió en 2020. Por lo que intentamos identificar todos los outliers.

Con la función tso obtuvimos los outliers de nuestra serie, pero pensamos que los puntos atípicos estaban mal ubicados por lo que se decidió ubicar unos nuevos y comparar.

Section 4

Resultados del ajuste, diagnóstico y medidas remediales para obtener un mejor ajuste.

Resultados del ajuste, diagnóstico y medidas remediales para obtener un mejor ajuste.

```
ts.no.val <- ts(
    data$visitors[1:(nrow(data) - 5)],
    start = c(1990, 1),
    frequency = 12
  )
```

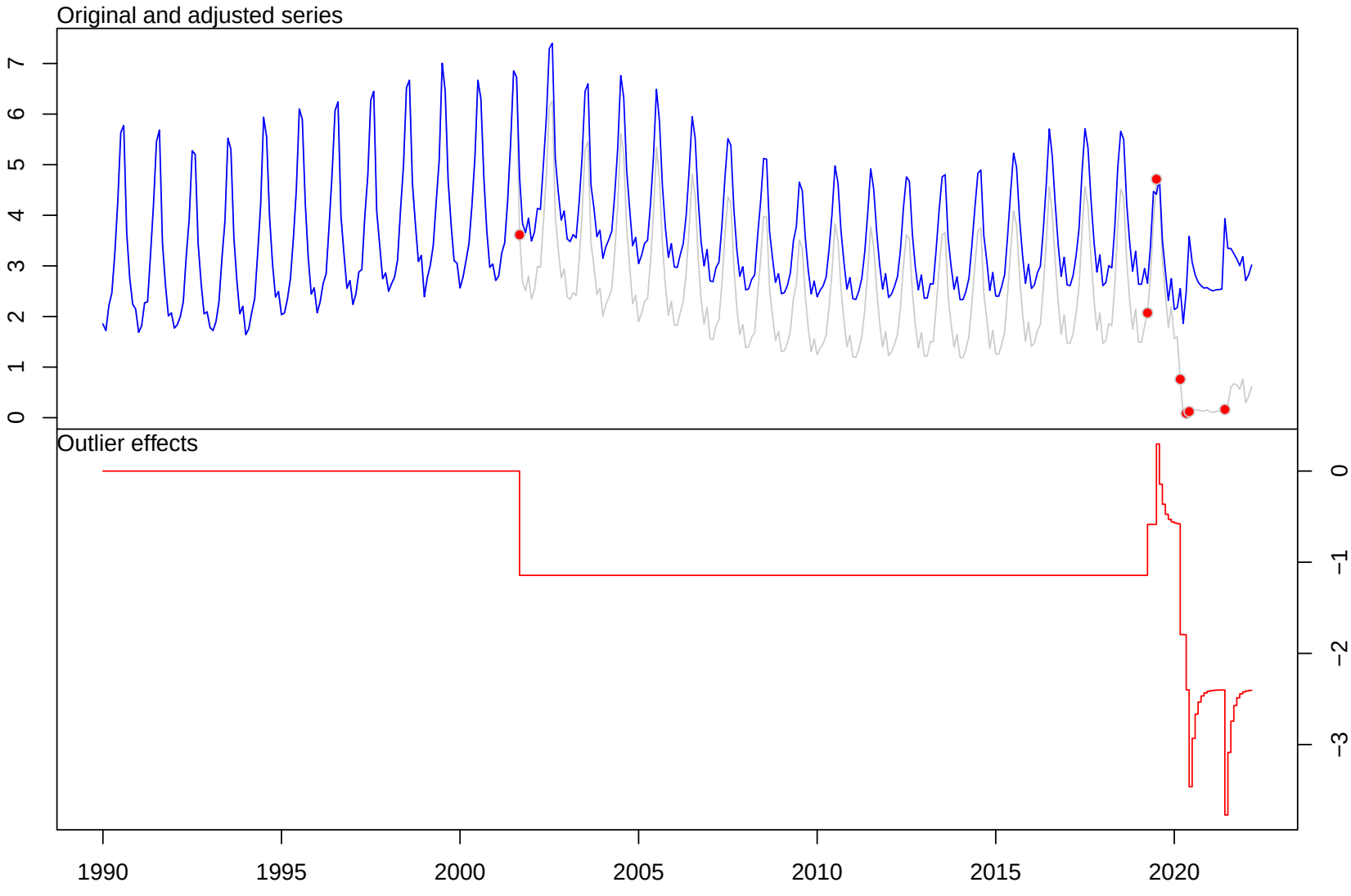
Datos atípicos

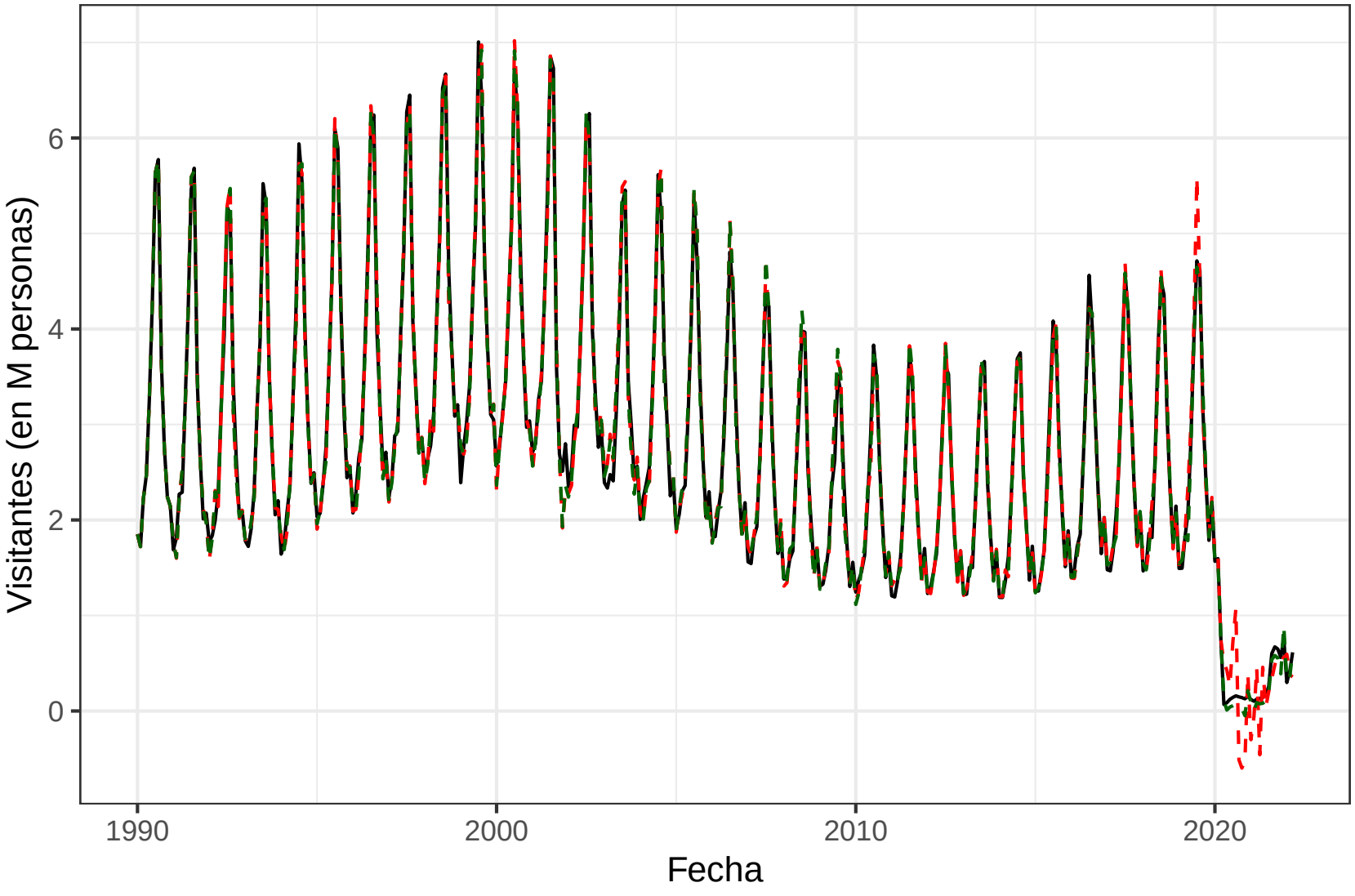
```
cov.df <- data.frame(
  ls.sep.2001 = as.integer(seq_along(len.total) >= 141),
  ls.mar.2020 = as.integer(seq_along(len.total) >= 363),
  ls.abr.2020 = as.integer(seq_along(len.total) >= 364),
  ao.may.2020 = as.integer(seq_along(len.total) == 365),
  ao.jun.2020 = as.integer(seq_along(len.total) == 366),
  ao.jul.2020 = as.integer(seq_along(len.total) == 367),
  ao.ago.2020 = as.integer(seq_along(len.total) == 368),
  ao.sep.2020 = as.integer(seq_along(len.total) == 369),
  ao.oct.2020 = as.integer(seq_along(len.total) == 370),
  ao.nov.2020 = as.integer(seq_along(len.total) == 371),
  ao.dic.2020 = as.integer(seq_along(len.total) == 372),
  ao.ene.2021 = as.integer(seq_along(len.total) == 373),
  ao.feb.2021 = as.integer(seq_along(len.total) == 374),
  ao.mar.2021 = as.integer(seq_along(len.total) == 375),
  ao.abr.2021 = as.integer(seq_along(len.total) == 376),
  ao.may.2021 = as.integer(seq_along(len.total) == 377),
  ao.jun.2021 = as.integer(seq_along(len.total) == 378),
  ao.jul.2021 = as.integer(seq_along(len.total) == 379),
  tc.ago.2021 = tc.ago.2021
  )
```

```
# Usando las covariables establecidas por nosotros
model.arima <-
  auto.arima(ts.no.val, seasonal = TRUE,
             xreg = as.matrix(cov.df[1:387,]))
```

```
# Usando la libreria tso para encontrar los outliers
model.tso <- tso(
  ts.no.val,
  delta = 0.5, # Se usa el delta que minimiza
               # el AIC encontrado usando un grid
  types = c("AO", "LS", "TC")
);

xreg.tso <- outliers.effects(model.tso$outliers, 392 + 12)
xreg.tso
```

```
##         LS141 LS352        TC355 LS363 LS365        TC366
##    [1,]     0     0 0.000000e+00     0     0 0.000000e+00
##    [2,]     0     0 0.000000e+00     0     0 0.000000e+00
##    [3,]     0     0 0.000000e+00     0     0 0.000000e+00
##    [4,]     0     0 0.000000e+00     0     0 0.000000e+00
##    [5,]     0     0 0.000000e+00     0     0 0.000000e+00
##    [6,]     0     0 0.000000e+00     0     0 0.000000e+00
```


Original and adjusted series
0
1
2
3
4
5
6
7
Outlier effects
0
−1
−2
−3
1990
1995
2000
2005
2010
2015
2020


Visitantes (en M personas)
6
4
2
0
1990
2000
2010
2020
Fecha

Section 5

Supuestos de los modelos.

Modelo Sarimax

```
Series: ts.no.val
Regression with ARIMA(2,0,0)(0,1,2)[12] errors

Coefficients:
         ar1     ar2     sma1     sma2  ls.sep.2001  ls.mar.2020  ls.abr.2020  ao.may.2020  ao.jun.2020  ao.jul.2020  ao.ago.2020
      0.5790  0.3457  -0.2164  -0.2497      -1.2322      -1.0809      -0.9576      -0.7606      -1.8106      -2.6870      -2.5562
s.e.  0.0518  0.0510   0.0522   0.0533       0.1047       0.1309       0.1529       0.1482       0.1596       0.1782       0.1887
      ao.sep.2020  ao.oct.2020  ao.nov.2020  ao.dic.2020  ao.ene.2021  ao.feb.2021  ao.mar.2021  ao.abr.2021  ao.may.2021  ao.jun.2021
          -1.2803      -0.6201      -0.1291      -0.3459       0.1726       0.0686      -0.1427      -0.3630      -1.1388      -2.2133
s.e.       0.1728       0.1715       0.1737       0.1762       0.1780       0.1791       0.1858       0.1911       0.2164       0.2241
      ao.jul.2021  tc.ago.2021
          -3.0098      -2.5330
s.e.       0.2202       0.2227

sigma^2 = 0.02083:  log likelihood = 203.7
AIC=-359.4   AICc=-355.97   BIC=-265.15
```

Figure 3: Modelo Sarimax

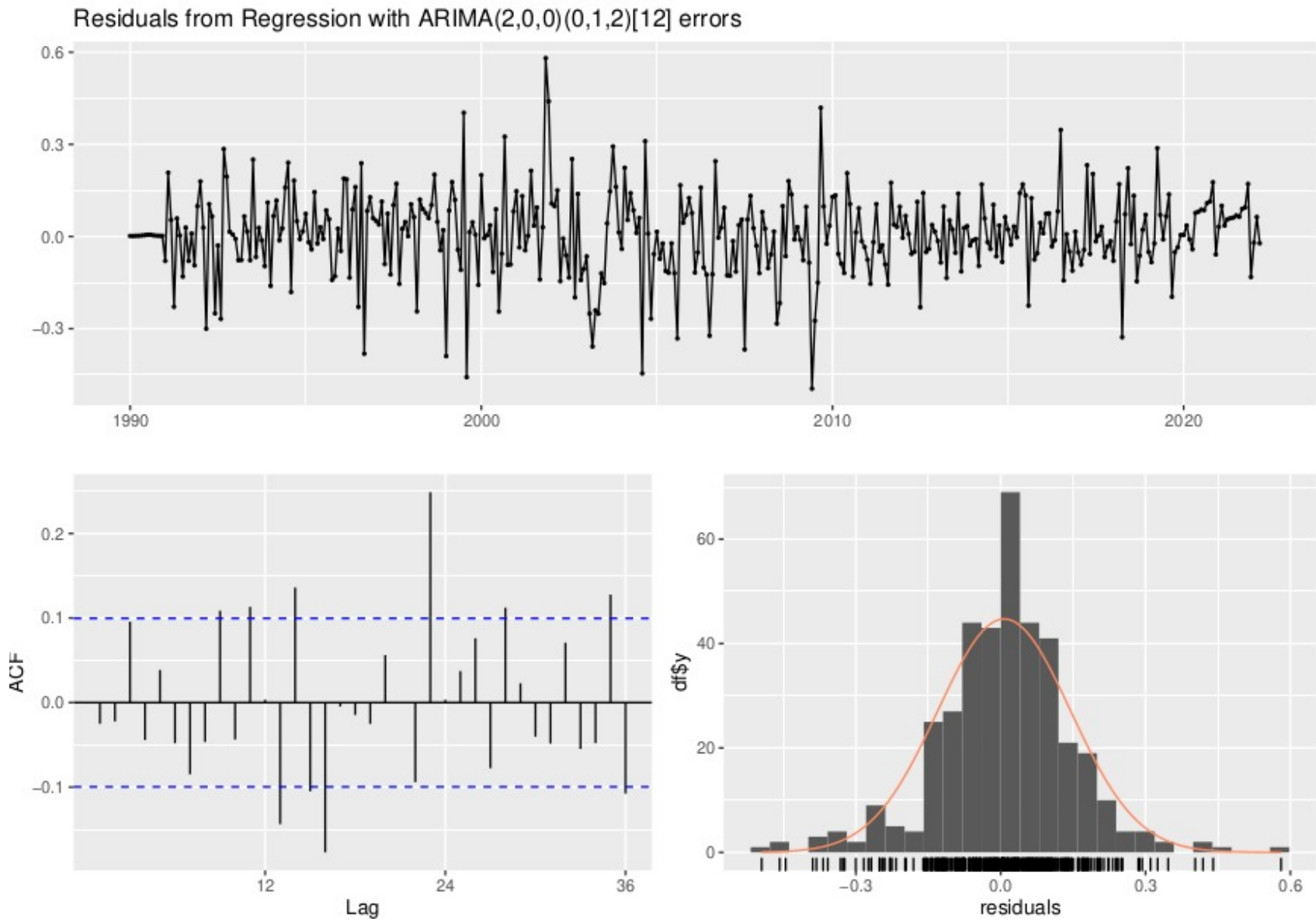
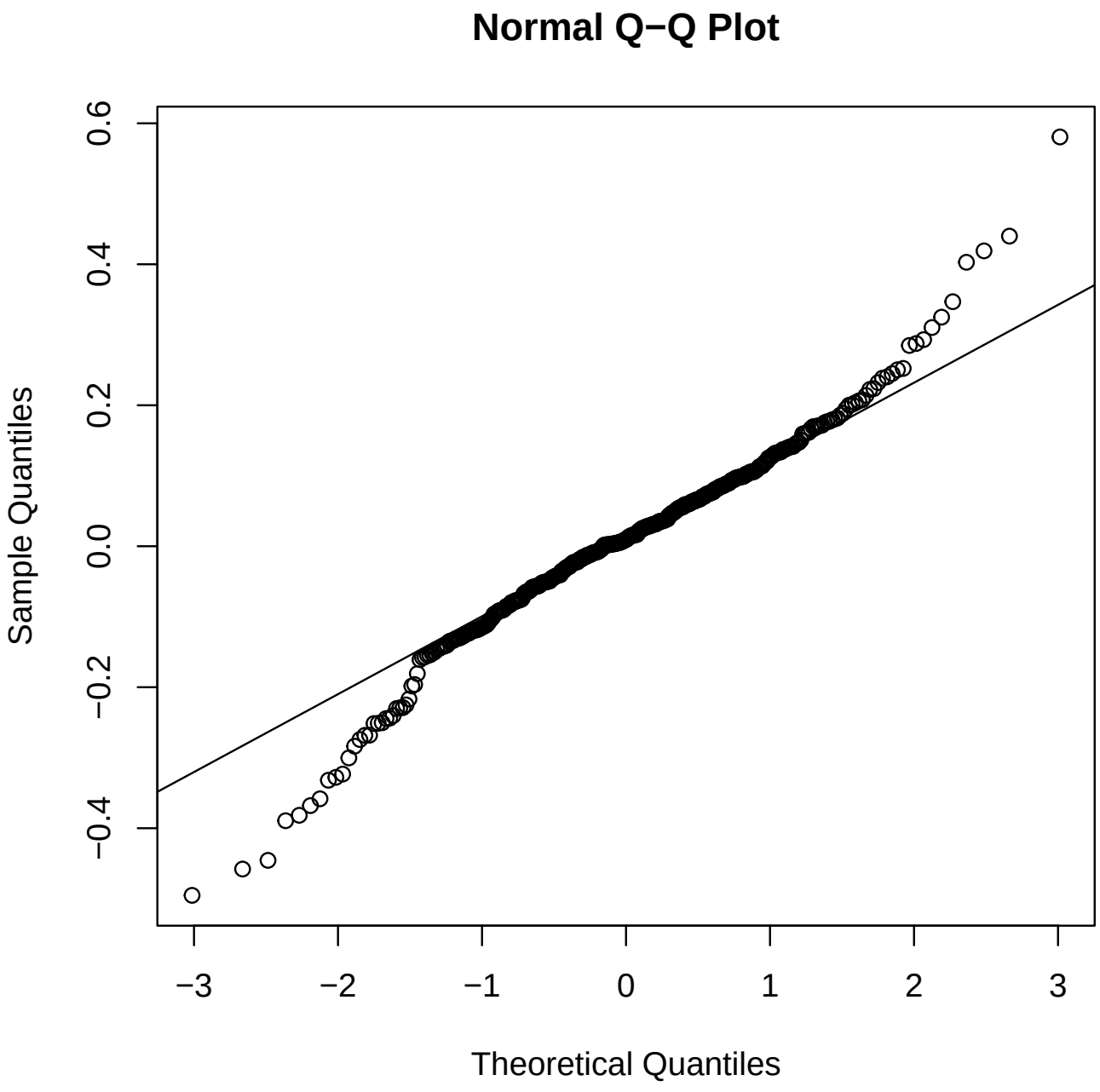


Figure 4: checkresidual Modelo Sarimax

```
        Ljung-Box test

data:  Residuals from Regression with ARIMA(2,0,0)(0,1,2)
[12] errors
Q* = 84.527, df = 20, p-value = 6.603e-10

Model df: 4.   Total lags used: 24
```

Figure 5: Ljung-Box test Modelo Sarimax


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
0.6
0.4
0.2
0.0
−0.2
−0.4
−3
−2
−1
0
1
2
3

```
##
##  Jarque Bera Test
##
## data:  model.arima$residuals
## X-squared = 60.278, df = 2, p-value = 8.149e-14

##
##  Shapiro-Wilk normality test
##
## data:  model.arima$residuals
## W = 0.97249, p-value = 1.069e-06
```

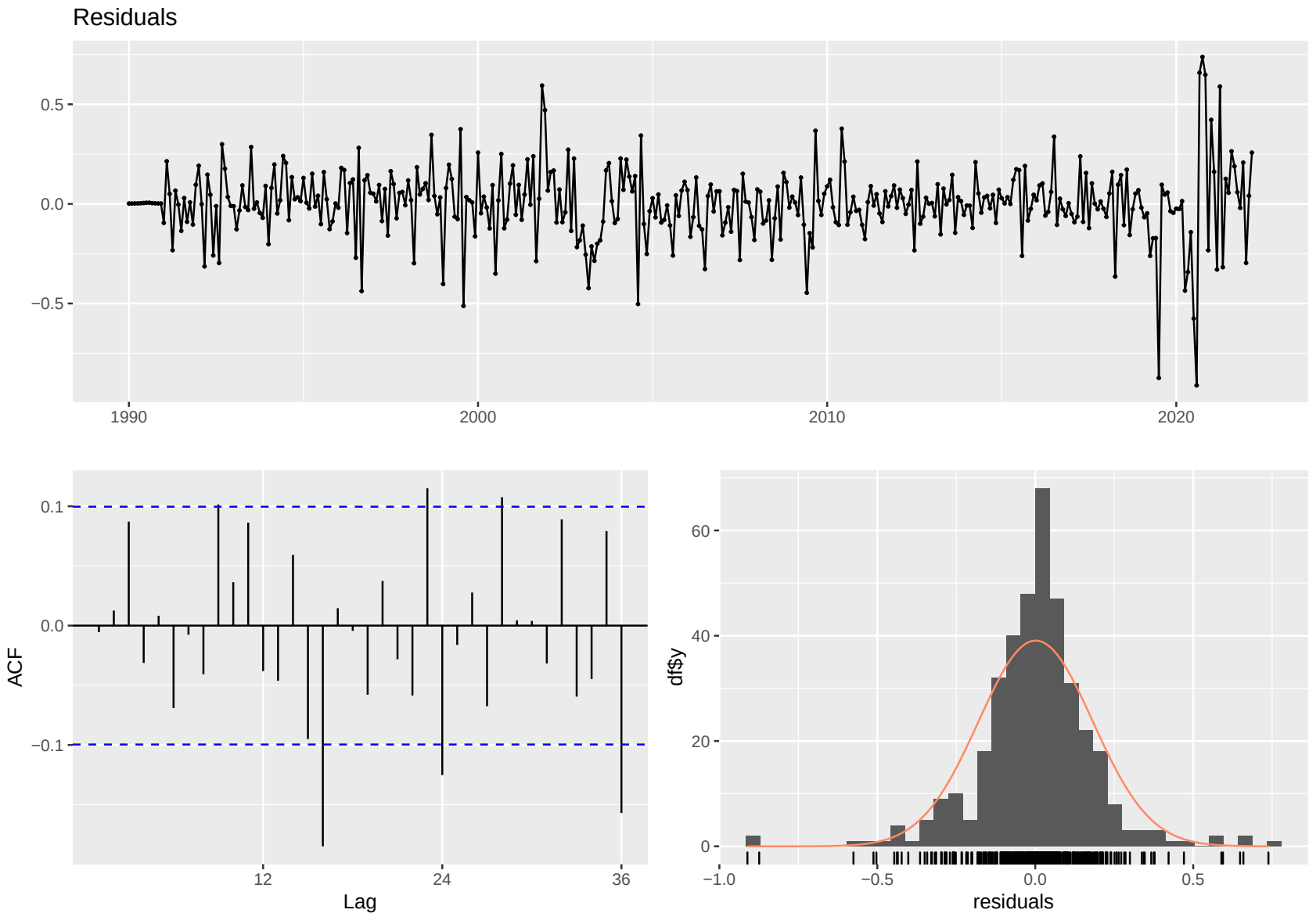
Modelo tso

```
Series: ts.no.val
Regression with ARIMA(2,0,0)(0,1,0)[12] errors

Coefficients:
          ar1     ar2    LS141   LS352   TC355    LS363
      0.5690  0.2886  -1.1440  0.5592  0.8820  -1.2118
s.e.  0.0544  0.0531   0.1287  0.1318  0.1249   0.1298
        LS365    TC366    TC378
      -0.6041  -1.0630  -1.3714
s.e.   0.1342   0.1764   0.2483

sigma^2 = 0.03456:  log likelihood = 102.78
AIC=-185.56   AICc=-184.95   BIC=-146.29
```

Figure 6: Modelo tso

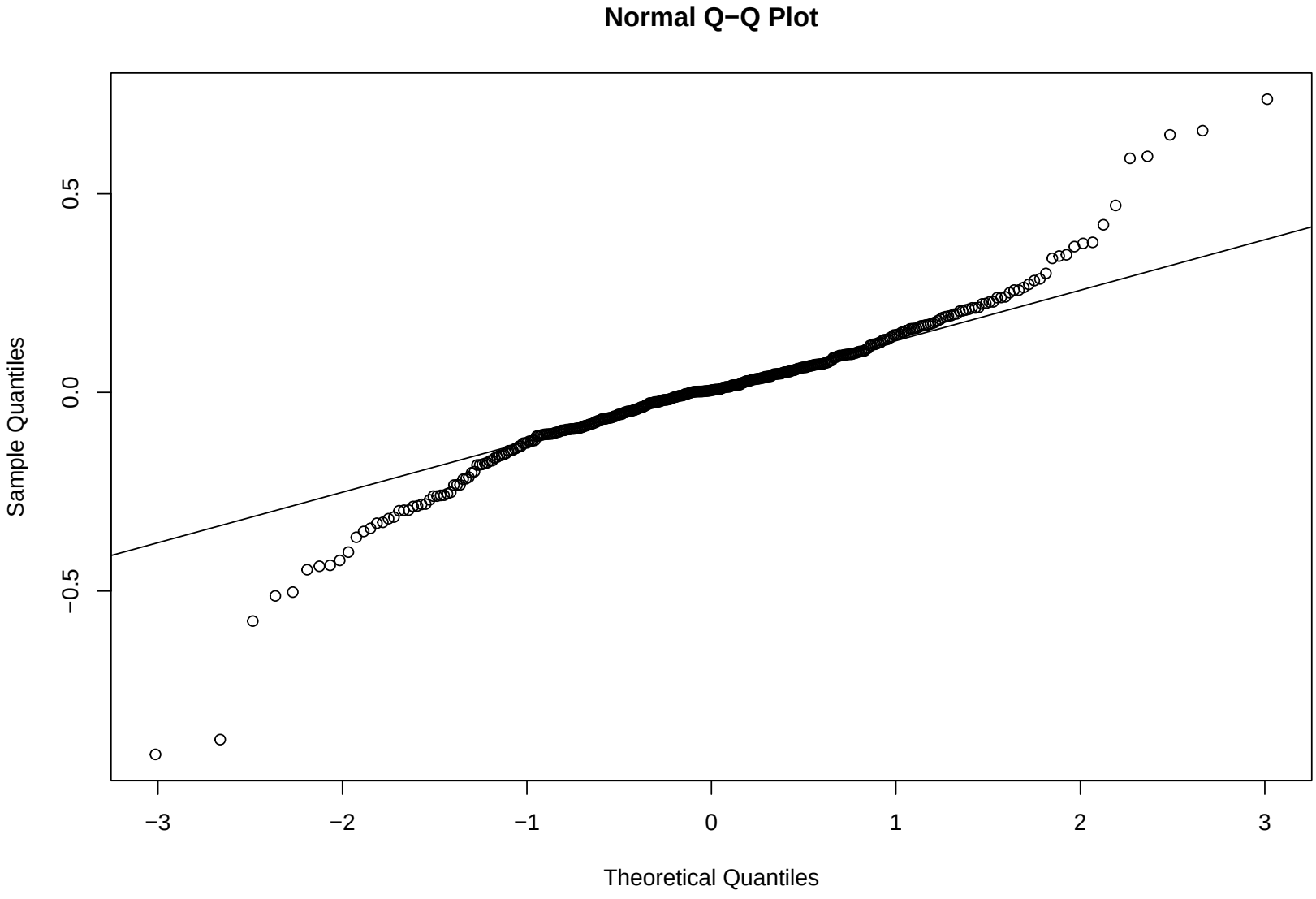
	type <S3: AsIs>	ind <S3: AsIs>	time <S3: AsIs>	coefhat <S3: AsIs>	tstat <S3: AsIs>
1	LS	141	2001:09	-1.1440	-8.892
2	LS	352	2019:04	0.5592	4.242
3	TC	355	2019:07	0.8820	7.059
4	LS	363	2020:03	-1.2118	-9.336
5	LS	365	2020:05	-0.6041	-4.501
6	TC	366	2020:06	-1.0630	-6.027
7	TC	378	2021:06	-1.3714	-5.523

Figure 7: Outliers Modelo tso


Residuals
0.5
0.0
−0.5
1990
2000
2010
2020
0.1
0.0
−0.1
ACF
12
24
36
Lag
60
40
20
0
df$y
−1.0
−0.5
0.0
0.5
residuals

```
        Ljung-Box test

data:  Residuals from Regression with ARIMA(2,0,0)(0,1,0)[12] errors
Q* = 49.783, df = 22, p-value = 0.0006274

Model df: 2.   Total lags used: 24
```

Figure 8: Ljung-Box Test modelo tso


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
0.5
0.0
−0.5
−3
−2
−1
0
1
2
3

```
##
##  Jarque Bera Test
##
## data:  model.tso$fit$residuals
## X-squared = 330.9, df = 2, p-value < 2.2e-16

##
##  Shapiro-Wilk normality test
##
## data:  model.tso$fit$residuals
## W = 0.93133, p-value = 2.367e-12
```

Section 6

Realización y validación de las predicciones.

Realización y validación de las predicciones.

```
# MSE para los ultimos 5 datos no
# considerados en el entrenamiento
paste("MSE manual: ", mean((data[388:392, ]$visitors - prec
```

```
## [1] "MSE manual:  0.269935869786739"
```

```
paste("MSE tso: ", mean((data[388:392, ]$visitors - pred.ts
```

```
## [1] "MSE tso:  1.14155438076962"
```

```
model.arima$coef |> length()
```

```
## [1] 23
```

```
model.tso$fit$coef |> length()
```

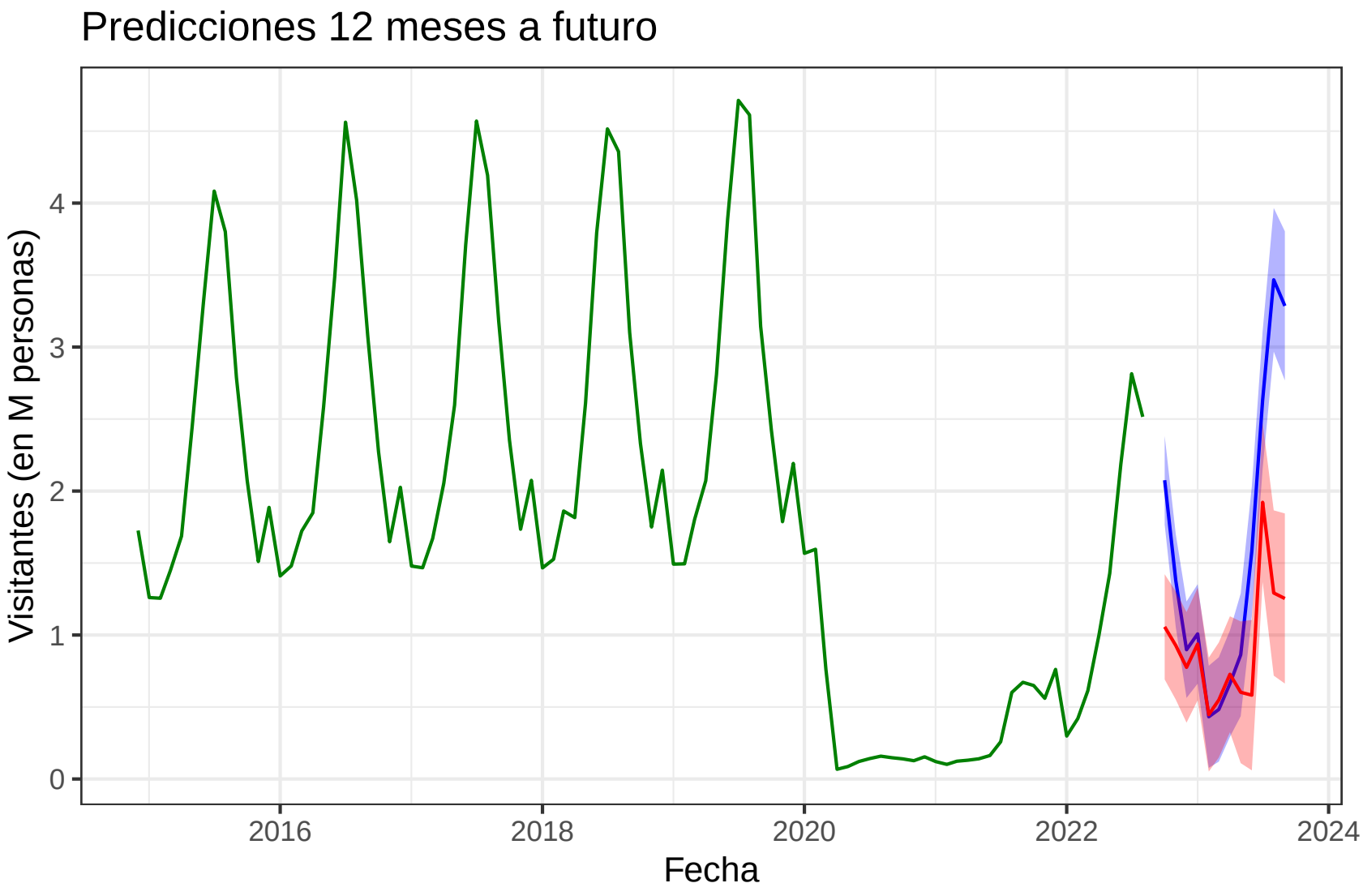
```
## [1] 9
```

```
model.arima$bic
```

```
## [1] -265.1544
```

```
model.tso$fit$bic
```

```
## [1] -146.2887
```


Predicciones 12 meses a futuro
Visitantes (en M personas)
0
1
2
3
4
2016
2018
2020
2022
2024
Fecha

Section 7

Conclusiones y recomendaciones.

Conclusiones y recomendaciones.

En busca de un modelo que nos pueda ayudar a hacer predicciones sobre los visitantes que entran a Canadá se obtienen dos modelos, los cuales se encuentran teniendo en cuenta los datos atípicos e intervenciones.

Conclusiones y recomendaciones.

En busca de un modelo que nos pueda ayudar a hacer predicciones sobre los visitantes que entran a Canadá se obtienen dos modelos, los cuales se encuentran teniendo en cuenta los datos atípicos e intervenciones.

Se obtiene dos modelos, uno con la función **tso**, la cual intenta encontrar los puntos atípicos automáticamente y el otro en el que asignamos manualmente los puntos atípicos y los usamos como covariables en el modelo.

Conclusiones y recomendaciones.

En busca de un modelo que nos pueda ayudar a hacer predicciones sobre los visitantes que entran a Canadá se obtienen dos modelos, los cuales se encuentran teniendo en cuenta los datos atípicos e intervenciones.

Se obtiene dos modelos, uno con la función **tso**, la cual intenta encontrar los puntos atípicos automáticamente y el otro en el que asignamos manualmente los puntos atípicos y los usamos como covariables en el modelo.

Se observa que a pesar de que ambos modelos no cumplen de los supuestos necesarios el modelo en el que asignamos manualmente los puntos atípicos tiene mejor MSE de prueba, AIC y BIC.

Section 8

Citaciones

Citaciones

- Statistics Canada. Table 24-10-0050-01 Non-resident visitors entering Canada, by country of residence. **DOI:** https://doi.org/10.25318/2410005001-eng