

PEC 3: Desing and Implementation

UOC - Alumno: Álvaro Rodríguez Sans

May 2020 - Delivery 23/05/2021

Índex

1 Data load	6
2 Initial descriptive statistics and visualization	7
2.1 Data types and modifications	7
2.1.1 EM3 review	7
2.1.2 EM3 data transformation	8
2.1.3 EM3 transpose and dates missing generation	8
2.1.4 EM3 review missing values & impute	9
2.1.5 Google review	24
2.1.6 Google autonomous-communities & provinces	27
2.1.7 Google data transformation	30
2.1.8 Google review missing values & impute	33
2.1.9 CNE review	68
2.1.10 CNE review missing values & impute	69
2.1.11 CNE data transformation	77
2.2 Datasets combinations	79
2.2.1 CNE_tec_cas	79
2.2.2 GOG_CNE	80
2.2.3 Total GOG_CNE + EM3	81
2.3 Visual analysis	91
2.3.1 Dataframe plots (Málaga, Sevilla and Cádiz)	91
2.3.2 Time-series plots (Barcelona, Madrid, Málaga, Sevilla and Cádiz)	94
2.3.3 Correlation plots (from dataframe)	99
2.3.4 PCA (Barcelona)	105
2.3.5 Review normality (Barcelona)	113
2.3.6 Final plots (Barcelona and others)	117
3 ARIMA - fpp3 library	129
3.1 ACF and PACF (Barcelona, Madrid, Málaga, Córdoba and Cádiz)	129
3.2 STL (Seasonal and Trend decomposition using Loess - Barcelona, Madrid, Málaga, Cádiz and Sevilla)	139
3.3 Model and Forecast (Barcelona, Madrid, Málaga, Sevilla and Cádiz)	189
3.3.1 Univariate (7, 14, 17 days) Barcelona, Madrid, Málaga, Sevilla and Cádiz	189
3.3.1.1 Barcelona	190
3.3.1.2 Madrid	200
3.3.1.3 Málaga	210
3.3.1.4 Cádiz	220
3.3.1.5 Sevilla	230
3.3.2 Multivariate (7, 14, 17 days) + residential_percent_change (Google) + Total (INE - EM3)	240

3.3.2.1	Barcelona	240
3.3.2.2	Madrid	249
3.3.2.3	Málaga	257
3.3.2.4	Cádiz	266
3.3.2.5	Sevilla	275
3.3.3	Multivariate (7, 14, 17 days) + All mobility	284
3.3.3.1	Barcelona	285
3.3.3.2	Málaga	296

Bibliography

306

This is an R Markdown Notebook. When you execute code within the notebook, the results appear beneath the code. Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Ctrl+Shift+Enter*.

Add a new chunk by clicking the *Insert Chunk* button on the toolbar or by pressing *Ctrl+Alt+I*. When you save the notebook, an HTML file containing the code and output will be saved alongside it (click the *Preview* button or press *Ctrl+Shift+K* to preview the HTML file).

The preview shows you a rendered HTML copy of the contents of the editor. Consequently, unlike *Knit*, *Preview* does not run any R code chunks. Instead, the output of the chunk when it was last run in the editor is displayed.

The bibliographic references used for this practice have been: (Baayen 2008; Hothorn and Everitt 2014; Hyndman and Athanasopoulos 2021; Liviano Solas and Pujol Jover nd; Teator 2011; Vegas Lozano nd).

```
# At section - Data types and modifications
if(!require(knitr)){
  install.packages('knitr', repos='http://cran.us.r-project.org')
  library(knitr)}

## Loading required package: knitr
if(!require(latexpdf)){
  install.packages('latexpdf', repos='http://cran.us.r-project.org')
  library(latexpdf)}

## Loading required package: latexpdf
if(!require(latex2exp)){
  install.packages('latex2exp', repos='http://cran.us.r-project.org')
  library(latex2exp)}

## Loading required package: latex2exp
if(!require(recipes)){
  install.packages('recipes', repos='http://cran.us.r-project.org')
  library(recipes)}

## Loading required package: recipes
## Loading required package: dplyr
## Registered S3 methods overwritten by 'tibble':
##   method      from
##   format.tbl pillar
##   print.tbl  pillar
##
## Attaching package: 'dplyr'
```

```

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

##
## Attaching package: 'recipes'

## The following object is masked from 'package:stats':
##
##     step

if(!require(timetk)){
  install.packages('timetk', repos='http://cran.us.r-project.org')
  library(timetk)}

## Loading required package: timetk

if(!require(data.table)){
  install.packages('data.table', repos='http://cran.us.r-project.org')
  library(data.table)}

## Loading required package: data.table

##
## Attaching package: 'data.table'

## The following object is masked from 'package:timetk':
##
##     :=
## The following objects are masked from 'package:dplyr':
##
##     between, first, last

if(!require(tidyverse)){
  install.packages("tidyverse", repos='http://cran.us.r-project.org')
  library(tidyverse)}

## Loading required package: tidyverse

## -- Attaching packages ----- tidyverse 1.3.0 --
## v ggplot2 3.3.3     v purrr   0.3.3
## v tibble   3.0.0     v stringr 1.4.0
## v tidyverse 1.3.0    vforcats 0.5.0
## v readr    1.3.1

## -- Conflicts ----- tidyverse_conflicts() --
## x data.table::between() masks dplyr::between()
## x dplyr::filter()      masks stats::filter()
## x data.table::first()   masks dplyr::first()
## x stringr::fixed()     masks recipes::fixed()
## x dplyr::lag()         masks stats::lag()
## x data.table::last()    masks dplyr::last()
## x purrr::transpose()   masks data.table::transpose()

```

```

if(!require(VIM)){
  install.packages('VIM', repos='http://cran.us.r-project.org')
  library(VIM)}

## Loading required package: VIM
## Loading required package: colorspace
## Loading required package: grid
## VIM is ready to use.
## Since version 4.0.0 the GUI is in its own package VIMGUI.
##
##           Please use the package to use the new (and old) GUI.
## Suggestions and bug-reports can be submitted at: https://github.com/alexkowa/VIM/issues
##
## Attaching package: 'VIM'

## The following object is masked from 'package:recipes':
##
##     prepare

## The following object is masked from 'package:datasets':
##
##     sleep

if(!require(imputeTS)){
  install.packages("imputeTS", repos='http://cran.us.r-project.org')
  library(imputeTS)}

## Loading required package: imputeTS
## Registered S3 method overwritten by 'quantmod':
##   method           from
##   as.zoo.data.frame zoo

if(!require(xts)){
  install.packages("xts", repos='http://cran.us.r-project.org')
  library(xts)}

## Loading required package: xts
## Loading required package: zoo
##
## Attaching package: 'zoo'

## The following object is masked from 'package:imputeTS':
##
##     na.locf

## The following objects are masked from 'package:base':
##   as.Date, as.Date.numeric
##
## Attaching package: 'xts'

## The following objects are masked from 'package:data.table':
##

```

```

##      first, last
## The following objects are masked from 'package:dplyr':
##
##      first, last
if(!require(tsbox)){
  install.packages("tsbox", repos='http://cran.us.r-project.org')
  library(tsbox)}

## Loading required package: tsbox
# At section - Visual analysis
if(!require(fpp3)){
  install.packages("fpp3", repos='http://cran.us.r-project.org')
  library(fpp3)}

## Loading required package: fpp3
## -- Attaching packages ----- fpp3 0.4.0 --
## v lubridate     1.7.8      v feasts       0.2.1
## v tsibble       1.0.0      v fable        0.3.0
## v tsibbledata   0.3.0

## -- Conflicts ----- fpp3_conflicts --
## x data.table::between() masks dplyr::between()
## x lubridate::date()    masks base::date()
## x dplyr::filter()     masks stats::filter()
## x xts::first()        masks data.table::first(), dplyr::first()
## x lubridate::hour()   masks data.table::hour()
## x tsibble::index()   masks zoo::index()
## x tsibble::intersect() masks base::intersect()
## x tsibble::interval() masks lubridate::interval()
## x lubridate::isoweek() masks data.table::isoweek()
## x tsibble::key()     masks data.table::key()
## x dplyr::lag()        masks stats::lag()
## x xts::last()         masks data.table::last(), dplyr::last()
## x lubridate::mday()   masks data.table::mday()
## x lubridate::minute() masks data.table::minute()
## x lubridate::month()  masks data.table::month()
## x lubridate::quarter() masks data.table::quarter()
## x lubridate::second() masks data.table::second()
## x tsibble::setdiff()  masks base::setdiff()
## x purrr::transpose()  masks data.table::transpose()
## x tsibble::union()   masks base::union()
## x lubridate::wday()   masks data.table::wday()
## x lubridate::week()   masks data.table::week()
## x lubridate::yday()   masks data.table::yday()
## x lubridate::year()   masks data.table::year()

if(!require(corrplot)){
  install.packages('corrplot', repos='http://cran.us.r-project.org')
  library(corrplot)}

## Loading required package: corrplot
## corrplot 0.84 loaded

```

```

if(!require(DescTools)){
  install.packages("DescTools", repos='http://cran.us.r-project.org')
  library(DescTools)}

## Loading required package: DescTools

##
## Attaching package: 'DescTools'

## The following objects are masked from 'package:fabletools':
## 
##     MAE, MAPE, MSE, RMSE

## The following object is masked from 'package:data.table':
## 
##     %like%

# At LSTM section

knitr::opts_chunk$set(echo = TRUE)

```

1 Data load

Data is loaded from the sources stated at PEC1 and PEC2 (CNE, INE and Google).

- CNE-Covid-19
- INE-Covid-19
- Google-Covid-19

```

#library(dplyr)
# Source INE
EM3 <- read.csv('EM3-Movimiento de personas por provincias.csv',
                 header=TRUE,
                 sep = ";",
                 stringsAsFactors = FALSE)

# Source Google
Google <- read.csv('Google-2020_ES_Region_Mobility_Report.csv',
                     header=TRUE,
                     sep = ";",
                     stringsAsFactors = FALSE)

# Source CNE (here sep is ",")
CNE_tecnica <- read.csv('CNE-casos_tecnica_provincia.csv',
                        header=TRUE,
                        sep = ",",
                        stringsAsFactors = FALSE)
CNE_casos <- read.csv('CNE-casos_hosp_uci_def_sexo_edad_provres.csv',
                      header=TRUE,
                      sep = ",",
                      stringsAsFactors = FALSE)

```

2 Initial descriptive statistics and visualization

2.1 Data types and modifications

We are going to check the **type of variable** that corresponds to each of the variables (numerical, factor, etc.) and **missing data / values or other anomalies** in each dataset.

2.1.1 EM3 review

Here we have the mobility of people by provinces (we can see 146 rows by province, that correspond to days). In order to facilitate the comparison, a valid reference date for the mobility of the population should be considered. The “normal” date for this study, has been considered as the one that results from the average of the days 18 (Monday) to 21 (Thursday) of November 2019. It is indicated in the tables as the reference date 18/11/2019.

```
# Source INE
summary(EM3)

##  Zonas.de.movilidad Periodo           Total
##  Length:9198          Length:9198      Length:9198
##  Class :character    Class :character  Class :character
##  Mode  :character    Mode  :character  Mode  :character

head(str(EM3,vec.len=2))

## 'data.frame':  9198 obs. of  3 variables:
## $ Zonas.de.movilidad: chr  "Almería" "Almería" ...
## $ Periodo            : chr  "30/12/2020" "27/12/2020" ...
## $ Total              : chr  "17,17" "11,53" ...
## NULL

table(EM3$Zonas.de.movilidad)

##          Albacete Alicante/Alacant Almería
##             146          146        146
##          Araba/Álava Asturias       Ávila
##             146          146        146
##          Badajoz   Balears, Illes Barcelona
##             146          146        146
##          Bizkaia      Burgos       Cáceres
##             146          146        146
##          Cádiz       Cantabria Castellón/Castelló
##             146          146        146
##          Ceuta       Ciudad Real Córdoba
##             146          146        146
##          Coruña, A     Cuenca      Formentera
##             146          146        146
##          Fuerteventura Gipuzkoa     Girona
##             146          146        146
##          Gomera, La    Gran Canaria Granada
##             146          146        146
##          Guadalajara   Hierro, El Huelva
##             146          146        146
##          Huesca        Ibiza       Jaén
##             146          146        146
##          Lanzarote      León       Lleida
```

##	146	146	146
##	Lugo	Madrid	Málaga
##	146	146	146
##	Mallorca	Melilla	Menorca
##	146	146	146
##	Murcia	Navarra	Ourense
##	146	146	146
##	Palencia	Palma, La	Palmas, Las
##	146	146	146
##	Pontevedra	Rioja, La	Salamanca
##	146	146	146
## Santa Cruz de Tenerife		Segovia	Sevilla
##	146	146	146
##	Soria	Tarragona	Tenerife
##	146	146	146
##	Teruel	Toledo	Valencia/València
##	146	146	146
##	Valladolid	Zamora	Zaragoza
##	146	146	146

2.1.2 EM3 data transformation

We are going to **transform**:

- “Total” from “character” to “numerical”
- “Periodo” from “character” to “date”

```
EM3$Total <- sub(", ", ".", EM3$Total)
EM3$Total <- as.numeric(EM3$Total)
EM3$Periodo <- as.Date(EM3$Periodo, format="%d/%m/%Y")
head(EM3)
```

```
##   Zonas.de.movilidad Periodo Total
## 1 Almería 2020-12-30 17.17
## 2 Almería 2020-12-27 11.53
## 3 Almería 2020-12-23 17.81
## 4 Almería 2020-12-20 12.13
## 5 Almería 2020-12-16 18.28
## 6 Almería 2020-12-13 11.97
```

2.1.3 EM3 transpose and dates missing generation

Due to the nature of this dataset we have to transpose it in order to analyze the missing values by province and impute them. There **are some dates that are not provided by EM3 study**.

```
#library(data.table)
# Transpose dataframe
EM3_t<-dcast(EM3, Periodo~Zonas.de.movilidad) #, fill=NA)

#library(tidyverse)
# Create dates missing (for time series).
# Note: According INE, some "dates" are not provided.
# We have to generate them
EM3_t<-EM3_t %>%
  complete(Periodo = seq.Date(min(Periodo), max(Periodo), by="day"))
```

```

# Filter the interest period according INE EM3 study
# "2019-11-18" is the reference date EM3 study (for us it is excluded)
EM3_t<- EM3_t %>%
  filter(Periodo <= "2019-11-18" | Periodo >= "2020-03-16")

EM3_t

## # A tibble: 291 x 64
##   Periodo    Albacete `Alicante/Alaca~ Almería `Araba/Álava` Asturias Ávila
##   <date>     <dbl>        <dbl>      <dbl>      <dbl>      <dbl>      <dbl>
## 1 2019-11-18  25.2       28.1      24.4      31.9      29.9      26.6
## 2 2020-03-16   9.9       14.4      11.0      15.9      13.1      9.44
## 3 2020-03-17   NA        NA        NA        NA        NA        NA
## 4 2020-03-18   9.51      13.4      7.28      14.5      12.0      9.17
## 5 2020-03-19   NA        NA        NA        NA        NA        NA
## 6 2020-03-20   8.75      12.0      6.87      11.9      11.3      8.69
## 7 2020-03-21   NA        NA        NA        NA        NA        NA
## 8 2020-03-22   4.5       6.14      4.19      6.46      5.64      4.53
## 9 2020-03-23   NA        NA        NA        NA        NA        NA
## 10 2020-03-24  9.02      10.9      8.98      13.3      11.2      8.26
## # ... with 281 more rows, and 57 more variables: Badajoz <dbl>, `Balears,
## # Illes` <dbl>, Barcelona <dbl>, Bizkaia <dbl>, Burgos <dbl>, Cáceres <dbl>,
## # Cádiz <dbl>, Cantabria <dbl>, `Castellón/Castelló` <dbl>, Ceuta <dbl>,
## # `Ciudad Real` <dbl>, Córdoba <dbl>, `Coruña, A` <dbl>, Cuenca <dbl>,
## # Formentera <dbl>, Fuerteventura <dbl>, Gipuzkoa <dbl>, Girona <dbl>,
## # `Gomera, La` <dbl>, `Gran Canaria` <dbl>, Granada <dbl>, Guadalajara <dbl>,
## # `Hierro, El` <dbl>, Huelva <dbl>, Huesca <dbl>, Ibiza <dbl>, Jaén <dbl>,
## # Lanzarote <dbl>, León <dbl>, Lleida <dbl>, Lugo <dbl>, Madrid <dbl>,
## # Málaga <dbl>, Mallorca <dbl>, Melilla <dbl>, Menorca <dbl>, Murcia <dbl>,
## # Navarra <dbl>, Ourense <dbl>, Palencia <dbl>, `Palma, La` <dbl>, `Palmas,
## # Las` <dbl>, Pontevedra <dbl>, `Rioja, La` <dbl>, Salamanca <dbl>, `Santa
## # Cruz de Tenerife` <dbl>, Segovia <dbl>, Sevilla <dbl>, Soria <dbl>,
## # Tarragona <dbl>, Tenerife <dbl>, Teruel <dbl>, Toledo <dbl>,
## # `Valencia/València` <dbl>, Valladolid <dbl>, Zamora <dbl>, Zaragoza <dbl>

```

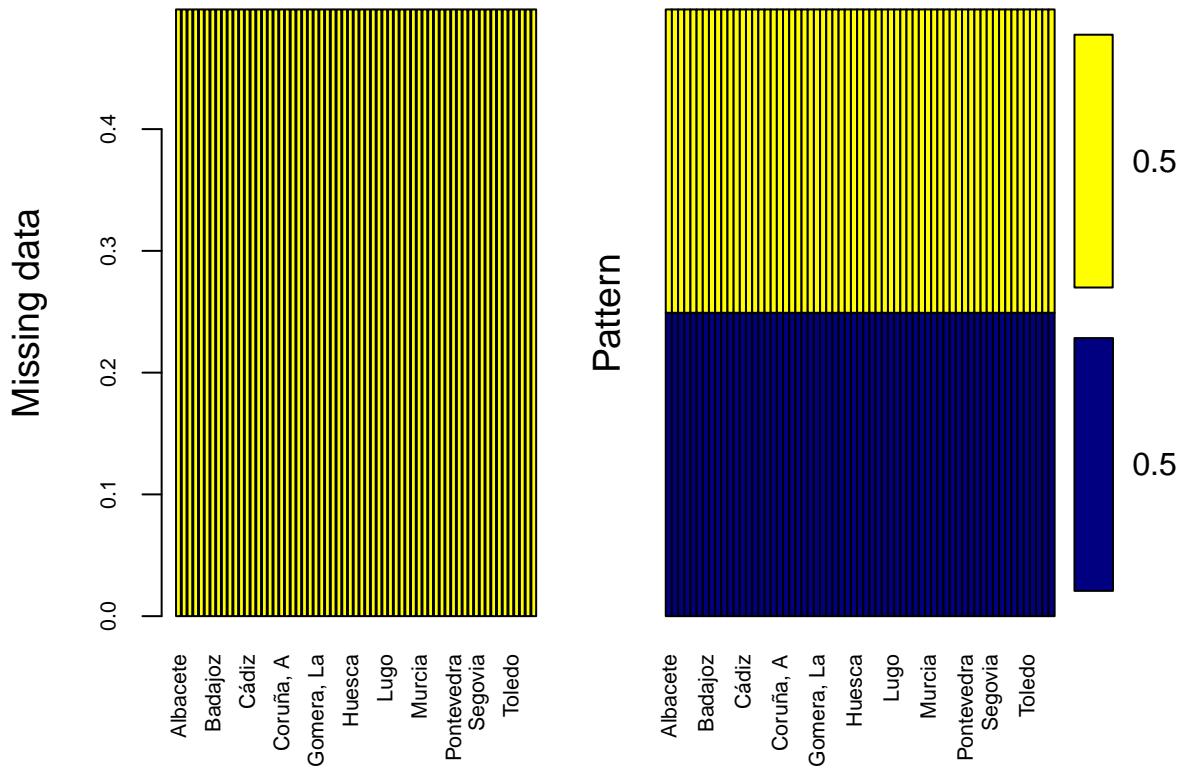
2.1.4 EM3 review missing values & impute

We check the missing values by province (we are close to have 150 by province).

```

library(VIM)
aggr(EM3_t[,-1], col=c('navyblue','yellow'),
  numbers=TRUE, sortVars=TRUE,
  labels=names(EM3_t[,-1]), cex.axis=.7,
  gap=3, ylab=c("Missing data", "Pattern"))

```



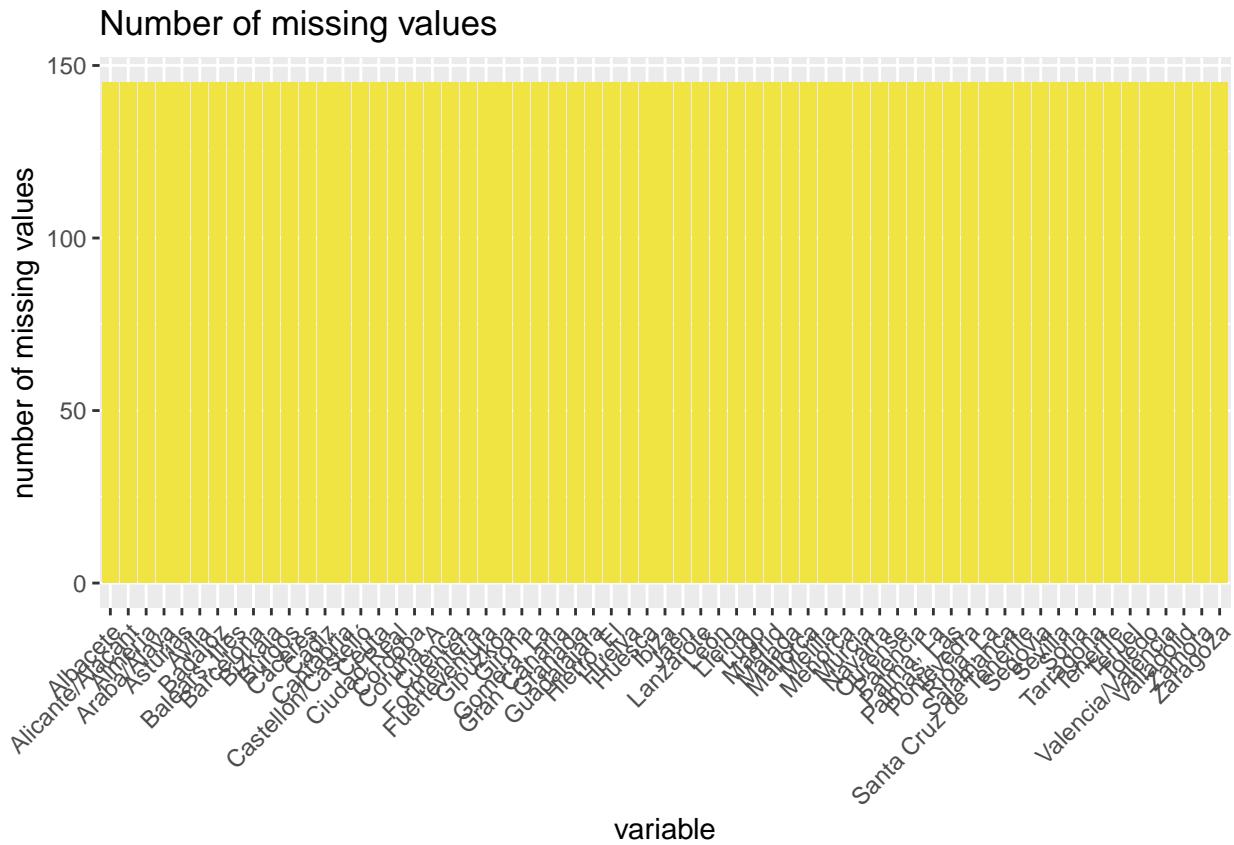
```
##  
##  Variables sorted by number of missings:  
##          Variable    Count  
##          Albacete 0.4982818  
##          Alicante/Alacant 0.4982818  
##          Almería 0.4982818  
##          Araba/Álava 0.4982818  
##          Asturias 0.4982818  
##          Ávila 0.4982818  
##          Badajoz 0.4982818  
##          Balears, Illes 0.4982818  
##          Barcelona 0.4982818  
##          Bizkaia 0.4982818  
##          Burgos 0.4982818  
##          Cáceres 0.4982818  
##          Cádiz 0.4982818  
##          Cantabria 0.4982818  
##          Castellón/Castelló 0.4982818  
##          Ceuta 0.4982818  
##          Ciudad Real 0.4982818  
##          Córdoba 0.4982818  
##          Coruña, A 0.4982818  
##          Cuenca 0.4982818  
##          Formentera 0.4982818  
##          Fuerteventura 0.4982818  
##          Gipuzkoa 0.4982818
```

```

##                  Girona 0.4982818
##          Gomera, La 0.4982818
##      Gran Canaria 0.4982818
##          Granada 0.4982818
##      Guadalajara 0.4982818
##          Hierro, El 0.4982818
##          Huelva 0.4982818
##          Huesca 0.4982818
##          Ibiza 0.4982818
##          Jaén 0.4982818
##      Lanzarote 0.4982818
##          León 0.4982818
##          Lleida 0.4982818
##          Lugo 0.4982818
##          Madrid 0.4982818
##          Málaga 0.4982818
##          Mallorca 0.4982818
##          Melilla 0.4982818
##          Menorca 0.4982818
##          Murcia 0.4982818
##          Navarra 0.4982818
##          Ourense 0.4982818
##          Palencia 0.4982818
##          Palma, La 0.4982818
##      Palmas, Las 0.4982818
##          Pontevedra 0.4982818
##          Rioja, La 0.4982818
##          Salamanca 0.4982818
##      Santa Cruz de Tenerife 0.4982818
##          Segovia 0.4982818
##          Sevilla 0.4982818
##          Soria 0.4982818
##          Tarragona 0.4982818
##          Tenerife 0.4982818
##          Teruel 0.4982818
##          Toledo 0.4982818
##      Valencia/València 0.4982818
##          Valladolid 0.4982818
##          Zamora 0.4982818
##          Zaragoza 0.4982818

EM3_t %>%
  gather(key = "key", value = "val") %>%
  mutate(is.missing = is.na(val)) %>%
  group_by(key, is.missing) %>%
  summarise(num.missing = n()) %>%
  filter(is.missing==T) %>%
  select(-is.missing) %>%
  arrange(desc(num.missing)) %>%
  ggplot() +
  geom_bar(aes(x=key, y=num.missing), stat = 'identity', fill="#F0E442") +
  labs(x='variable', y="number of missing values",
       title='Number of missing values') +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

```



We impute the missing values following the principles stated for imputeTS. Thanks to this approach we almost double the amount of data for analysis by province (It was selected “na_seadec” function due to it covers seasonality aspects -weekdays/weekends in our case-).

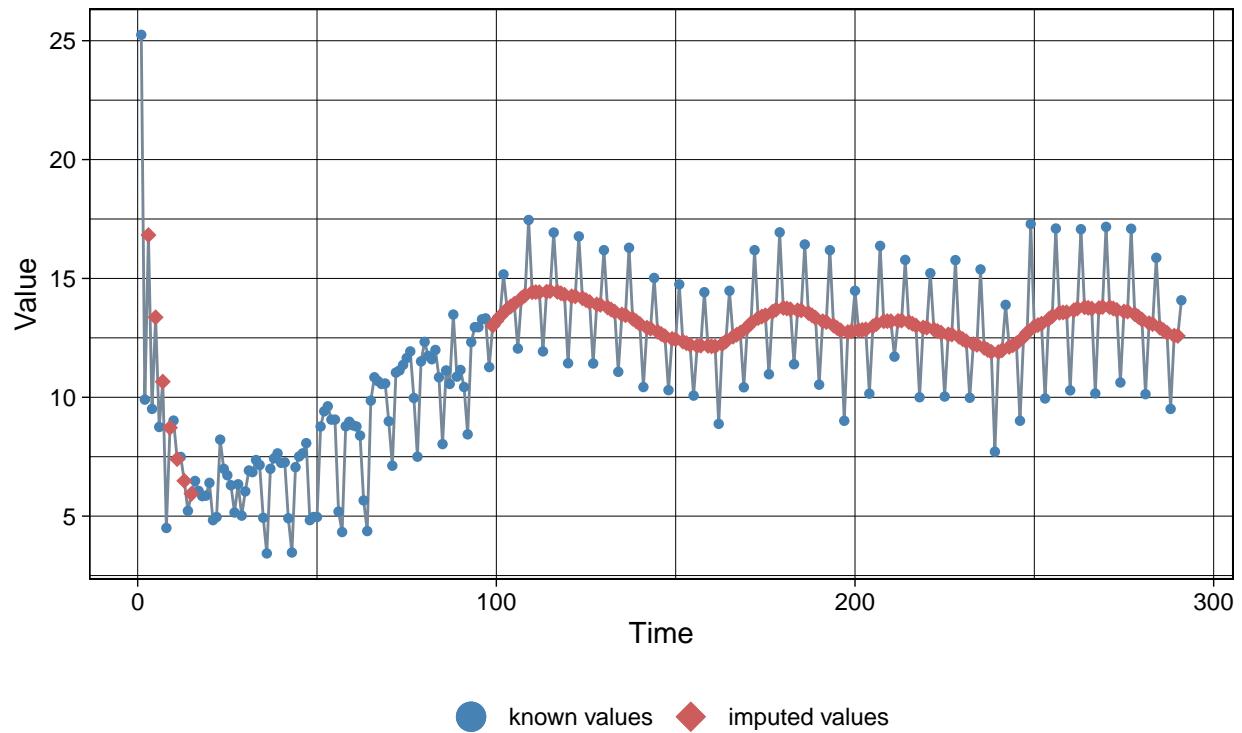
It is needed to transform the dataframe to a time series object.

```
# Used to convert dataframe to ts object
#library(xts)
EM3_t_ts<-xts(EM3_t[,-1],EM3_t$Periodo)

# Impute the missing values with na_kalman, na_seadec, na_interpolation & na_seasplit
#library(imputeTS)
imp <- na_kalman(EM3_t_ts[,1])
ggplot_na_imputations(EM3_t_ts[,1], imp)
```

Imputed Values

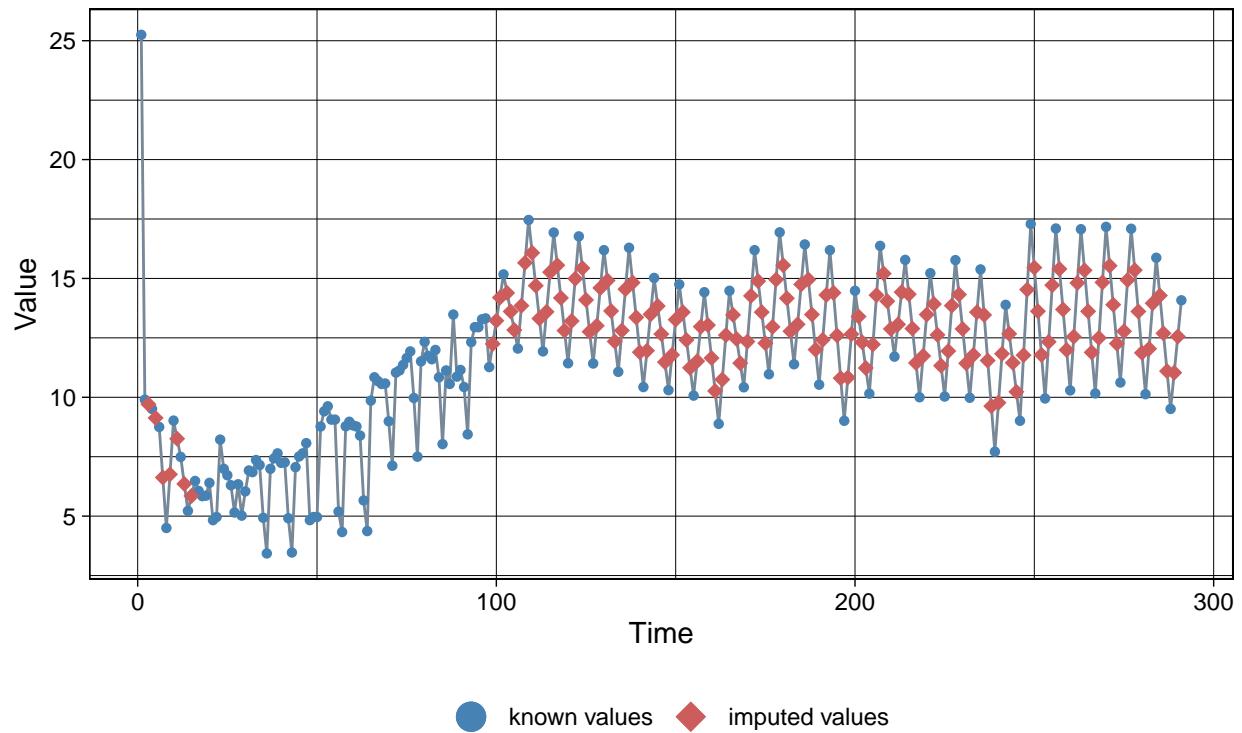
Visualization of missing value replacements



```
imp2 <- na_seadec(EM3_t_ts[,1])
ggplot_na_imputations(EM3_t_ts[,1], imp2)
```

Imputed Values

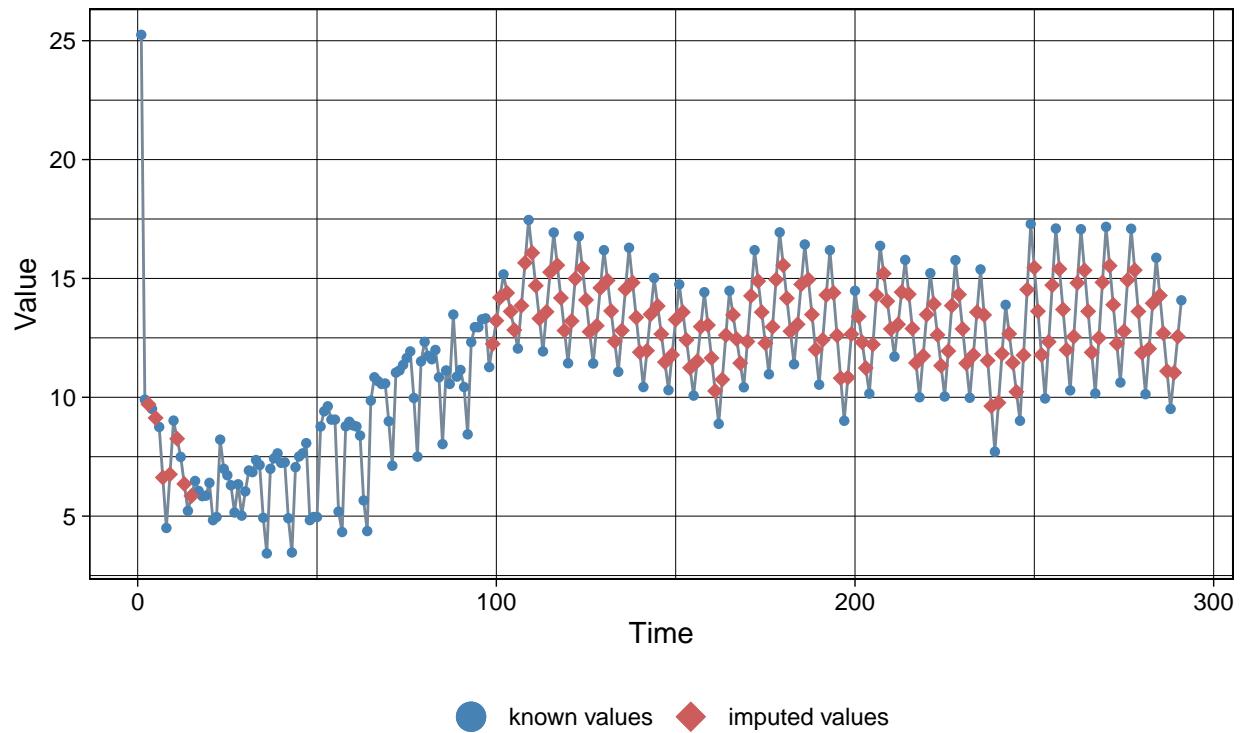
Visualization of missing value replacements



```
imp3 <- na_seasplit(EM3_t_ts[,1])
ggplot_na_imputations(EM3_t_ts[,1], imp3)
```

Imputed Values

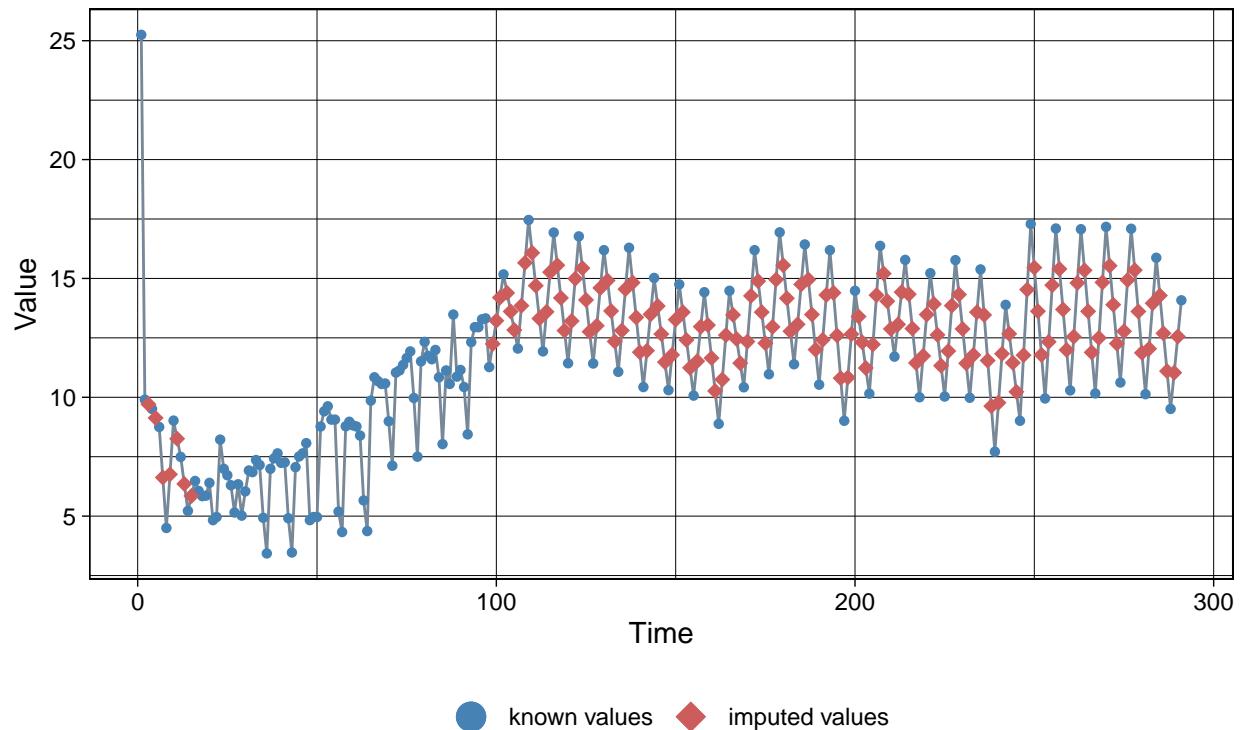
Visualization of missing value replacements



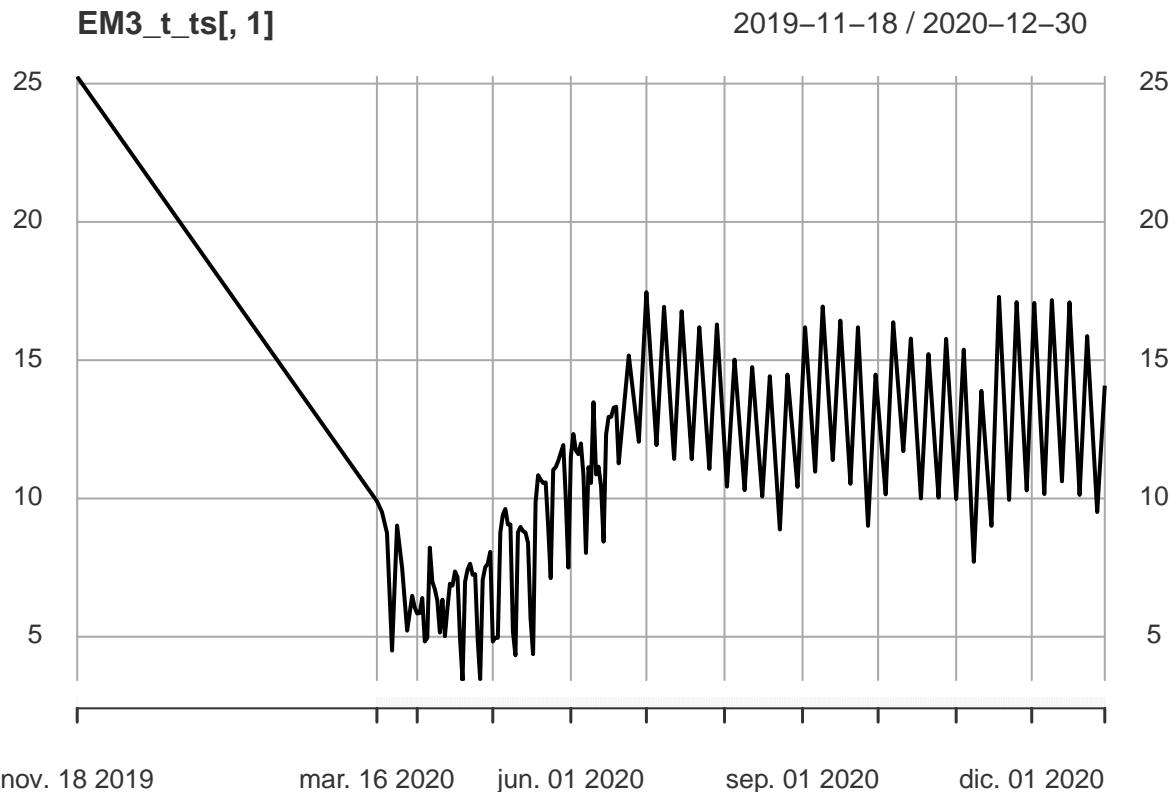
```
imp4 <- na_interpolation(EM3_t_ts[,1])
ggplot_na_imputations(EM3_t_ts[,1], imp4)
```

Imputed Values

Visualization of missing value replacements



```
# We select na_seadec for the dataset
EM3_t_ts <- na_seadec(EM3_t_ts)
plot(EM3_t_ts[,1])
```

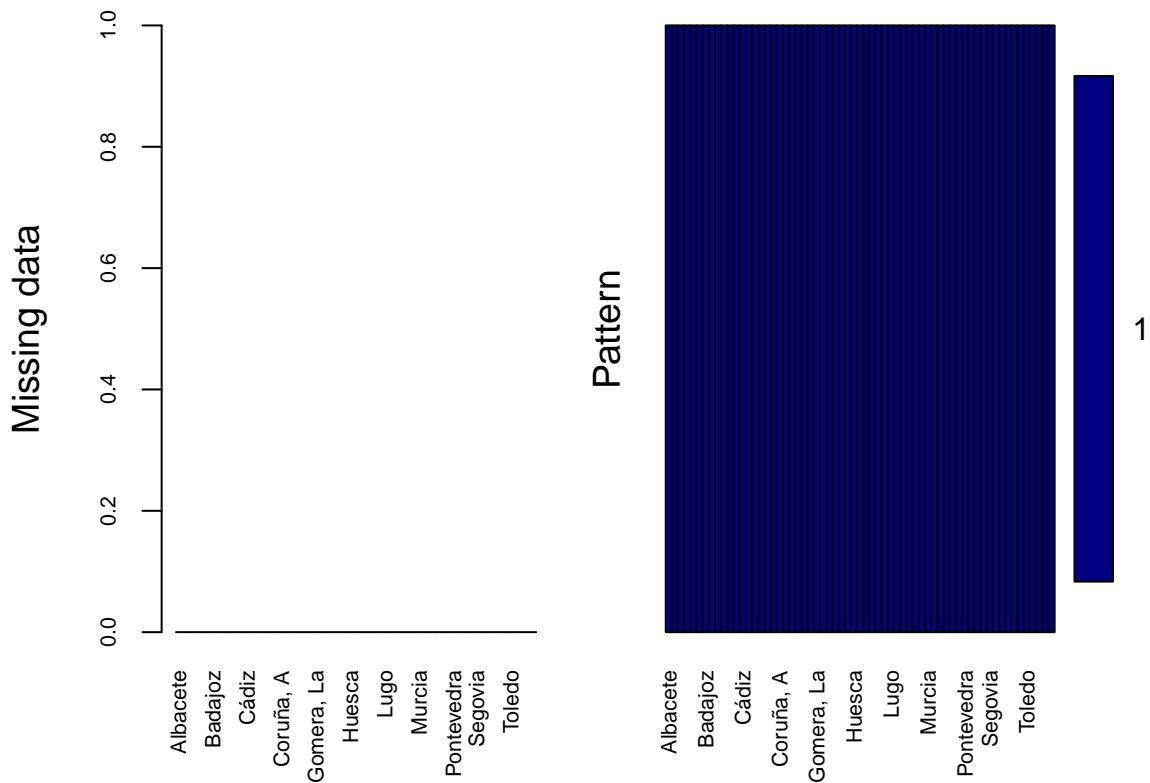


```
# We convert the time series object to a dataframe
library(tsbox)
EM3 <- ts_df(EM3_t_ts)

names(EM3)[names(EM3) == "id"] <- "Zonas.de.movilidad"
names(EM3)[names(EM3) == "time"] <- "Periodo"
names(EM3)[names(EM3) == "value"] <- "Total"

# Transpose dataframe
EM3_t<-dcast(EM3, Periodo~Zonas.de.movilidad, fill=NA)

# We check again missing values (result should be zero)
aggr(EM3_t[,-1], col=c('navyblue','yellow'),
  numbers=TRUE, sortVars=TRUE,
  labels=names(EM3_t[,-1]), cex.axis=.7,
  gap=3, ylab=c("Missing data","Pattern"))
```



```
##  
##  Variables sorted by number of missings:  
##  
##          Variable Count  
##          Albacete      0  
##          Alicante/Alacant 0  
##          Almería       0  
##          Araba/Álava    0  
##          Asturias      0  
##          Ávila          0  
##          Badajoz       0  
##          Balears, Illes  0  
##          Barcelona     0  
##          Bizkaia       0  
##          Burgos         0  
##          Cáceres        0  
##          Cádiz          0  
##          Cantabria     0  
##          Castellón/Castelló 0  
##          Ceuta          0  
##          Ciudad Real   0  
##          Córdoba        0  
##          Coruña, A       0  
##          Cuenca         0  
##          Formentera    0  
##          Fuerteventura  0  
##          Gipuzkoa      0
```

```

##          Girona      0
##      Gomera, La      0
##      Gran Canaria    0
##          Granada      0
##      Guadalajara      0
##      Hierro, El       0
##          Huelva      0
##          Huesca      0
##          Ibiza       0
##          Jaén        0
##      Lanzarote      0
##          León        0
##          Lleida      0
##          Lugo         0
##          Madrid      0
##          Málaga      0
##      Mallorca      0
##          Melilla     0
##          Menorca     0
##          Murcia      0
##          Navarra     0
##          Ourense     0
##          Palencia     0
##          Palma, La     0
##      Palmas, Las      0
##          Pontevedra   0
##          Rioja, La     0
##          Salamanca    0
##      Santa Cruz de Tenerife 0
##          Segovia     0
##          Sevilla     0
##          Soria       0
##          Tarragona    0
##          Tenerife     0
##          Teruel       0
##          Toledo       0
##      Valencia/València 0
##          Valladolid   0
##          Zamora      0
##          Zaragoza     0

EM3_t %>%
  gather(key = "key", value = "val") %>%
  mutate(is.missing = is.na(val)) %>%
  group_by(key, is.missing) %>%
  summarise(num.missing = n()) %>%
  filter(is.missing==T) %>%
  select(-is.missing) %>%
  arrange(desc(num.missing)) %>%
  ggplot() +
  geom_bar(aes(x=key, y=num.missing), stat = 'identity', fill="#F0E442") +
  labs(x='variable', y="number of missing values",
       title='Number of missing values') +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

```

Number of missing values

number of missing values

variable

```
head(str(EM3_t,vec.len=2))

## 'data.frame': 291 obs. of 64 variables:
## $ Periodo : Date, format: "2019-11-18" "2020-03-16" ...
## $ Albacete : num 25.2 9.9 ...
## $ Alicante/Alacant : num 28.1 14.4 ...
## $ Almería : num 24.4 11 ...
## $ Araba/Álava : num 31.9 15.9 ...
## $ Asturias : num 29.9 13.1 ...
## $ Ávila : num 26.62 9.44 ...
## $ Badajoz : num 19.5 10.7 ...
## $ Balears, Illes : num 25.9 12.3 ...
## $ Barcelona : num 34.9 14.6 ...
## $ Bizkaia : num 33.9 17.9 ...
## $ Burgos : num 24.9 11.3 ...
## $ Cáceres : num 21.37 9.08 ...
## $ Cádiz : num 24.5 10.7 ...
## $ Cantabria : num 31.5 14.5 ...
## $ Castellón/Castelló : num 26.9 11.7 ...
## $ Ceuta : num 31.9 12.3 ...
## $ Ciudad Real : num 17.07 8.29 ...
## $ Córdoba : num 22.4 11.9 ...
## $ Coruña, A : num 26.8 15.8 ...
## $ Cuenca : num 20.27 8.54 ...
## $ Formentera : num 12.98 3.08 ...
## $ Fuerteventura : num 11.9 5.76 5.31 4.86 4.62 ...
```

```

## $ Gipuzkoa : num 31.3 14.1 ...
## $ Girona : num 25.6 10.9 ...
## $ Gomera, La : num 11.43 4.26 ...
## $ Gran Canaria : num 31.4 16 ...
## $ Granada : num 24.2 11.4 ...
## $ Guadalajara : num 31.3 12.8 ...
## $ Hierro, El : num 14.82 5.75 ...
## $ Huelva : num 21.8 11.4 ...
## $ Huesca : num 25.5 10.3 ...
## $ Ibiza : num 20.3 10.1 ...
## $ Jaén : num 18.71 8.81 ...
## $ Lanzarote : num 25.1 13.6 ...
## $ León : num 29.8 13.2 ...
## $ Lleida : num 28.4 11.9 ...
## $ Lugo : num 23.6 12.4 ...
## $ Madrid : num 36.7 13.9 ...
## $ Málaga : num 23.7 11.4 ...
## $ Mallorca : num 28.1 13.3 ...
## $ Melilla : num 35 12.5 ...
## $ Menorca : num 16.26 7.21 ...
## $ Murcia : num 24.6 12.4 ...
## $ Navarra : num 30.6 13.7 ...
## $ Ourense : num 27.5 15.6 ...
## $ Palencia : num 31.1 12.9 ...
## $ Palma, La : num 24.6 13.1 ...
## $ Palmas, Las : num 28.5 14.6 ...
## $ Pontevedra : num 26.2 17.1 ...
## $ Rioja, La : num 28 12.1 ...
## $ Salamanca : num 28.1 12.9 ...
## $ Santa Cruz de Tenerife: num 28 13.7 ...
## $ Segovia : num 29 12.6 ...
## $ Sevilla : num 25.9 12.3 ...
## $ Soria : num 17.67 7.85 ...
## $ Tarragona : num 28.4 11.4 ...
## $ Tenerife : num 28.9 14.1 ...
## $ Teruel : num 16.35 6.77 ...
## $ Toledo : num 25.7 12.3 ...
## $ Valencia/València : num 31 15.3 ...
## $ Valladolid : num 29 14.3 ...
## $ Zamora : num 26 11.5 ...
## $ Zaragoza : num 32.3 14.8 ...

## NULL

summary(EM3_t)

##      Periodo          Albacete        Alicante/Alacant       Almería
## Min.   :2019-11-18  Min.   : 3.430  Min.   : 5.72  Min.   : 4.10
## 1st Qu.:2020-05-26  1st Qu.: 9.738  1st Qu.:14.42  1st Qu.:11.31
## Median :2020-08-07  Median :11.960  Median :17.19  Median :13.84
## Mean   :2020-08-06  Mean   :11.592  Mean   :16.16  Mean   :13.19
## 3rd Qu.:2020-10-18  3rd Qu.:13.873  3rd Qu.:18.70  3rd Qu.:15.56
## Max.   :2020-12-30  Max.   :25.250  Max.   :28.12  Max.   :24.35
##      Araba/Álava      Asturias       Ávila           Badajoz
## Min.   : 6.30  Min.   : 5.24  Min.   : 4.51  Min.   : 4.90

```

##	1st Qu.:14.48	1st Qu.:11.93	1st Qu.:10.10	1st Qu.:11.29
##	Median :18.09	Median :16.28	Median :12.71	Median :13.10
##	Mean :17.52	Mean :15.65	Mean :11.93	Mean :12.74
##	3rd Qu.:21.11	3rd Qu.:19.41	3rd Qu.:14.26	3rd Qu.:14.79
##	Max. :31.92	Max. :29.94	Max. :26.62	Max. :19.46
##	Balears, Illes	Barcelona	Bizkaia	Burgos
##	Min. : 3.92	Min. : 6.42	Min. : 7.53	Min. : 3.97
##	1st Qu.:13.09	1st Qu.:13.95	1st Qu.:15.41	1st Qu.:10.74
##	Median :16.14	Median :17.51	Median :19.70	Median :13.06
##	Mean :15.29	Mean :17.05	Mean :19.21	Mean :12.68
##	3rd Qu.:18.46	3rd Qu.:20.25	3rd Qu.:23.06	3rd Qu.:15.26
##	Max. :25.95	Max. :34.90	Max. :33.93	Max. :24.88
##	Cáceres	Cádiz	Cantabria	Castellón/Castelló
##	Min. : 4.02	Min. : 5.42	Min. : 5.84	Min. : 4.95
##	1st Qu.:10.34	1st Qu.:12.59	1st Qu.:13.40	1st Qu.:12.59
##	Median :12.39	Median :15.98	Median :17.57	Median :15.00
##	Mean :11.81	Mean :14.92	Mean :16.93	Mean :14.43
##	3rd Qu.:13.84	3rd Qu.:17.82	3rd Qu.:20.63	3rd Qu.:16.76
##	Max. :21.37	Max. :24.46	Max. :31.50	Max. :26.93
##	Ceuta	Ciudad Real	Córdoba	Coruña, A
##	Min. : 4.87	Min. : 3.080	Min. : 5.68	Min. : 6.02
##	1st Qu.:11.90	1st Qu.: 8.291	1st Qu.:11.49	1st Qu.:12.73
##	Median :13.93	Median :10.050	Median :13.69	Median :15.84
##	Mean :13.34	Mean : 9.597	Mean :13.41	Mean :15.63
##	3rd Qu.:15.34	3rd Qu.:11.453	3rd Qu.:15.67	3rd Qu.:18.96
##	Max. :31.88	Max. :17.070	Max. :22.38	Max. :26.84
##	Cuenca	Formentera	Fuerteventura	Gipuzkoa
##	Min. : 3.010	Min. : 0.830	Min. : 1.140	Min. : 4.71
##	1st Qu.: 9.298	1st Qu.: 2.520	1st Qu.: 5.745	1st Qu.:11.95
##	Median :11.620	Median : 3.562	Median : 7.490	Median :15.49
##	Mean :10.843	Mean : 4.559	Mean : 6.954	Mean :14.94
##	3rd Qu.:12.892	3rd Qu.: 7.320	3rd Qu.: 8.503	3rd Qu.:18.27
##	Max. :20.270	Max. :12.980	Max. :11.900	Max. :31.31
##	Girona	Gomera, La	Gran Canaria	Granada
##	Min. : 4.16	Min. : 1.340	Min. : 5.70	Min. : 5.71
##	1st Qu.:10.24	1st Qu.: 4.860	1st Qu.:15.07	1st Qu.:11.38
##	Median :14.60	Median : 6.635	Median :18.21	Median :14.72
##	Mean :13.45	Mean : 6.111	Mean :17.54	Mean :14.07
##	3rd Qu.:16.65	3rd Qu.: 7.416	3rd Qu.:20.93	3rd Qu.:16.66
##	Max. :25.56	Max. :11.430	Max. :31.44	Max. :24.25
##	Guadalajara	Hierro, El	Huelva	Huesca
##	Min. : 4.89	Min. : 1.560	Min. : 5.51	Min. : 4.24
##	1st Qu.:12.19	1st Qu.: 6.645	1st Qu.:10.94	1st Qu.:11.12
##	Median :14.93	Median : 8.470	Median :13.65	Median :13.77
##	Mean :14.52	Mean : 7.933	Mean :13.23	Mean :12.96
##	3rd Qu.:17.29	3rd Qu.: 9.710	3rd Qu.:15.62	3rd Qu.:15.14
##	Max. :31.33	Max. :14.820	Max. :21.79	Max. :25.48
##	Ibiza	Jaén	Lanzarote	León
##	Min. : 2.960	Min. : 4.17	Min. : 4.73	Min. : 5.56
##	1st Qu.: 9.411	1st Qu.: 9.43	1st Qu.:12.87	1st Qu.:12.58
##	Median :11.550	Median :11.77	Median :14.89	Median :15.65
##	Mean :12.017	Mean :11.24	Mean :14.22	Mean :14.92
##	3rd Qu.:15.325	3rd Qu.:13.22	3rd Qu.:16.58	3rd Qu.:17.50
##	Max. :20.650	Max. :18.71	Max. :25.07	Max. :29.80

```

##      Lleida          Lugo        Madrid       Málaga
## Min.   : 5.16   Min.   : 4.88   Min.   : 6.03   Min.   : 4.56
## 1st Qu.:11.73  1st Qu.:10.94  1st Qu.:13.21  1st Qu.:12.25
## Median :14.89  Median :13.64  Median :16.58  Median :15.41
## Mean   :14.25  Mean   :13.11  Mean   :16.19  Mean   :14.51
## 3rd Qu.:16.81  3rd Qu.:15.65  3rd Qu.:19.51  3rd Qu.:17.51
## Max.   :28.38  Max.   :23.60  Max.   :36.70  Max.   :23.71
##      Mallorca       Melilla     Menorca      Murcia
## Min.   : 4.36   Min.   : 6.48   Min.   : 1.590  Min.   : 5.13
## 1st Qu.:14.31  1st Qu.:13.60  1st Qu.: 8.134  1st Qu.:11.81
## Median :17.52  Median :15.85  Median :10.110  Median :14.19
## Mean   :16.44  Mean   :15.37  Mean   :10.846  Mean   :13.97
## 3rd Qu.:19.72  3rd Qu.:17.39  3rd Qu.:15.880  3rd Qu.:16.45
## Max.   :28.06  Max.   :35.05  Max.   :19.490  Max.   :24.65
##      Navarra        Ourense    Palencia    Palma, La
## Min.   : 5.33   Min.   : 6.43   Min.   : 5.05   Min.   : 5.28
## 1st Qu.:13.45  1st Qu.:12.87  1st Qu.:12.18  1st Qu.:12.96
## Median :16.51  Median :15.62  Median :14.78  Median :15.39
## Mean   :15.89  Mean   :15.35  Mean   :14.40  Mean   :14.79
## 3rd Qu.:18.87  3rd Qu.:18.06  3rd Qu.:16.94  3rd Qu.:17.03
## Max.   :30.61  Max.   :27.48  Max.   :31.07  Max.   :24.63
##      Palmas, Las    Pontevedra  Rioja, La   Salamanca
## Min.   : 5.11   Min.   : 6.15   Min.   : 4.74   Min.   : 6.02
## 1st Qu.:13.98  1st Qu.:13.41  1st Qu.:12.51  1st Qu.:12.61
## Median :16.65  Median :17.23  Median :15.78  Median :15.07
## Mean   :15.98  Mean   :16.63  Mean   :15.01  Mean   :14.54
## 3rd Qu.:18.99  3rd Qu.:20.39  3rd Qu.:17.73  3rd Qu.:17.00
## Max.   :28.53  Max.   :26.22  Max.   :28.00  Max.   :28.12
##      Santa Cruz de Tenerife  Segovia      Sevilla      Soria
## Min.   : 5.20   Min.   : 4.96   Min.   : 5.44   Min.   : 3.03
## 1st Qu.:13.71  1st Qu.:11.83  1st Qu.:12.13  1st Qu.: 8.78
## Median :16.54  Median :14.48  Median :14.96  Median :10.95
## Mean   :15.99  Mean   :13.80  Mean   :14.73  Mean   :10.19
## 3rd Qu.:18.98  3rd Qu.:16.27  3rd Qu.:17.70  3rd Qu.:12.20
## Max.   :28.02  Max.   :29.04  Max.   :25.95  Max.   :17.67
##      Tarragona       Tenerife     Teruel      Toledo
## Min.   : 4.93   Min.   : 5.32   Min.   : 1.950  Min.   : 4.15
## 1st Qu.:10.64  1st Qu.:14.13  1st Qu.: 6.582  1st Qu.:11.00
## Median :14.95  Median :17.02  Median : 8.680  Median :13.52
## Mean   :13.90  Mean   :16.42  Mean   : 8.210  Mean   :13.01
## 3rd Qu.:16.90  3rd Qu.:19.58  3rd Qu.: 9.999  3rd Qu.:15.57
## Max.   :28.38  Max.   :28.87  Max.   :16.350  Max.   :25.70
##      Valencia/València  Valladolid  Zamora      Zaragoza
## Min.   : 6.17   Min.   : 5.55   Min.   : 4.82   Min.   : 5.80
## 1st Qu.:15.39  1st Qu.:12.60  1st Qu.:10.79  1st Qu.:13.44
## Median :18.16  Median :15.95  Median :13.09  Median :16.45
## Mean   :17.53  Mean   :15.71  Mean   :12.79  Mean   :16.37
## 3rd Qu.:20.38  3rd Qu.:19.10  3rd Qu.:14.95  3rd Qu.:19.37
## Max.   :30.97  Max.   :28.99  Max.   :25.99  Max.   :32.26

```

```
table(EM3$Zonas.de.movilidad)
```

```

##          Albacete      Alicante/Alacant      Almería
##                291                  291                  291

```

##	Araba/Álava		Asturias	Ávila
##	291		291	291
##	Badajoz	Balears, Illes		Barcelona
##	291		291	291
##	Bizkaia	Burgos		Cáceres
##	291	291		291
##	Cádiz	Cantabria	Castellón/Castelló	
##	291	291	291	291
##	Ceuta	Ciudad Real		Córdoba
##	291	291		291
##	Coruña, A	Cuenca	Formentera	
##	291	291	291	291
##	Fuerteventura	Gipuzkoa		Girona
##	291	291		291
##	Gomera, La	Gran Canaria		Granada
##	291	291		291
##	Guadalajara	Hierro, El		Huelva
##	291	291		291
##	Huesca	Ibiza		Jaén
##	291	291		291
##	Lanzarote	León		Lleida
##	291	291		291
##	Lugo	Madrid		Málaga
##	291	291		291
##	Mallorca	Melilla		Menorca
##	291	291		291
##	Murcia	Navarra		Ourense
##	291	291		291
##	Palencia	Palma, La	Palmas, Las	
##	291	291	291	
##	Pontevedra	Rioja, La	Salamanca	
##	291	291	291	
##	Santa Cruz de Tenerife	Segovia	Sevilla	
##	291	291	291	
##	Soria	Tarragona	Tenerife	
##	291	291	291	
##	Teruel	Toledo	Valencia/València	
##	291	291	291	
##	Valladolid	Zamora	Zaragoza	
##	291	291	291	

2.1.5 Google review

Here we have data mobility from autonomous-communities and provinces.

```
#Source Google
summary(Google)
```

```
## country_region_code country_region      sub_region_1      sub_region_2
## Length:24242       Length:24242       Length:24242       Length:24242
## Class :character   Class :character   Class :character   Class :character
## Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
```

```

##
## metro_area      iso_3166_2_code      census_fips_code   place_id
## Mode:logical    Length:24242        Mode:logical       Length:24242
## NA's:24242      Class :character  NA's:24242        Class :character
##                      Mode  :character          Mode  :character
##
##
##
##
##      date            retail_and_recreation_percent_change_from_baseline
## Length:24242        Min.   :-97.00
## Class :character    1st Qu.:-53.00
## Mode  :character    Median :-32.00
##                      Mean   :-36.42
##                      3rd Qu.:-17.00
##                      Max.   : 71.00
##                      NA's   :56
## grocery_and_pharmacy_percent_change_from_baseline
## Min.   :-96.000
## 1st Qu.:-18.000
## Median : -4.000
## Mean   : -9.973
## 3rd Qu.:  3.000
## Max.   :194.000
## NA's   :396
## parks_percent_change_from_baseline
## Min.   :-94.0000
## 1st Qu.:-30.0000
## Median : -5.0000
## Mean   : -0.0038
## 3rd Qu.: 22.0000
## Max.   :543.0000
## NA's   :305
## transit_stations_percent_change_from_baseline
## Min.   :-100.00
## 1st Qu.:-46.00
## Median : -30.00
## Mean   : -32.63
## 3rd Qu.:-16.00
## Max.   : 177.00
## NA's   :832
## workplaces_percent_change_from_baseline
## Min.   :-92.0
## 1st Qu.:-37.0
## Median :-24.0
## Mean   :-26.7
## 3rd Qu.:-13.0
## Max.   : 55.0
## NA's   :42
## residential_percent_change_from_baseline
## Min.   :-10.000
## 1st Qu.:  4.000
## Median :  7.000
## Mean   :  9.419

```

```

## 3rd Qu.: 13.000
## Max. : 48.000
## NA's :267
head(str(Google,vec.len=1))

## 'data.frame': 24242 obs. of 15 variables:
## $ country_region_code : chr "ES" ...
## $ country_region : chr "Spain" ...
## $ sub_region_1 : chr "" ...
## $ sub_region_2 : chr "" ...
## $ metro_area : logi NA ...
## $ iso_3166_2_code : chr ""
## $ census_fips_code : logi NA ...
## $ place_id : chr "ChIJi7xhMnjjQgwR7KNoB5Qs7KY" ...
## $ date : chr "15/02/2020" ...
## $ retail_and_recreation_percent_change_from_baseline: int 2 2 ...
## $ grocery_and_pharmacy_percent_change_from_baseline : int -1 3 ...
## $ parks_percent_change_from_baseline : int 26 13 ...
## $ transit_stations_percent_change_from_baseline : int 8 5 ...
## $ workplaces_percent_change_from_baseline : int 0 -1 ...
## $ residential_percent_change_from_baseline : int -2 -2 ...

## NULL
table(Google$sub_region_1)

##
##          Andalusia      Aragon      Asturias
## 385           3465        1540         385
## Balearic Islands Basque Country Canary Islands Cantabria
## 385           1540        1155         385
## Castile-La Mancha Castile and LeÃ³n Catalonia Ceuta
## 2310           3850        1925         378
## Community of Madrid Extremadura Galicia La Rioja
## 385            1155        1925         385
## Melilla           Navarre Region of Murcia Valencian Community
## 379             385        385          1540
table(Google$sub_region_2)

##
##          A CoruÃ±a      Ãvillava
## 7687           385          385
## Ãvila           Albacete Alicante
## 385            385          385
## AlmerÃ¡a          Badajoz Barcelona
## 385            385          385
## Biscay           Burgos CÃ¡ceres
## 385            385          385
## CÃ¡diz           CÃ³rdoba CastellÃ³n
## 385            385          385
## Ciudad Real       Cuenca Gipuzkoa
## 385            385          385
## Girona           Granada Guadalajara
## 385            385          385
## Huelva            Huesca JaÃ©n

```

```

##          385          385          385          385
##      Las Palmas      LeÃ³n      Lleida
##          385          385          385          385
##      Lugo      MÃ¡laga      Palencia
##          385          385          385          385
##      Pontevedra Province of Ourense      Salamanca
##          385          385          385          385
## Santa Cruz de Tenerife      Segovia      Seville
##          385          385          385          385
##      Soria      Tarragona      Teruel
##          385          385          385          385
##      Toledo      Valencia      Valladolid
##          385          385          385          385
##      Zamora      Zaragoza
##          385          385          385          385

table(Google$iso_3166_2_code)

##
##          ES-A ES-AB ES-AL ES-AN ES-AR ES-AS ES-AV   ES-B ES-BA ES-BI ES-BU   ES-C
##      385  385  385  385  385  385  385  385  385  385  385  385  385
##  ES-CA ES-CB ES-CC ES-CE ES-CL ES-CM ES-CN ES-CO ES-CR ES-CS ES-CT ES-CU ES-EX
##      385  385  385  378  385  385  385  385  385  385  385  385  385
##  ES-GA ES-GC ES-GI ES-GR ES-GU   ES-H ES-HU ES-IB   ES-J ES-L ES-LE ES-LU ES-MA
##      385  385  385  385  385  385  385  385  385  385  385  385  385
##  ES-MC ES-MD ES-ML ES-NC ES-OR   ES-P ES-PO ES-PV   ES-RI ES-SA ES-SE ES-SG ES-SO
##      385  385  379  385  385  385  385  385  385  385  385  385  385
##  ES-SS ES-T  ES-TE ES-TF ES-T0   ES-V ES-VA ES-VC   ES-VI ES-Z ES-ZA
##      385  385  385  385  385  385  385  385  385  385  385  385

```

2.1.6 Google autonomous-communities & provinces

We check data grouped by autonomous communities and provinces.

```
Google %>% group_by(sub_region_1) %>% tally()
```

```

## # A tibble: 20 x 2
##   sub_region_1     n
##   <chr>        <int>
## 1 ""            385
## 2 "Andalusia"  3465
## 3 "Aragon"       1540
## 4 "Asturias"     385
## 5 "Balearic Islands" 385
## 6 "Basque Country" 1540
## 7 "Canary Islands" 1155
## 8 "Cantabria"     385
## 9 "Castile-La Mancha" 2310
## 10 "Castile and LeÃ³n" 3850
## 11 "Catalonia"    1925
## 12 "Ceuta"         378
## 13 "Community of Madrid" 385
## 14 "Extremadura"   1155
## 15 "Galicia"        1925
## 16 "La Rioja"       385
## 17 "Melilla"        379

```

```

## 18 "Navarre"          385
## 19 "Region of Murcia" 385
## 20 "Valencian Community" 1540
Google %>% group_by(sub_region_1) %>% count(sub_region_2)

## # A tibble: 63 x 3
## # Groups:   sub_region_1 [20]
##       sub_region_1 sub_region_2     n
##       <chr>        <chr>      <int>
## 1   ""           ""           385
## 2 "Andalusia"   "AlmerÃa"    385
## 3 "Andalusia"   "CÃ¡diz"     385
## 4 "Andalusia"   "CÃ³rdoba"   385
## 5 "Andalusia"   "Granada"   385
## 6 "Andalusia"   "Huelva"     385
## 7 "Andalusia"   "JaÃ±on"     385
## 8 "Andalusia"   "MÃ¡laga"    385
## 9 "Andalusia"   "Seville"   385
## 10 "Andalusia"  ... with 53 more rows

```

In Spain there are **autonomous communities (AC)** and **autonomous cities (C)** that are considered as **provinces (Pr)** as well. This is the case for:

- AC - Asturias, Principality - Pr - Asturias
- AC - Balears, Illes - Pr - Balears, Illes
- AC - Cantabria - Pr - Cantabria
- AC - Madrid, Community - Pr - Madrid
- AC - Murcia, Region - Pr - Murcia
- AC - Navarra, Foral Community - Pr - Navarra
- AC - Rioja, La - Pr - Rioja, La
- C - Ceuta - C/Pr - Ceuta
- C - Melilla - C/Pr - Melilla

In this data set, the empty values in the “sub_region_2” column, for the autonomous communities mentioned, will be replaced by the value contained in the “sub_region_1” column (A). Also we are going to modify the names of the provinces that have special characters in order to adopt the INE standards (B). See note.

Note The following links states the provinces in Spain INE CCAA and its ISO codes are going to be used as tables of reference.

```

# Modification provinces - A
Google$sub_region_2[Google$sub_region_1=="Balearic Islands"] <- "Balears, Illes"
Google$iso_3166_2_code[Google$sub_region_2=="Balears, Illes"] <- "PM"

Google$sub_region_2[Google$sub_region_1=="Asturias"] <- "Asturias"
Google$iso_3166_2_code[Google$sub_region_2=="Asturias"] <- "0"

Google$sub_region_2[Google$sub_region_1=="Cantabria"] <- "Cantabria"
Google$iso_3166_2_code[Google$sub_region_2=="Cantabria"] <- "S"

Google$sub_region_2[Google$sub_region_1=="Community of Madrid"] <- "Madrid"
Google$iso_3166_2_code[Google$sub_region_2=="Madrid"] <- "M"

Google$sub_region_2[Google$sub_region_1=="Region of Murcia"] <- "Murcia"
Google$iso_3166_2_code[Google$sub_region_2=="Murcia"] <- "MU"

```

```

Google$sub_region_2[Google$sub_region_1=="Navarre"] <- "Navarra"
Google$iso_3166_2_code[Google$sub_region_2=="Navarra"] <- "NA"

Google$sub_region_2[Google$sub_region_1=="La Rioja"] <- "Rioja, La"
Google$iso_3166_2_code[Google$sub_region_2=="Rioja, La"] <- "LO"

Google$sub_region_2[Google$sub_region_1=="Ceuta"] <- "Ceuta"
Google$iso_3166_2_code[Google$sub_region_2=="Ceuta"] <- "CE"

Google$sub_region_2[Google$sub_region_1=="Melilla"] <- "Melilla"
Google$iso_3166_2_code[Google$sub_region_2=="Melilla"] <- "ML"

# Modification provinces - B
Google$sub_region_2[Google$sub_region_2=="A Coruña, A"] <- "Coruña, A"
Google$sub_region_2[Google$sub_region_2=="Á\u00f1alava"] <- "Araba/Álava"
Google$sub_region_2[Google$sub_region_2=="Á\u00f1ivila"] <- "Ávila"
Google$sub_region_2[Google$sub_region_2=="Alicante"] <- "Alicante/Alacant"
Google$sub_region_2[Google$sub_region_2=="Biscay"] <- "Bizkaia"
Google$sub_region_2[Google$sub_region_2=="Cáceres"] <- "Cáceres"
Google$sub_region_2[Google$sub_region_2=="Cádiz"] <- "Cádiz"
Google$sub_region_2[Google$sub_region_2=="Córdoba"] <- "Córdoba"
Google$sub_region_2[Google$sub_region_2=="Castellón"] <- "Castellón/Castelló"
Google$sub_region_2[Google$sub_region_2=="Jaén"] <- "Jaén"
Google$sub_region_2[Google$sub_region_2=="Las Palmas"] <- "Palmas, Las"
Google$sub_region_2[Google$sub_region_2=="León"] <- "León"
Google$sub_region_2[Google$sub_region_2=="Málaga"] <- "Málaga"
Google$sub_region_2[Google$sub_region_2=="Province of Ourense"] <- "Ourense"
Google$sub_region_2[Google$sub_region_2=="Seville"] <- "Sevilla"
Google$sub_region_2[Google$sub_region_2=="Valencia"] <- "Valencia/València"
Google$sub_region_2 <- with(Google, ifelse(grepl("Almer", sub_region_2),
                                         "Almería", sub_region_2))

# Table check
table(Google$sub_region_2)

```

		Albacete	Alicante/Alacant
##	4235	385	385
##	Almería	Araba/Álava	Asturias
##	385	385	385
##	Ávila	Badajoz	Balears, Illes
##	385	385	385
##	Barcelona	Bizkaia	Burgos
##	385	385	385
##	Cáceres	Cádiz	Cantabria
##	385	385	385
##	Castellón/Castelló	Ceuta	Ciudad Real
##	385	378	385
##	Córdoba	Coruña, A	Cuenca
##	385	385	385
##	Gipuzkoa	Girona	Granada
##	385	385	385
##	Guadalajara	Huelva	Huesca
##	385	385	385

```

## Jaén León Lleida
## 385 385 385
## Lugo Madrid Málaga
## 385 385 385
## Melilla Murcia Navarra
## 379 385 385
## Ourense Palencia Palmas, Las
## 385 385 385
## Pontevedra Rioja, La Salamanca
## 385 385 385
## Santa Cruz de Tenerife Segovia Sevilla
## 385 385 385
## Soria Tarragona Teruel
## 385 385 385
## Toledo Valencia/València Valladolid
## 385 385 385
## Zamora Zaragoza
## 385 385

table(Google$iso_3166_2_code)

## CE ES-A ES-AB ES-AL ES-AN ES-AR ES-AV ES-B ES-BA ES-BI ES-BU ES-C
## 385 378 385 385 385 385 385 385 385 385 385 385 385
## ES-CA ES-CC ES-CL ES-CM ES-CN ES-CO ES-CR ES-CS ES-CT ES-CU ES-EX ES-GA ES-GC
## 385 385 385 385 385 385 385 385 385 385 385 385 385
## ES-GI ES-GR ES-GU ES-H ES-HU ES-J ES-L ES-LE ES-LU ES-MA ES-OR ES-P ES-PO
## 385 385 385 385 385 385 385 385 385 385 385 385 385
## ES-PV ES-SA ES-SE ES-SG ES-SO ES-SS ES-T ES-TE ES-TF ES-TO ES-V ES-VA ES-VC
## 385 385 385 385 385 385 385 385 385 385 385 385 385
## ES-VI ES-Z ES-ZA LO M ML MU NA 0 PM S
## 385 385 385 385 385 379 385 385 385 385 385 385

```

2.1.7 Google data transformation

We are going to **transform / eliminate**:

- A - Rows with “na” / “” in “sub_region_1” and “sub_region_2” columns are eliminated.
- B - Date column is transformed from “character” to “date.”
- C - Some columns are eliminated due to they are not adding value for us or they contain blanks (country_region_code, country_region, metro_area, census_fips_code, place_id).
- D - “ES-” is eliminated from “iso_3166_2_code” column.
- E - We changed from integer to numeric the integer columns.

```

# Transform / eliminate A
Google <- filter(Google, sub_region_1 != "", sub_region_2 != "" )

# Transform / eliminate B
Google$date <- as.Date(Google$date ,format="%d/%m/%Y")

# Transform / eliminate C
Google<-within(Google, rm(country_region_code,
                           country_region,
                           metro_area,
                           census_fips_code,
                           place_id))

```

```

# Transform / eliminate D
Google$iso_3166_2_code <- gsub("ES-", "", Google$iso_3166_2_code)

# We pass from integer to numeric
Google$retail_and_recreation_percent_change_from_baseline <-
  as.numeric(Google$retail_and_recreation_percent_change_from_baseline)
Google$grocery_and_pharmacy_percent_change_from_baseline <-
  as.numeric(Google$grocery_and_pharmacy_percent_change_from_baseline)
Google$parks_percent_change_from_baseline <-
  as.numeric(Google$parks_percent_change_from_baseline)
Google$transit_stations_percent_change_from_baseline <-
  as.numeric(Google$transit_stations_percent_change_from_baseline)
Google$workplaces_percent_change_from_baseline <-
  as.numeric(Google$workplaces_percent_change_from_baseline)
Google$residential_percent_change_from_baseline <-
  as.numeric(Google$residential_percent_change_from_baseline)

# Check table
head(Google, 5)

##   sub_region_1 sub_region_2 iso_3166_2_code      date
## 1 Andalusia     Almería          AL 2020-02-15
## 2 Andalusia     Almería          AL 2020-02-16
## 3 Andalusia     Almería          AL 2020-02-17
## 4 Andalusia     Almería          AL 2020-02-18
## 5 Andalusia     Almería          AL 2020-02-19
##   retail_and_recreation_percent_change_from_baseline
## 1                               5
## 2                             -2
## 3                              0
## 4                             -3
## 5                             -1
##   grocery_and_pharmacy_percent_change_from_baseline
## 1                           -3
## 2                             0
## 3                           -2
## 4                           -3
## 5                           -3
##   parks_percent_change_from_baseline
## 1                            40
## 2                           -2
## 3                            3
## 4                           -2
## 5                            3
##   transit_stations_percent_change_from_baseline
## 1                            10
## 2                             1
## 3                            5
## 4                            5
## 5                            4
##   workplaces_percent_change_from_baseline
## 1                             1
## 2                             1

```

```

## 3          3
## 4          3
## 5          3
##   residential_percent_change_from_baseline
## 1          -2
## 2          -1
## 3          -1
## 4           0
## 5           0

table(Google$sub_region_2)

##          Albacete      Alicante/Alacant      Almería
##            385              385              385
##      Araba/Álava          Asturias          Ávila
##            385              385              385
##      Badajoz      Balears, Illes      Barcelona
##            385              385              385
##      Bizkaia          Burgos          Cáceres
##            385              385              385
##      Cádiz          Cantabria Castellón/Castelló
##            385              385              385
##      Ceuta          Ciudad Real      Córdoba
##            378              385              385
##      Coruña, A          Cuenca          Gipuzkoa
##            385              385              385
##      Girona          Granada      Guadalajara
##            385              385              385
##      Huelva          Huesca          Jaén
##            385              385              385
##      León           Lleida          Lugo
##            385              385              385
##      Madrid          Málaga      Melilla
##            385              385              379
##      Murcia          Navarra      Ourense
##            385              385              385
##      Palencia      Palmas, Las      Pontevedra
##            385              385              385
##      Rioja, La      Salamanca Santa Cruz de Tenerife
##            385              385              385
##      Segovia          Sevilla          Soria
##            385              385              385
##      Tarragona          Teruel          Toledo
##            385              385              385
##      Valencia/València      Valladolid      Zamora
##            385              385              385
##      Zaragoza
##            385

table(Google$iso_3166_2_code)

##          A    AB    AL    AV    B    BA    BI    BU    C    CA    CC    CE    CO    CR    CS    CU    GC    GI    GR    GU
## 385 385 385 385 385 385 385 385 385 385 385 385 385 378 385 385 385 385 385 385 385 385 385

```

```

##   H   HU   J   L   LE   LO   LU   M   MA   ML   MU   NA   O   OR   P   PM   PO   S   SA   SE
## 385 385 385 385 385 385 385 385 385 385 379 385 385 385 385 385 385 385 385 385 385 385
##   SG   SO   SS   T   TE   TF   TO   V   VA   VI   Z   ZA
## 385 385 385 385 385 385 385 385 385 385 385 385

```

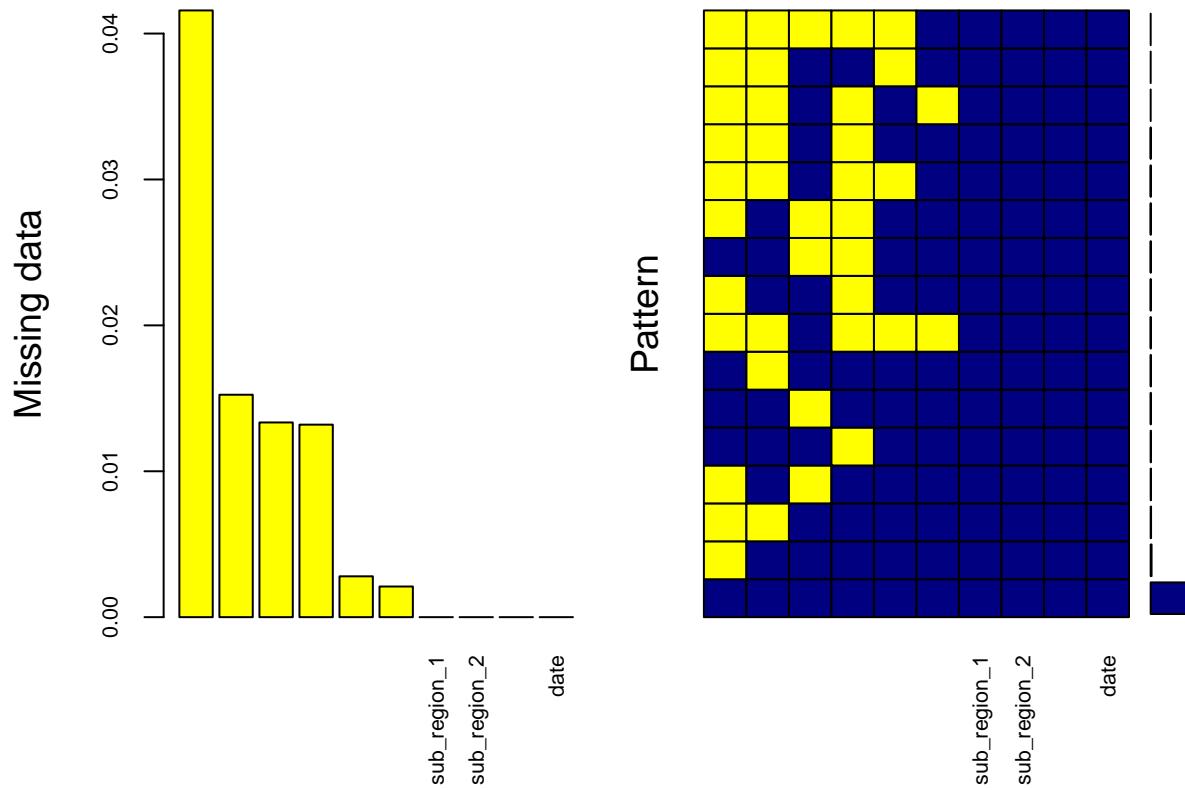
2.1.8 Google review missing values & impute

We check missing values.

```

aggr(Google, col=c('navyblue','yellow'),
      numbers=TRUE, sortVars=TRUE,
      labels=names(Google), cex.axis=.7,
      gap=3, ylab=c("Missing data","Pattern"))

```



```

##
##  Variables sorted by number of missings:
##                                         Variable      Count
## transit_stations_percent_change_from_baseline 0.041585445
## parks_percent_change_from_baseline 0.015244664
## residential_percent_change_from_baseline 0.013345329
## grocery_and_pharmacy_percent_change_from_baseline 0.013195382
## retail_and_recreation_percent_change_from_baseline 0.002799020
## workplaces_percent_change_from_baseline 0.002099265
## sub_region_1 0.000000000
## sub_region_2 0.000000000
## iso_3166_2_code 0.000000000
## date 0.000000000

```

```

Google %>%
  gather(key = "key", value = "val") %>%
  mutate(is.missing = is.na(val)) %>%
  group_by(key, is.missing) %>%
  summarise(num.missing = n()) %>%
  filter(is.missing==T) %>%
  select(-is.missing) %>%
  arrange(desc(num.missing)) %>%
  ggplot() +
  geom_bar(aes(x=key, y=num.missing), stat = 'identity', fill="#F0E442") +
  labs(x='variable', y="number of missing values",
       title='Number of missing values') +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

```



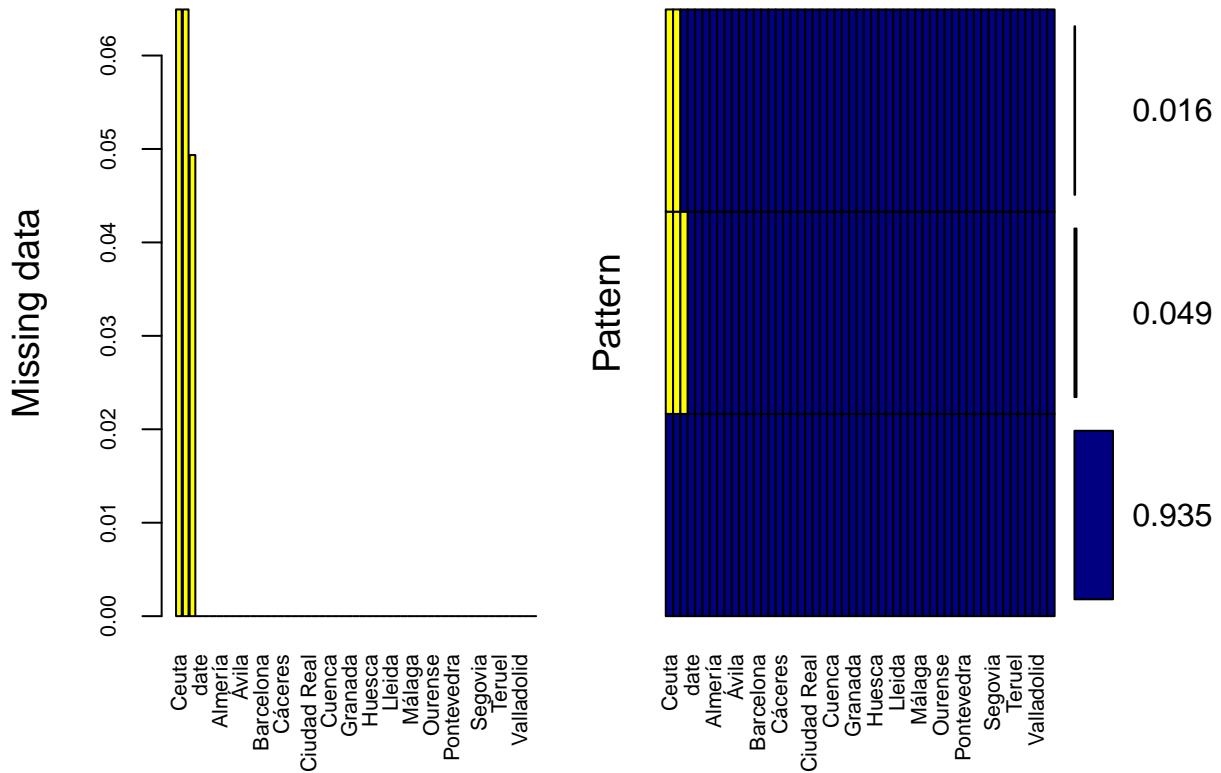
We generate 6 new dataframes from the 6 features stated in order to impute missing values by province using the approach stated at “imputeTS” library (and also used at EM3).

```

# Transpose dataframe
Google_retail<-Google[c(2,4,5)]
Google_t_retail<-dcast(Google_retail, date~sub_region_2, fill=NA)

# Visualize missing values
aggr(Google_t_retail, col=c('navyblue','yellow'),
      numbers=TRUE, sortVars=TRUE,
      labels=names(Google_t_retail), cex.axis=.7,
      gap=3, ylab=c("Missing data", "Pattern"))

```



```
##  
##  Variables sorted by number of missings:  
##  
##          Variable      Count  
##          Ceuta 0.06493506  
##          Melilla 0.06493506  
##          Soria 0.04935065  
##          date 0.00000000  
##          Albacete 0.00000000  
##          Alicante/Alacant 0.00000000  
##          Almería 0.00000000  
##          Araba/Álava 0.00000000  
##          Asturias 0.00000000  
##          Ávila 0.00000000  
##          Badajoz 0.00000000  
##          Balears, Illes 0.00000000  
##          Barcelona 0.00000000  
##          Bizkaia 0.00000000  
##          Burgos 0.00000000  
##          Cáceres 0.00000000  
##          Cádiz 0.00000000  
##          Cantabria 0.00000000  
##          Castellón/Castelló 0.00000000  
##          Ciudad Real 0.00000000  
##          Córdoba 0.00000000  
##          Coruña, A 0.00000000  
##          Cuenca 0.00000000
```

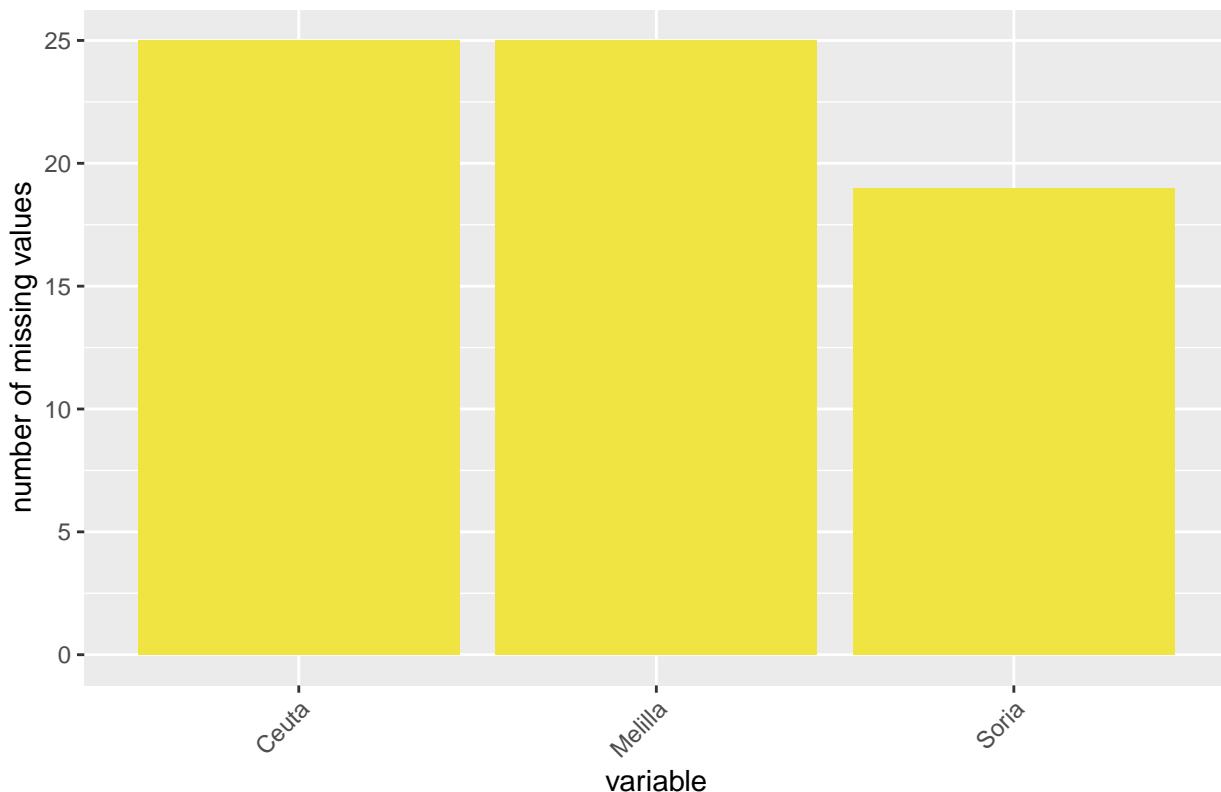
```

##          Gipuzkoa 0.00000000
##          Girona 0.00000000
##          Granada 0.00000000
##          Guadalajara 0.00000000
##          Huelva 0.00000000
##          Huesca 0.00000000
##          Jaén 0.00000000
##          León 0.00000000
##          Lleida 0.00000000
##          Lugo 0.00000000
##          Madrid 0.00000000
##          Málaga 0.00000000
##          Murcia 0.00000000
##          Navarra 0.00000000
##          Ourense 0.00000000
##          Palencia 0.00000000
##          Palmas, Las 0.00000000
##          Pontevedra 0.00000000
##          Rioja, La 0.00000000
##          Salamanca 0.00000000
## Santa Cruz de Tenerife 0.00000000
##          Segovia 0.00000000
##          Sevilla 0.00000000
##          Tarragona 0.00000000
##          Teruel 0.00000000
##          Toledo 0.00000000
## Valencia/València 0.00000000
##          Valladolid 0.00000000
##          Zamora 0.00000000
##          Zaragoza 0.00000000

Google_t_retail %>%
  gather(key = "key", value = "val") %>%
  mutate(is.missing = is.na(val)) %>%
  group_by(key, is.missing) %>%
  summarise(num.missing = n()) %>%
  filter(is.missing==T) %>%
  select(-is.missing) %>%
  arrange(desc(num.missing)) %>%
  ggplot() +
  geom_bar(aes(x=key, y=num.missing), stat = 'identity', fill="#F0E442") +
  labs(x='variable', y="number of missing values",
       title='Number of missing values') +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

```

Number of missing values

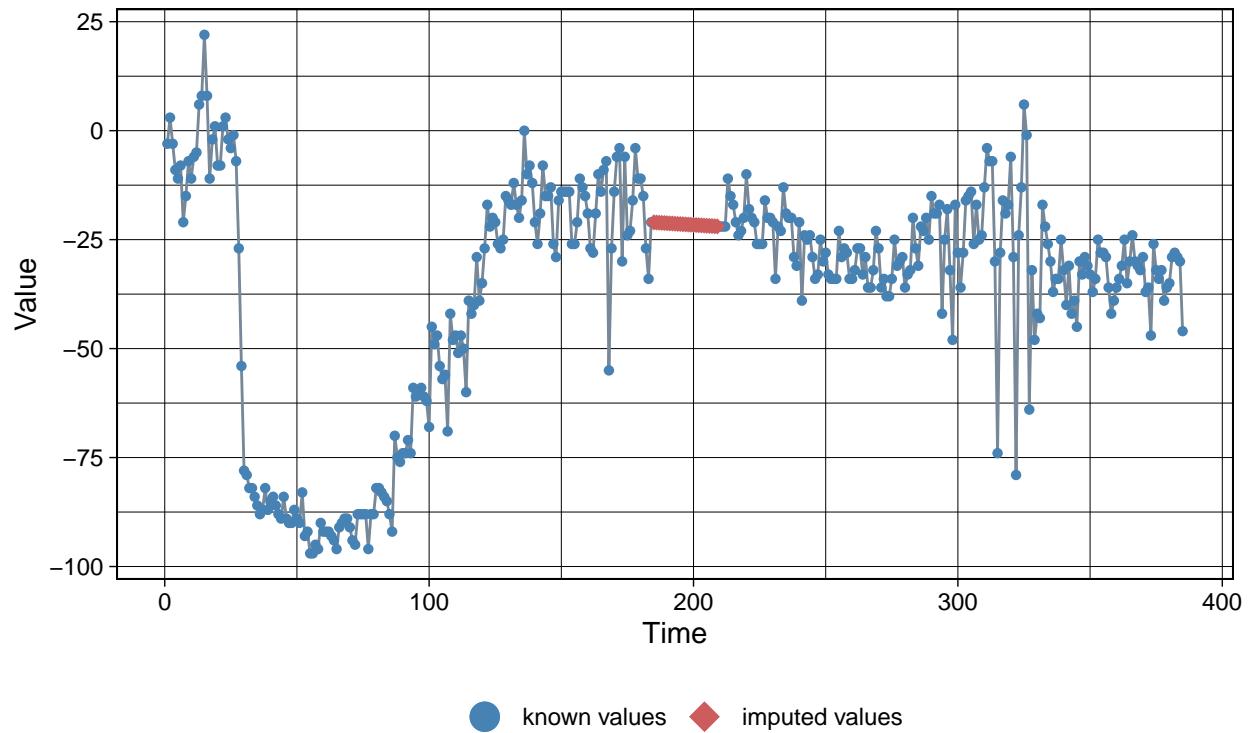


```
# Convert dataframe to ts object
Google_t_retail_ts<-xts(Google_t_retail[-1],Google_t_retail$date)

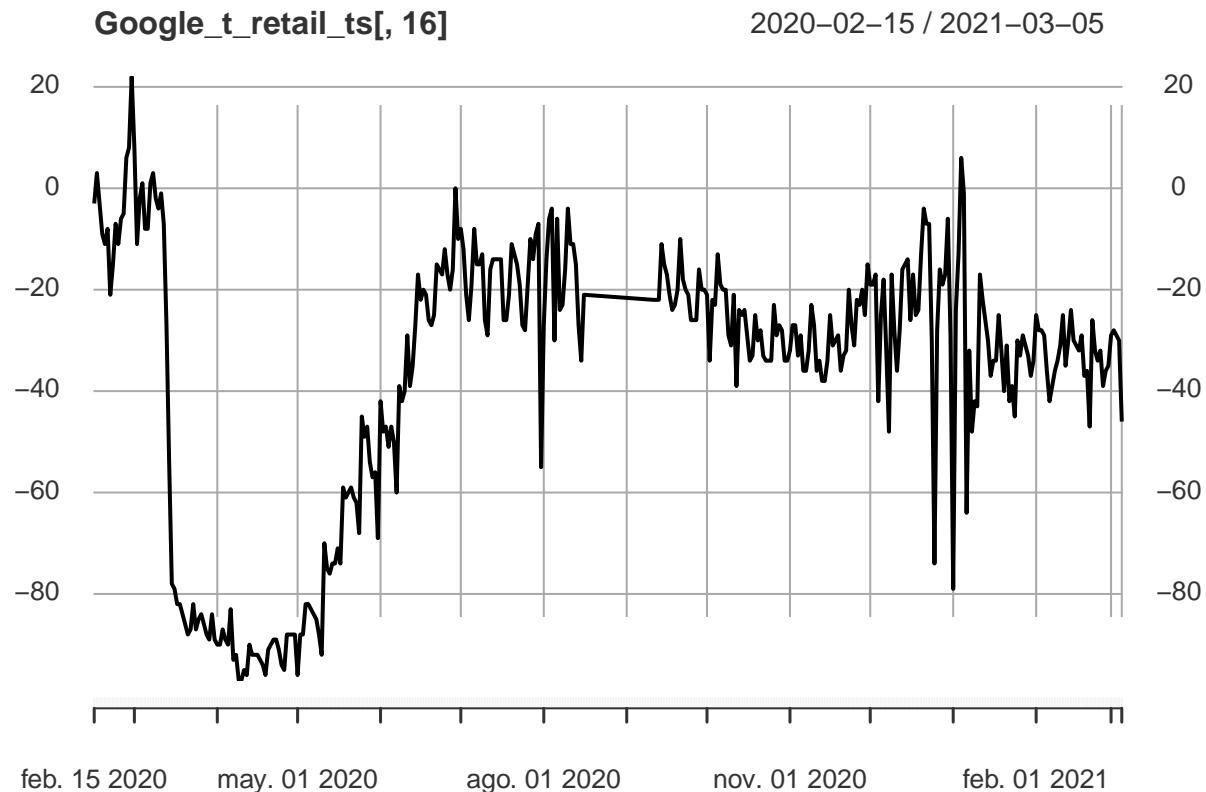
# Impute the missing values with na_seadec (i.e Ceuta)
imp5 <- na_seadec(Google_t_retail_ts[,16])
ggplot_na_imputations(Google_t_retail_ts[,16], imp5)
```

Imputed Values

Visualization of missing value replacements



```
# We select na_seadec for the dataset
Google_t_retail_ts <- na_seadec(Google_t_retail_ts)
plot(Google_t_retail_ts[,16])
```

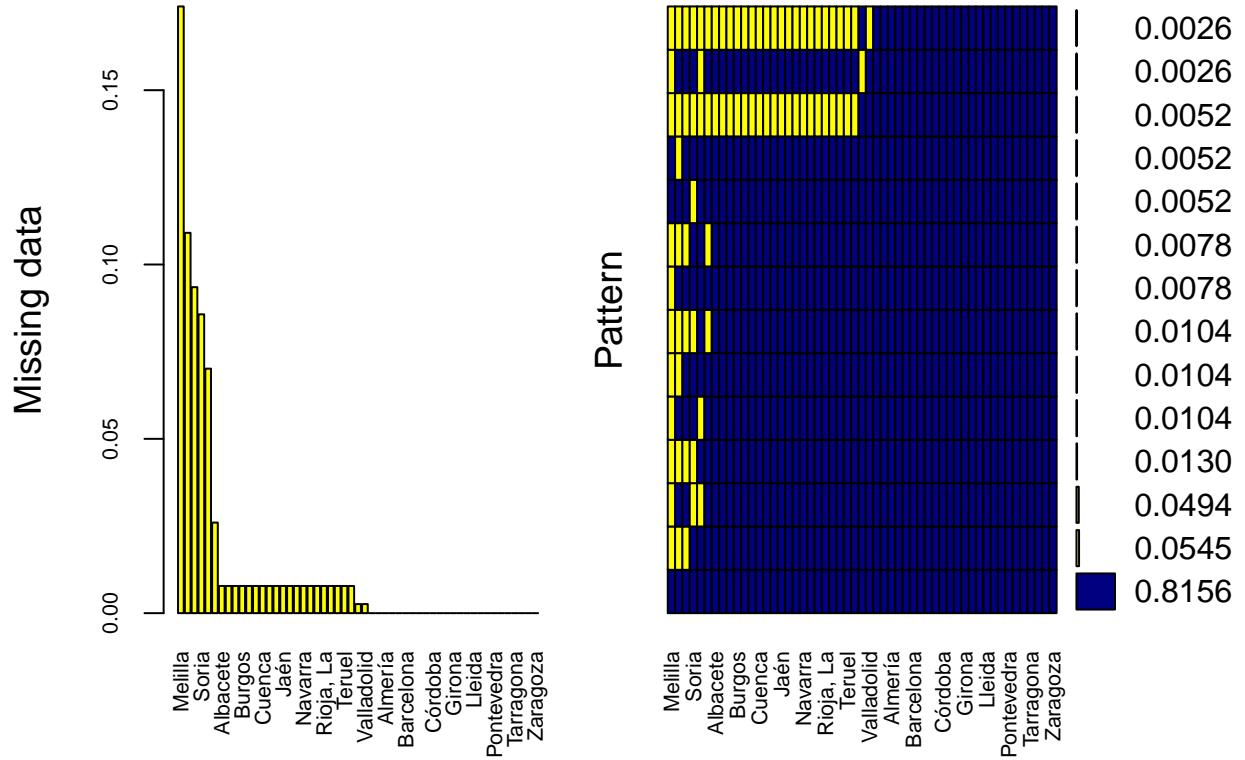


```
# We convert the time series object to a dataframe
Google_retail <- ts_df(Google_t_retail_ts)

names(Google_retail)[names(Google_retail) == "id"] <- "sub_region_2"
names(Google_retail)[names(Google_retail) == "time"] <- "Date"
names(Google_retail)[names(Google_retail) == "value"] <-
  "retail_and_recreation_percent_change_from_baseline"

#####
# Transpose dataframe
Google_grocery<-Google[,c(2,4,6)]
Google_t_grocery<-dcast(Google_grocery, date~sub_region_2, fill=NA)

# Visualize missing values
aggr(Google_t_grocery, col=c('navyblue','yellow'),
  numbers=TRUE, sortVars=TRUE,
  labels=names(Google_t_grocery), cex.axis=.7,
  gap=3, ylab=c("Missing data","Pattern"))
```



```
##  
##  Variables sorted by number of missings:  
##  
##          Variable      Count  
##          Melilla 0.174025974  
##          Asturias 0.109090909  
##          Murcia 0.093506494  
##          Soria 0.085714286  
##          Ceuta 0.070129870  
##          Cantabria 0.025974026  
##          Albacete 0.007792208  
##          Araba/Álava 0.007792208  
##          Ávila 0.007792208  
##          Burgos 0.007792208  
##          Cáceres 0.007792208  
##          Ciudad Real 0.007792208  
##          Cuenca 0.007792208  
##          Guadalajara 0.007792208  
##          Huesca 0.007792208  
##          Jaén 0.007792208  
##          León 0.007792208  
##          Lugo 0.007792208  
##          Navarra 0.007792208  
##          Ourense 0.007792208  
##          Palencia 0.007792208  
##          Rioja, La 0.007792208  
##          Salamanca 0.007792208
```

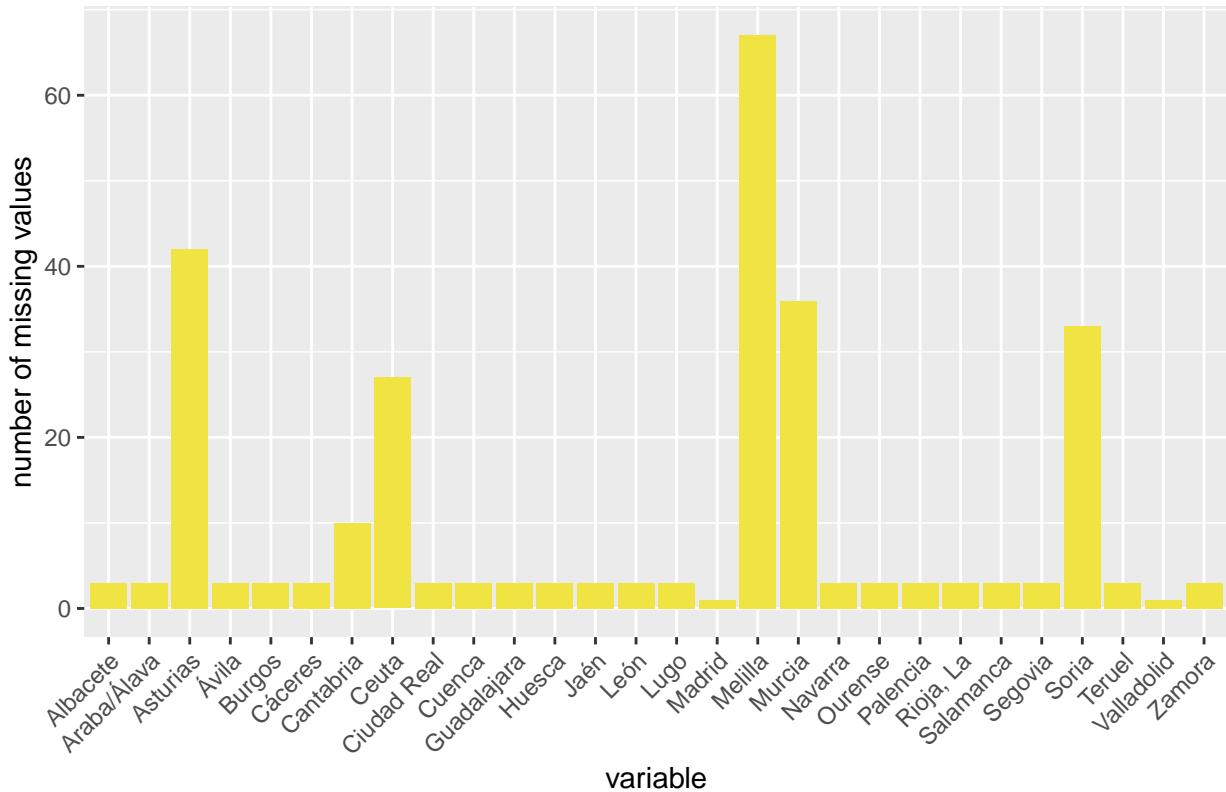
```

## Segovia 0.007792208
## Teruel 0.007792208
## Zamora 0.007792208
## Madrid 0.002597403
## Valladolid 0.002597403
## date 0.000000000
## Alicante/Alacant 0.000000000
## Almería 0.000000000
## Badajoz 0.000000000
## Balears, Illes 0.000000000
## Barcelona 0.000000000
## Bizkaia 0.000000000
## Cádiz 0.000000000
## Castellón/Castelló 0.000000000
## Córdoba 0.000000000
## Coruña, A 0.000000000
## Gipuzkoa 0.000000000
## Girona 0.000000000
## Granada 0.000000000
## Huelva 0.000000000
## Lleida 0.000000000
## Málaga 0.000000000
## Palmas, Las 0.000000000
## Pontevedra 0.000000000
## Santa Cruz de Tenerife 0.000000000
## Sevilla 0.000000000
## Tarragona 0.000000000
## Toledo 0.000000000
## Valencia/València 0.000000000
## Zaragoza 0.000000000

Google_t_grocery %>%
  gather(key = "key", value = "val") %>%
  mutate(is.missing = is.na(val)) %>%
  group_by(key, is.missing) %>%
  summarise(num.missing = n()) %>%
  filter(is.missing==T) %>%
  select(-is.missing) %>%
  arrange(desc(num.missing)) %>%
  ggplot() +
  geom_bar(aes(x=key, y=num.missing), stat = 'identity', fill="#F0E442") +
  labs(x='variable', y="number of missing values",
       title='Number of missing values') +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

```

Number of missing values

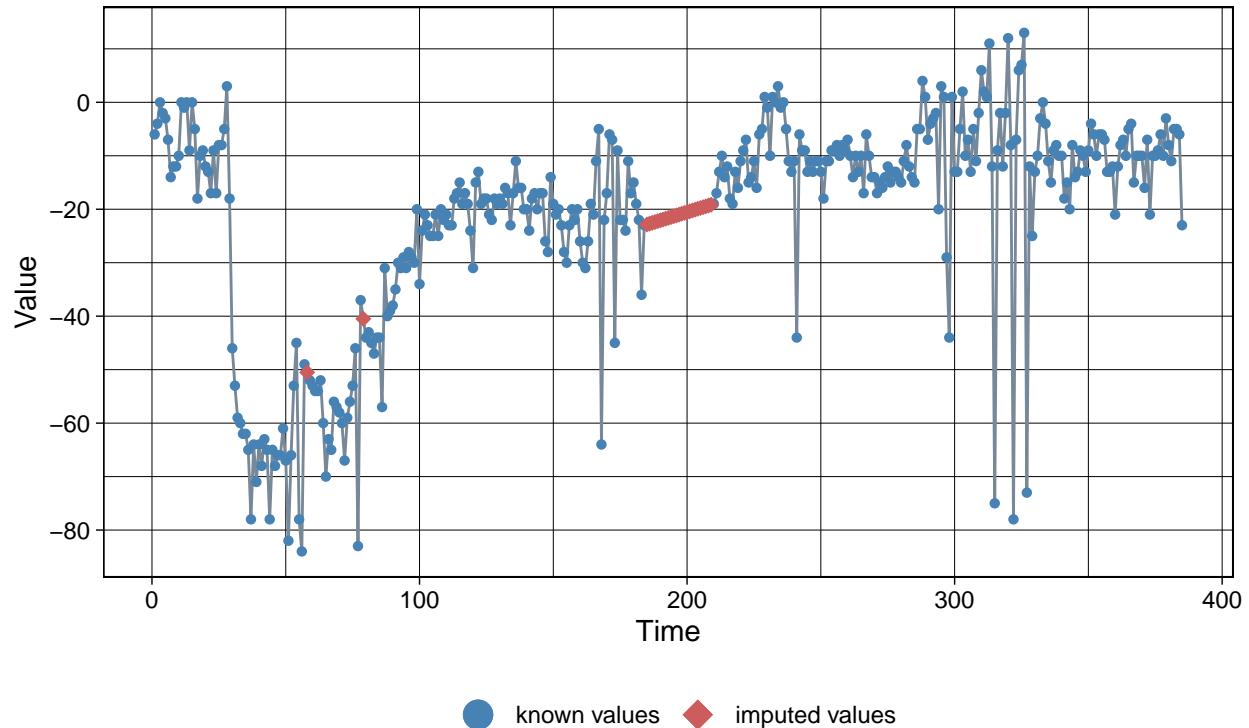


```
# Convert dataframe to ts object
Google_t_grocery_ts<-xts(Google_t_grocery[-1],Google_t_grocery$date)

# Impute the missing values with na_seadec (i.e Ceuta)
imp6 <- na_seadec(Google_t_grocery_ts[,16])
ggplot_na_imputations(Google_t_grocery_ts[,16], imp6)
```

Imputed Values

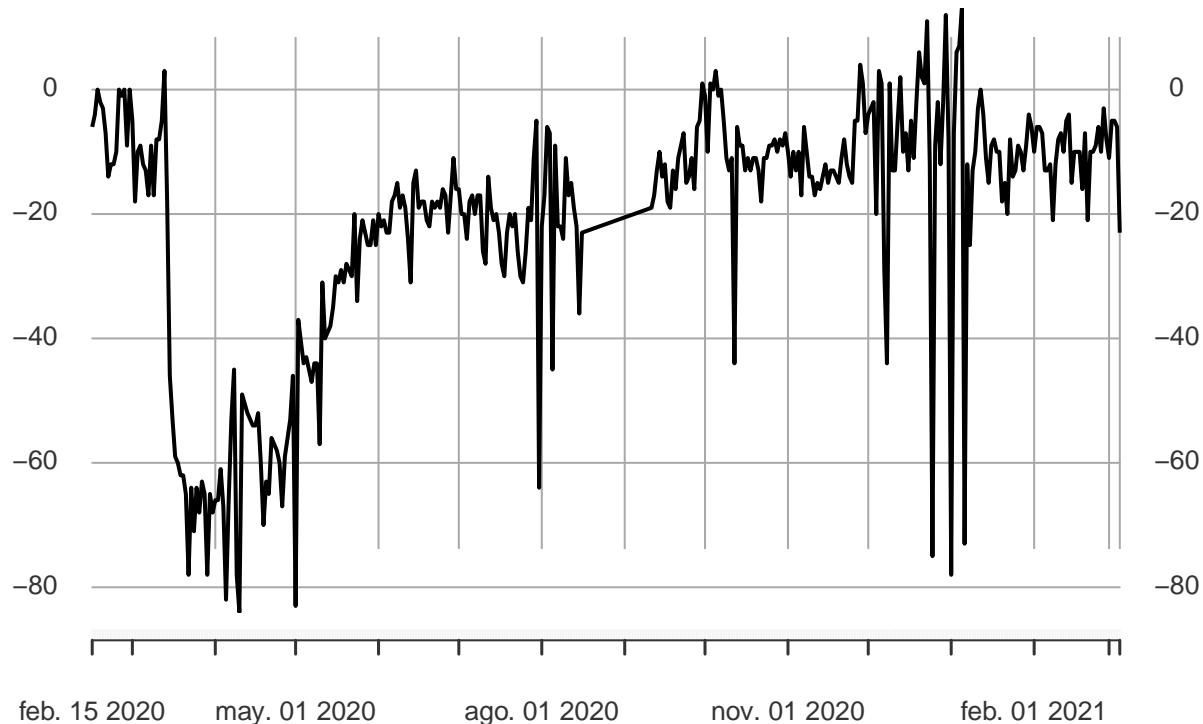
Visualization of missing value replacements



```
# We select na_seadec for the dataset
Google_t_grocery_ts <- na_seadec(Google_t_grocery_ts)
plot(Google_t_grocery_ts[,16])
```

Google_t_grocery_ts[, 16]

2020-02-15 / 2021-03-05

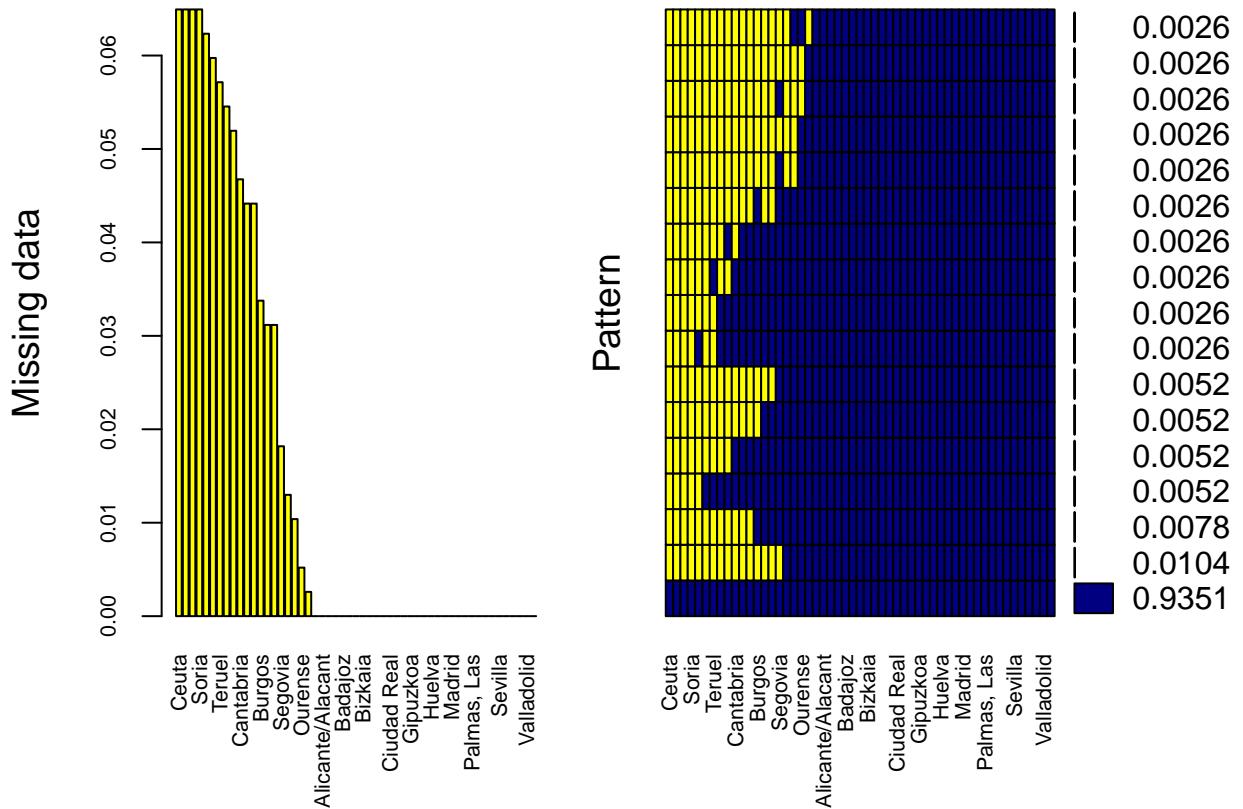


```
# We convert the time series object to a dataframe
Google_grocery <- ts_df(Google_t_grocery_ts)

names(Google_grocery)[names(Google_grocery) == "id"] <- "sub_region_2"
names(Google_grocery)[names(Google_grocery) == "time"] <- "Date"
names(Google_grocery)[names(Google_grocery) == "value"] <-
  "grocery_and_pharmacy_percent_change_from_baseline"

#####
# Transpose dataframe
Google_parks<-Google[,c(2,4,7)]
Google_t_parks<-dcast(Google_parks, date~sub_region_2, fill=NA)

# Visualize missing values
aggr(Google_t_parks, col=c('navyblue','yellow'),
  numbers=TRUE, sortVars=TRUE,
  labels=names(Google_t_parks), cex.axis=.7,
  gap=3, ylab=c("Missing data","Pattern"))
```



```
##  
## Variables sorted by number of missings:  
##           Variable      Count  
##             Ceuta 0.064935065  
##            Melilla 0.064935065  
##            Palencia 0.064935065  
##              Soria 0.064935065  
##            Cuenca 0.062337662  
##             Ávila 0.059740260  
##            Teruel 0.057142857  
##            Huesca 0.054545455  
##            Zamora 0.051948052  
##            Cantabria 0.046753247  
##             Lleida 0.044155844  
##            Rioja, La 0.044155844  
##            Burgos 0.033766234  
##            Guadalajara 0.031168831  
##              León 0.031168831  
##            Segovia 0.018181818  
##            Albacete 0.012987013  
##            Navarra 0.010389610  
##            Ourense 0.005194805  
##            Asturias 0.002597403  
##            date 0.000000000  
##            Alicante/Alacant 0.000000000  
##            Almería 0.000000000
```

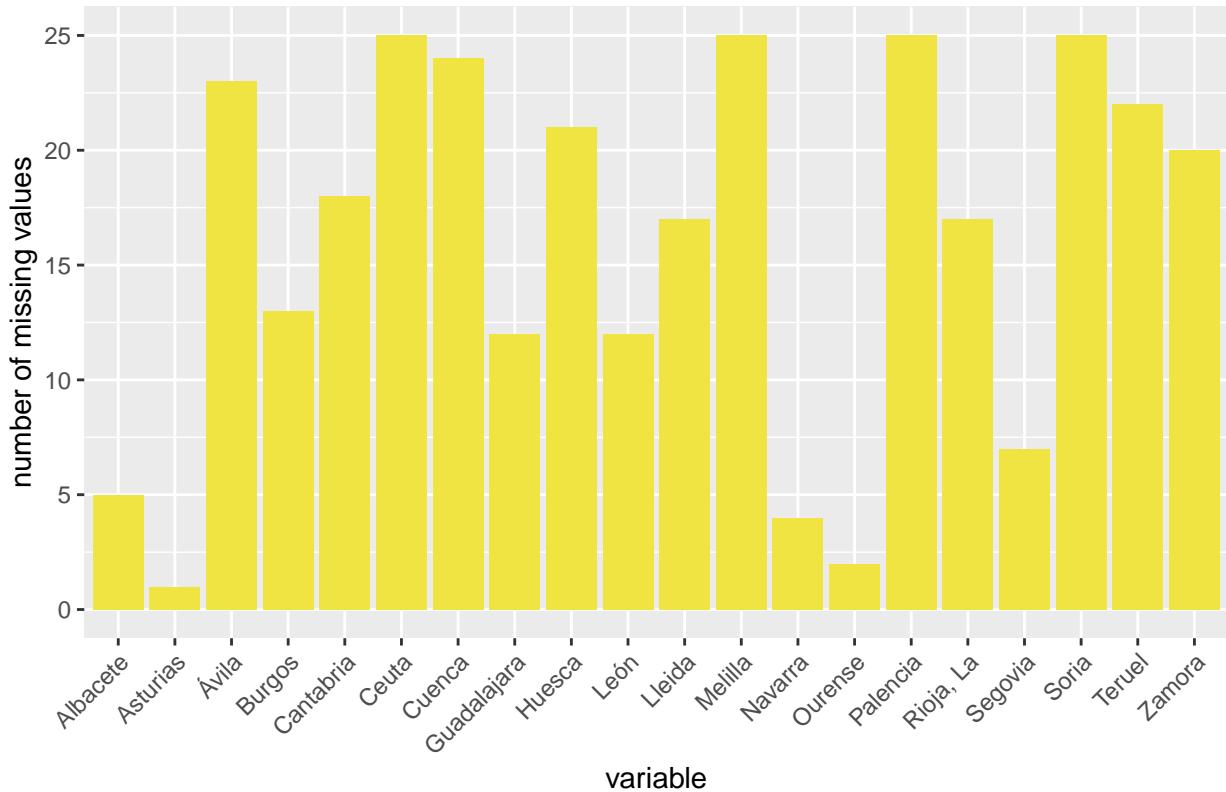
```

##          Araba/Álava 0.000000000
##          Badajoz 0.000000000
##          Balears, Illes 0.000000000
##          Barcelona 0.000000000
##          Bizkaia 0.000000000
##          Cáceres 0.000000000
##          Cádiz 0.000000000
##          Castellón/Castelló 0.000000000
##          Ciudad Real 0.000000000
##          Córdoba 0.000000000
##          Coruña, A 0.000000000
##          Gipuzkoa 0.000000000
##          Girona 0.000000000
##          Granada 0.000000000
##          Huelva 0.000000000
##          Jaén 0.000000000
##          Lugo 0.000000000
##          Madrid 0.000000000
##          Málaga 0.000000000
##          Murcia 0.000000000
##          Palmas, Las 0.000000000
##          Pontevedra 0.000000000
##          Salamanca 0.000000000
## Santa Cruz de Tenerife 0.000000000
##          Sevilla 0.000000000
##          Tarragona 0.000000000
##          Toledo 0.000000000
## Valencia/València 0.000000000
##          Valladolid 0.000000000
##          Zaragoza 0.000000000

Google_t_parks %>%
  gather(key = "key", value = "val") %>%
  mutate(is.missing = is.na(val)) %>%
  group_by(key, is.missing) %>%
  summarise(num.missing = n()) %>%
  filter(is.missing==T) %>%
  select(-is.missing) %>%
  arrange(desc(num.missing)) %>%
  ggplot() +
  geom_bar(aes(x=key, y=num.missing), stat = 'identity', fill="#F0E442") +
  labs(x='variable', y="number of missing values",
       title='Number of missing values') +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

```

Number of missing values

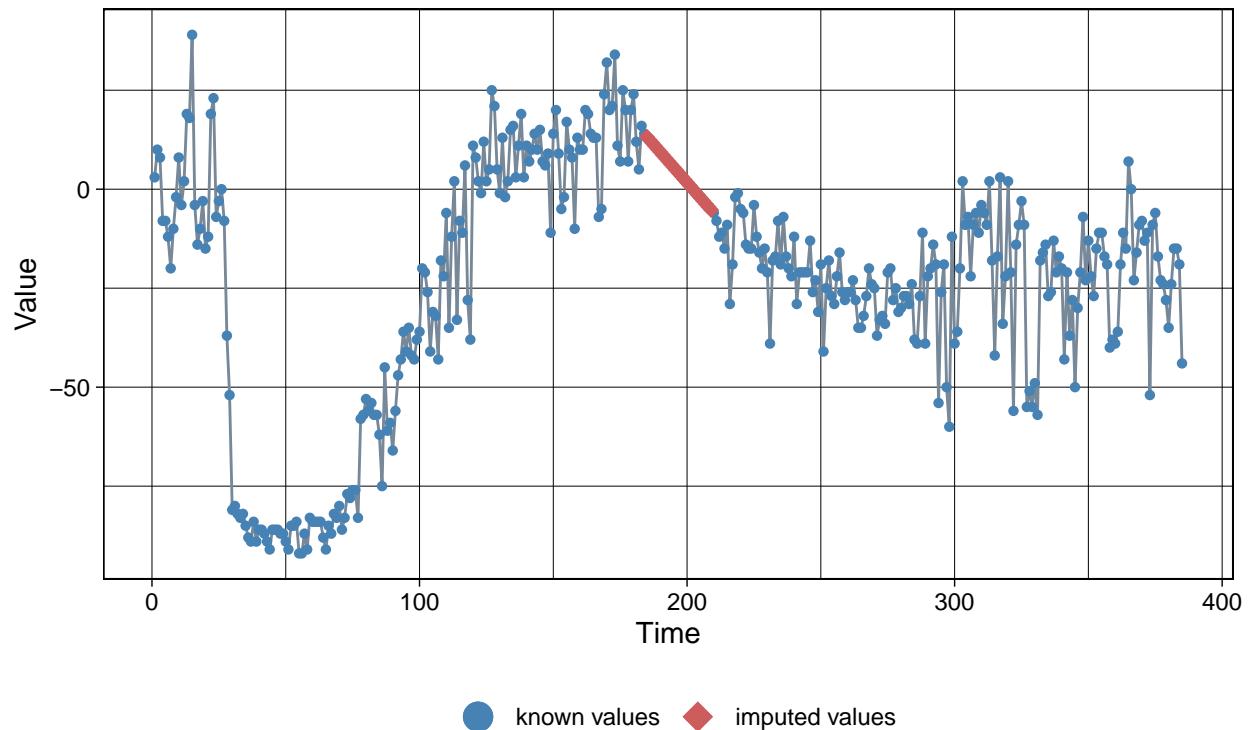


```
# Convert dataframe to ts object
Google_t_parks_ts<-xts(Google_t_parks[-1],Google_t_parks$date)

# Impute the missing values with na_seadec (i.e Ceuta)
imp7 <- na_seadec(Google_t_parks_ts[,16])
ggplot_na_imputations(Google_t_parks_ts[,16], imp7)
```

Imputed Values

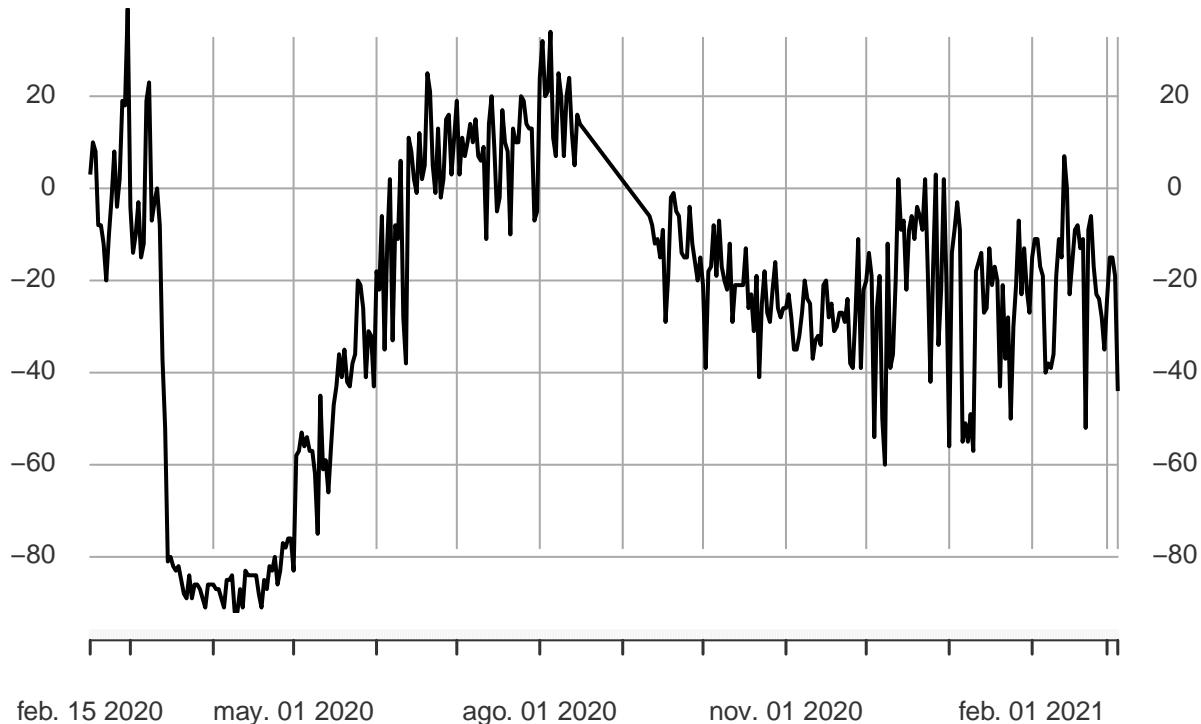
Visualization of missing value replacements



```
# We select na_seadec for the dataset
Google_t_parks_ts <- na_seadec(Google_t_parks_ts)
plot(Google_t_parks_ts[,16])
```

Google_t_parks_ts[, 16]

2020-02-15 / 2021-03-05

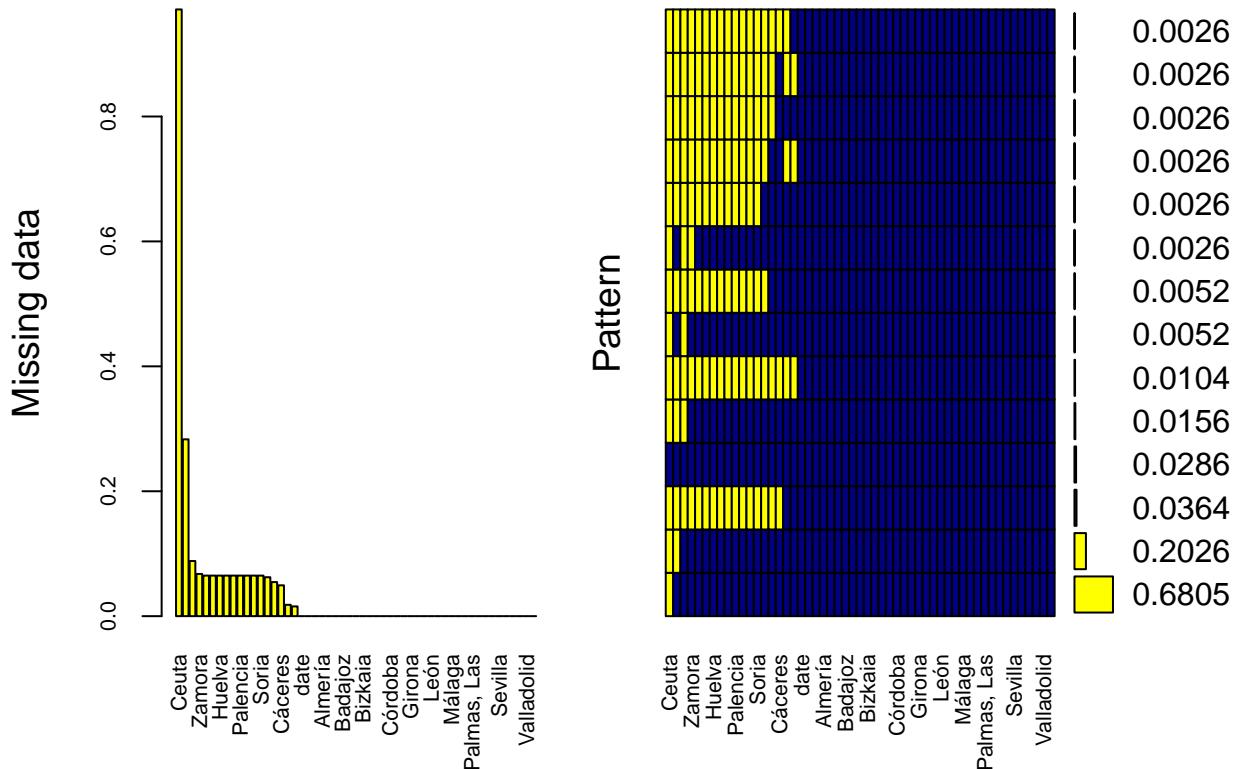


```
# We convert the time series object to a dataframe
Google_parks <- ts_df(Google_t_parks_ts)

names(Google_parks)[names(Google_parks) == "id"] <- "sub_region_2"
names(Google_parks)[names(Google_parks) == "time"] <- "Date"
names(Google_parks)[names(Google_parks) == "value"] <-
  "parks_percent_change_from_baseline"

#####
# Transpose dataframe
Google_transit<-Google[c(2,4,8)]
Google_t_transit<-dcast(Google_transit, date~sub_region_2, fill=NA)

# Visualize missing values
aggr(Google_t_transit, col=c('navyblue','yellow'),
  numbers=TRUE, sortVars=TRUE,
  labels=names(Google_t_transit), cex.axis=.7,
  gap=3, ylab=c("Missing data","Pattern"))
```



```
##  
##  Variables sorted by number of missings:  
##  
##          Variable      Count  
##              Ceuta 0.97142857  
##              Melilla 0.28311688  
##              Teruel 0.08831169  
##              Zamora 0.06753247  
##              Ávila 0.06493506  
##              Cuenca 0.06493506  
##              Huelva 0.06493506  
##              Huesca 0.06493506  
##              Lugo 0.06493506  
##              Palencia 0.06493506  
##              Rioja, La 0.06493506  
##              Segovia 0.06493506  
##              Soria 0.06493506  
##              Ourense 0.06233766  
##              Burgos 0.05454545  
##              Cáceres 0.04935065  
##              Ciudad Real 0.01818182  
##              Guadalajara 0.01558442  
##              date 0.00000000  
##              Albacete 0.00000000  
##              Alicante/Alacant 0.00000000  
##              Almería 0.00000000  
##              Araba/Álava 0.00000000
```

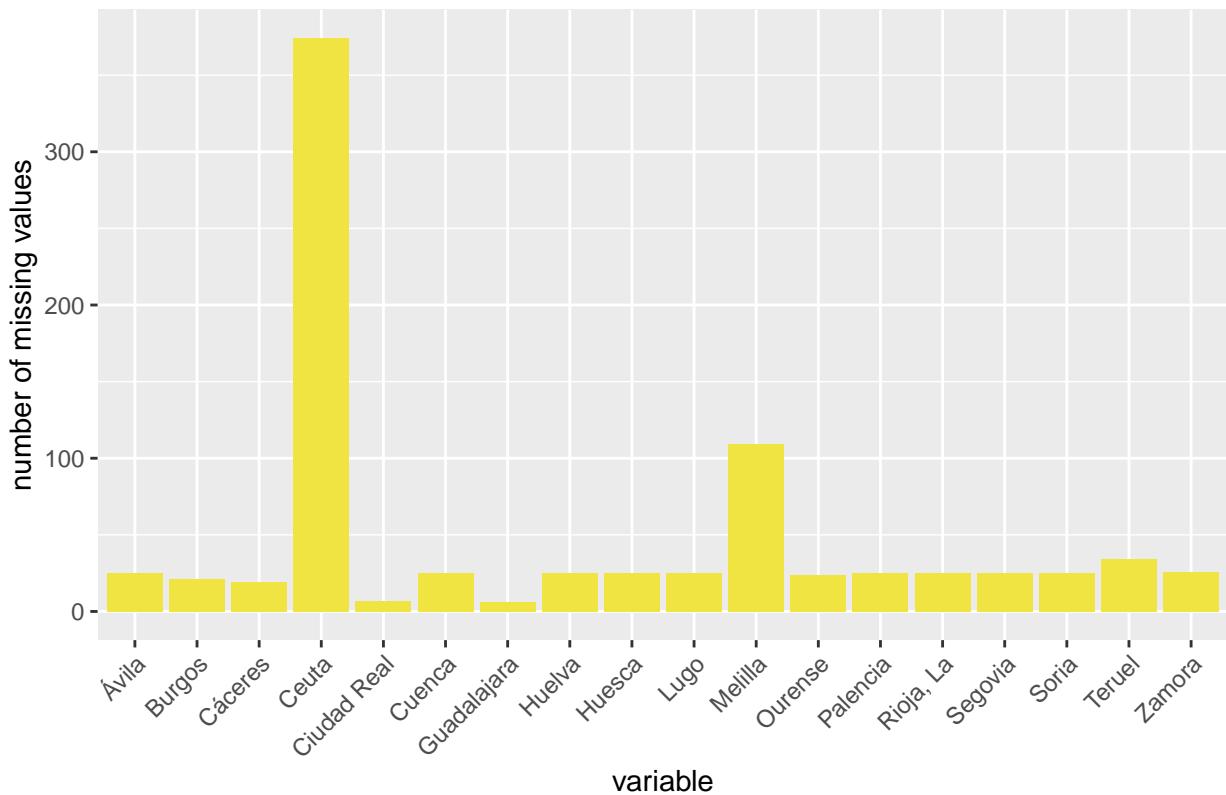
```

##          Asturias 0.00000000
##          Badajoz 0.00000000
##          Balears, Illes 0.00000000
##          Barcelona 0.00000000
##          Bizkaia 0.00000000
##          Cádiz 0.00000000
##          Cantabria 0.00000000
##          Castellón/Castelló 0.00000000
##          Córdoba 0.00000000
##          Coruña, A 0.00000000
##          Gipuzkoa 0.00000000
##          Girona 0.00000000
##          Granada 0.00000000
##          Jaén 0.00000000
##          León 0.00000000
##          Lleida 0.00000000
##          Madrid 0.00000000
##          Málaga 0.00000000
##          Murcia 0.00000000
##          Navarra 0.00000000
##          Palmas, Las 0.00000000
##          Pontevedra 0.00000000
##          Salamanca 0.00000000
## Santa Cruz de Tenerife 0.00000000
##          Sevilla 0.00000000
##          Tarragona 0.00000000
##          Toledo 0.00000000
## Valencia/València 0.00000000
##          Valladolid 0.00000000
##          Zaragoza 0.00000000

Google_t_transit %>%
  gather(key = "key", value = "val") %>%
  mutate(is.missing = is.na(val)) %>%
  group_by(key, is.missing) %>%
  summarise(num.missing = n()) %>%
  filter(is.missing==T) %>%
  select(-is.missing) %>%
  arrange(desc(num.missing)) %>%
  ggplot() +
  geom_bar(aes(x=key, y=num.missing), stat = 'identity', fill="#F0E442") +
  labs(x='variable', y="number of missing values",
       title='Number of missing values') +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

```

Number of missing values

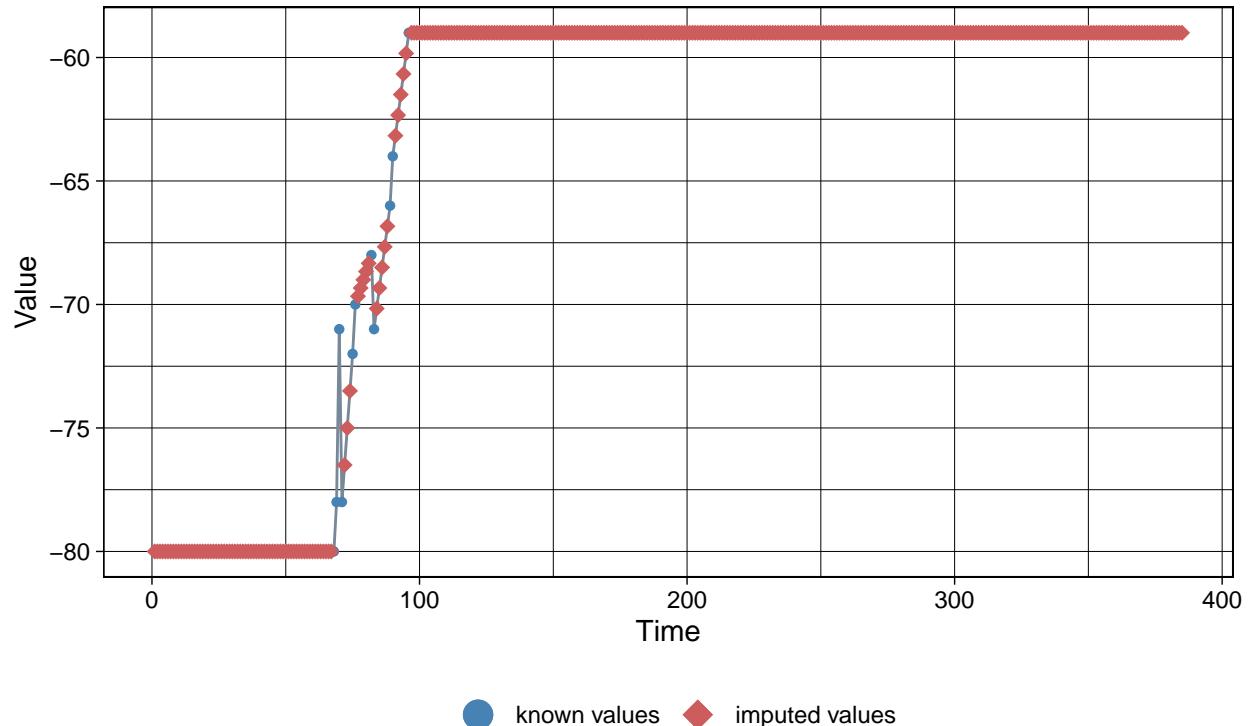


```
# Convert dataframe to ts object
Google_t_transit_ts<-xts(Google_t_transit[-1],Google_t_transit$date)

# Impute the missing values with na_seadec (i.e Ceuta)
imp8 <- na_seadec(Google_t_transit_ts[,16])
ggplot_na_imputations(Google_t_transit_ts[,16], imp8)
```

Imputed Values

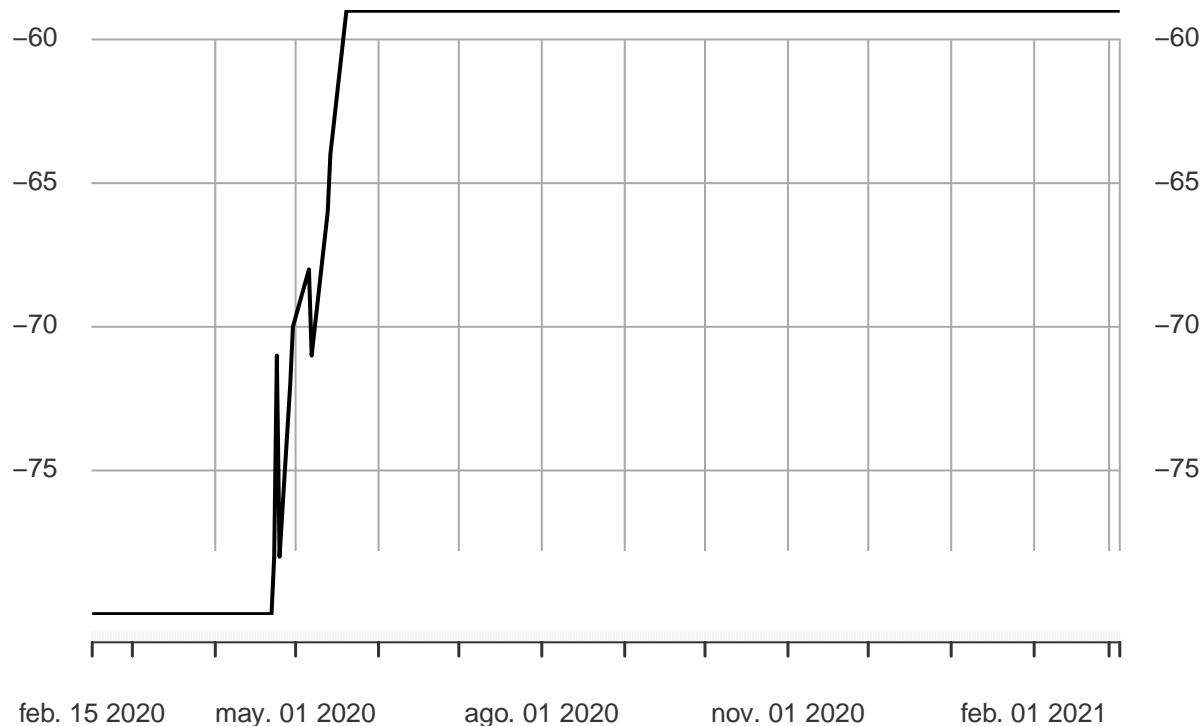
Visualization of missing value replacements



```
# We select na_seadec for the dataset
Google_t_transit_ts <- na_seadec(Google_t_transit_ts)
plot(Google_t_transit_ts[,16])
```

Google_t_transit_ts[, 16]

2020-02-15 / 2021-03-05

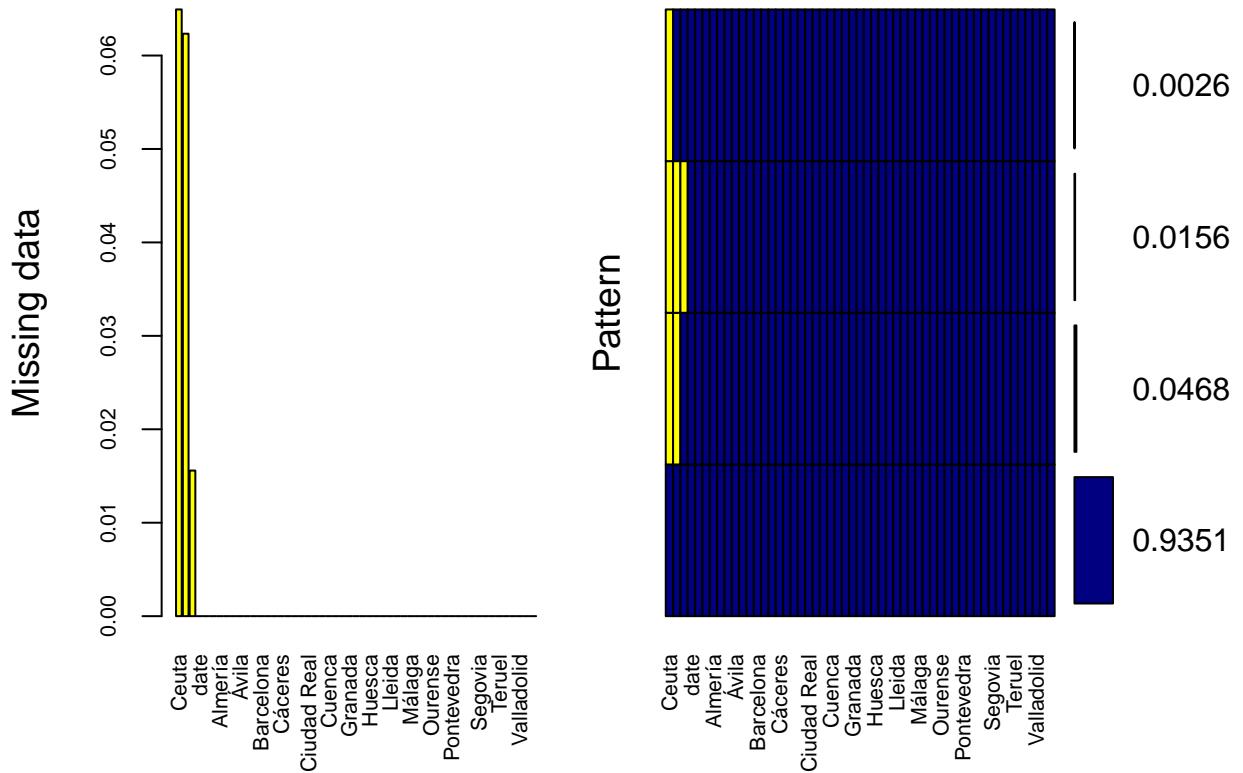


```
# We convert the time series object to a dataframe
Google_transit <- ts_df(Google_t_transit_ts)

names(Google_transit)[names(Google_transit) == "id"] <- "sub_region_2"
names(Google_transit)[names(Google_transit) == "time"] <- "Date"
names(Google_transit)[names(Google_transit) == "value"] <-
  "transit_stations_percent_change_from_baseline"

#####
# Transpose dataframe
Google_workplaces<-Google[c(2,4,9)]
Google_t_workplaces<-dcast(Google_workplaces, date~sub_region_2, fill=NA)

# Visualize missing values
aggr(Google_t_workplaces, col=c('navyblue','yellow'),
  numbers=TRUE, sortVars=TRUE,
  labels=names(Google_t_workplaces), cex.axis=.7,
  gap=3, ylab=c("Missing data","Pattern"))
```



```

## 
##  Variables sorted by number of missings:
##          Variable      Count
##            Ceuta 0.06493506
##            Melilla 0.06233766
##            Soria 0.01558442
##            date 0.00000000
##            Albacete 0.00000000
##            Alicante/Alacant 0.00000000
##            Almería 0.00000000
##            Araba/Álava 0.00000000
##            Asturias 0.00000000
##            Ávila 0.00000000
##            Badajoz 0.00000000
##            Balears, Illes 0.00000000
##            Barcelona 0.00000000
##            Bizkaia 0.00000000
##            Burgos 0.00000000
##            Cáceres 0.00000000
##            Cádiz 0.00000000
##            Cantabria 0.00000000
##            Castellón/Castelló 0.00000000
##            Ciudad Real 0.00000000
##            Córdoba 0.00000000
##            Coruña, A 0.00000000
##            Cuenca 0.00000000

```

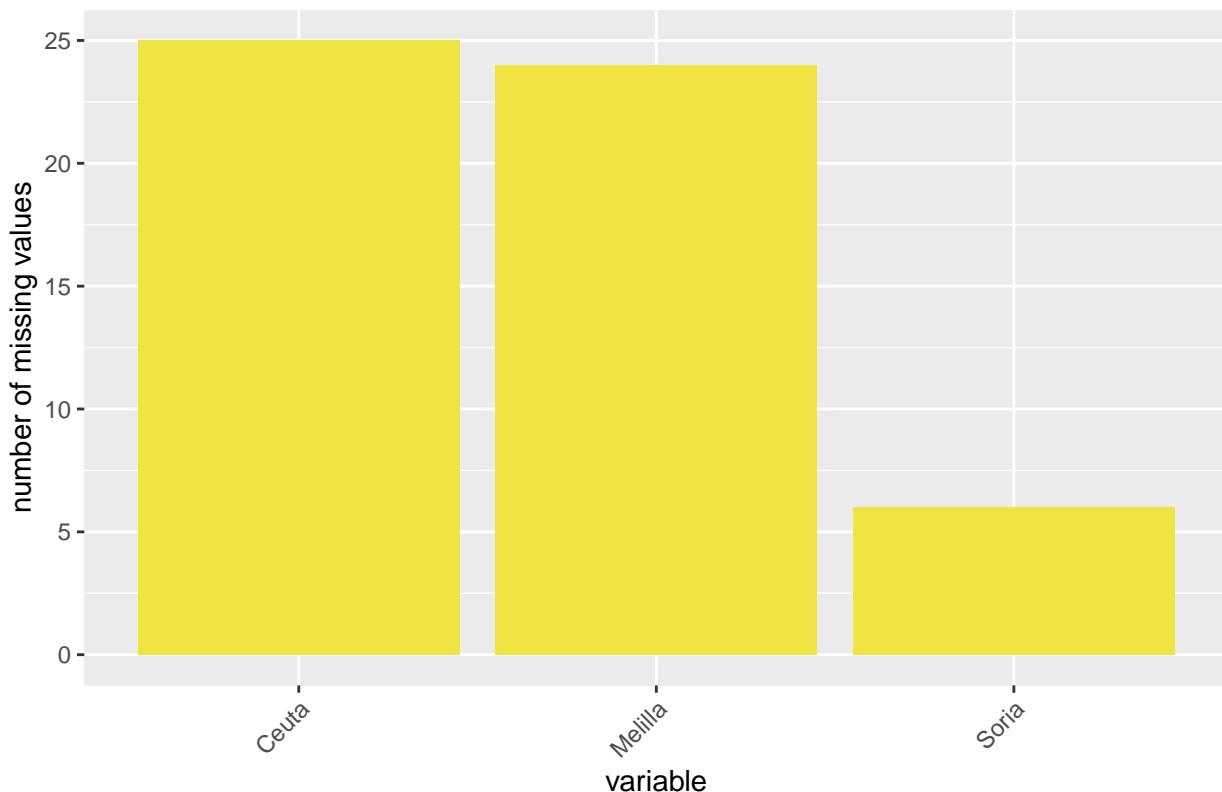
```

##          Gipuzkoa 0.00000000
##          Girona 0.00000000
##          Granada 0.00000000
##          Guadalajara 0.00000000
##          Huelva 0.00000000
##          Huesca 0.00000000
##          Jaén 0.00000000
##          León 0.00000000
##          Lleida 0.00000000
##          Lugo 0.00000000
##          Madrid 0.00000000
##          Málaga 0.00000000
##          Murcia 0.00000000
##          Navarra 0.00000000
##          Ourense 0.00000000
##          Palencia 0.00000000
##          Palmas, Las 0.00000000
##          Pontevedra 0.00000000
##          Rioja, La 0.00000000
##          Salamanca 0.00000000
## Santa Cruz de Tenerife 0.00000000
##          Segovia 0.00000000
##          Sevilla 0.00000000
##          Tarragona 0.00000000
##          Teruel 0.00000000
##          Toledo 0.00000000
## Valencia/València 0.00000000
##          Valladolid 0.00000000
##          Zamora 0.00000000
##          Zaragoza 0.00000000

Google_t_workplaces %>%
  gather(key = "key", value = "val") %>%
  mutate(is.missing = is.na(val)) %>%
  group_by(key, is.missing) %>%
  summarise(num.missing = n()) %>%
  filter(is.missing==T) %>%
  select(-is.missing) %>%
  arrange(desc(num.missing)) %>%
  ggplot() +
  geom_bar(aes(x=key, y=num.missing), stat = 'identity', fill="#F0E442") +
  labs(x='variable', y="number of missing values",
       title='Number of missing values') +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

```

Number of missing values

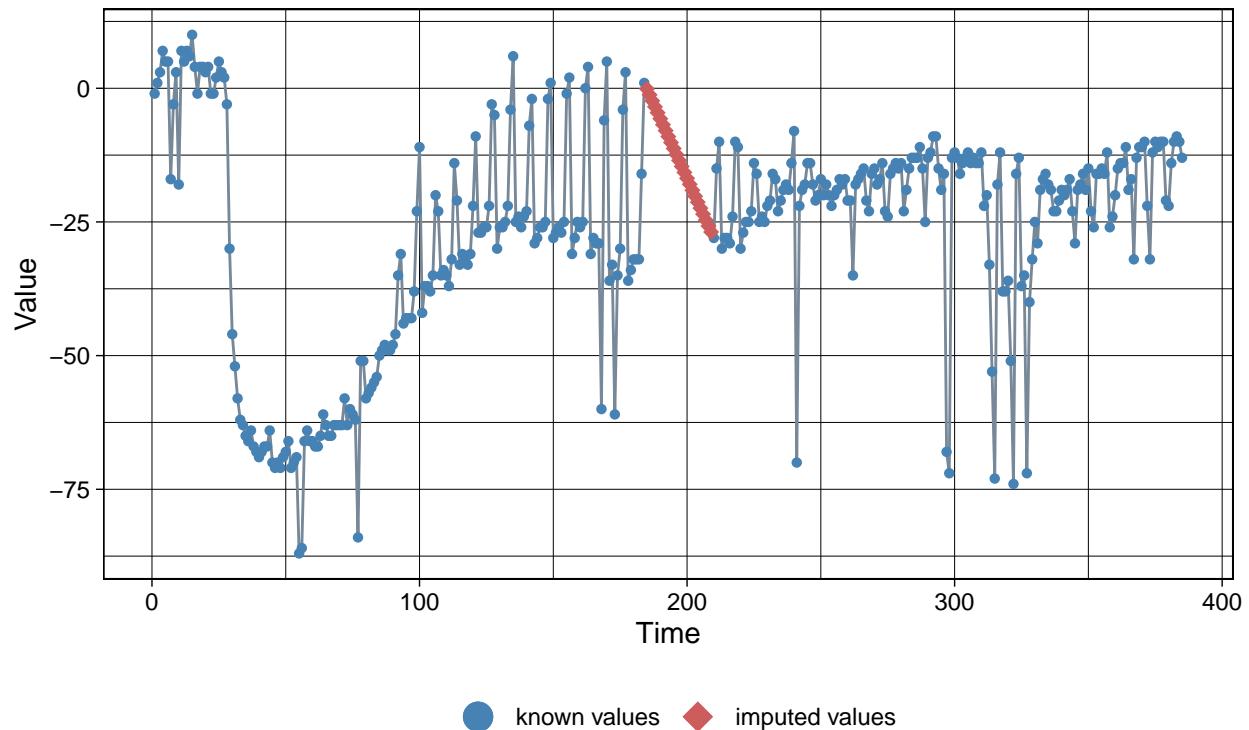


```
# Convert dataframe to ts object
Google_t_workplaces_ts<-xts(Google_t_workplaces[-1],Google_t_workplaces$date)

# Impute the missing values with na_seadec (i.e Ceuta)
imp9 <- na_seadec(Google_t_workplaces_ts[,16])
ggplot_na_imputations(Google_t_workplaces_ts[,16], imp9)
```

Imputed Values

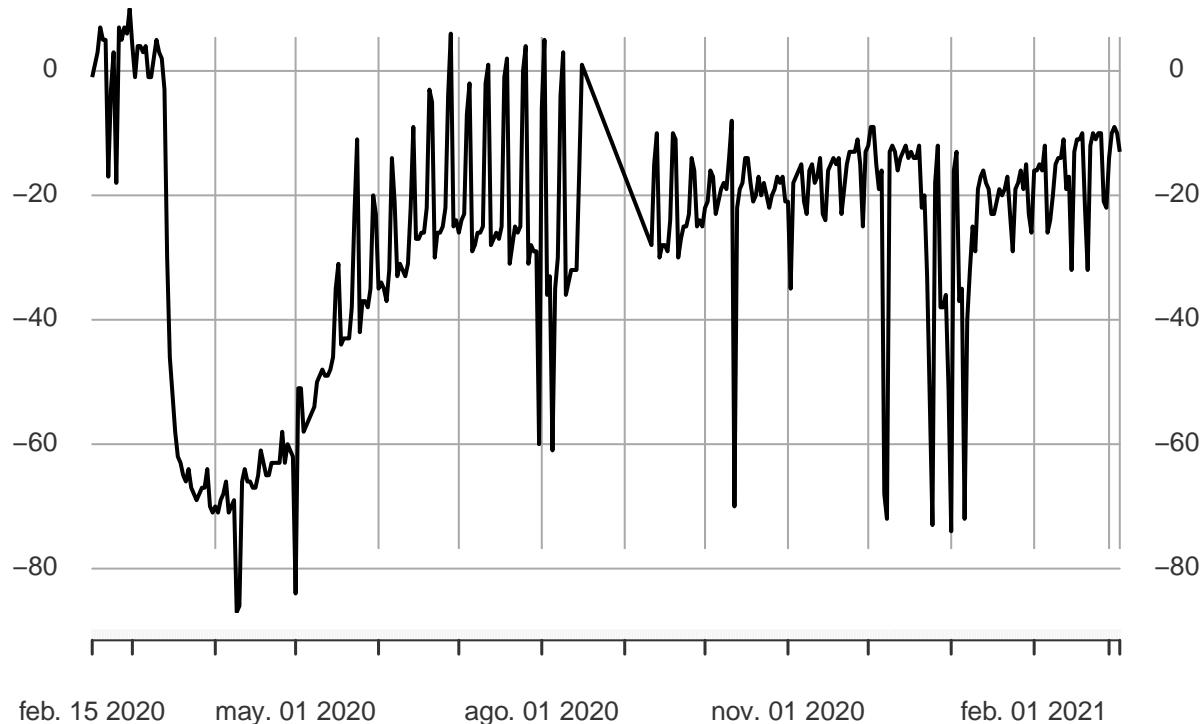
Visualization of missing value replacements



```
# We select na_seadec for the dataset
Google_t_workplaces_ts <- na_seadec(Google_t_workplaces_ts)
plot(Google_t_workplaces_ts[,16])
```

Google_t_workplaces_ts[, 16]

2020-02-15 / 2021-03-05

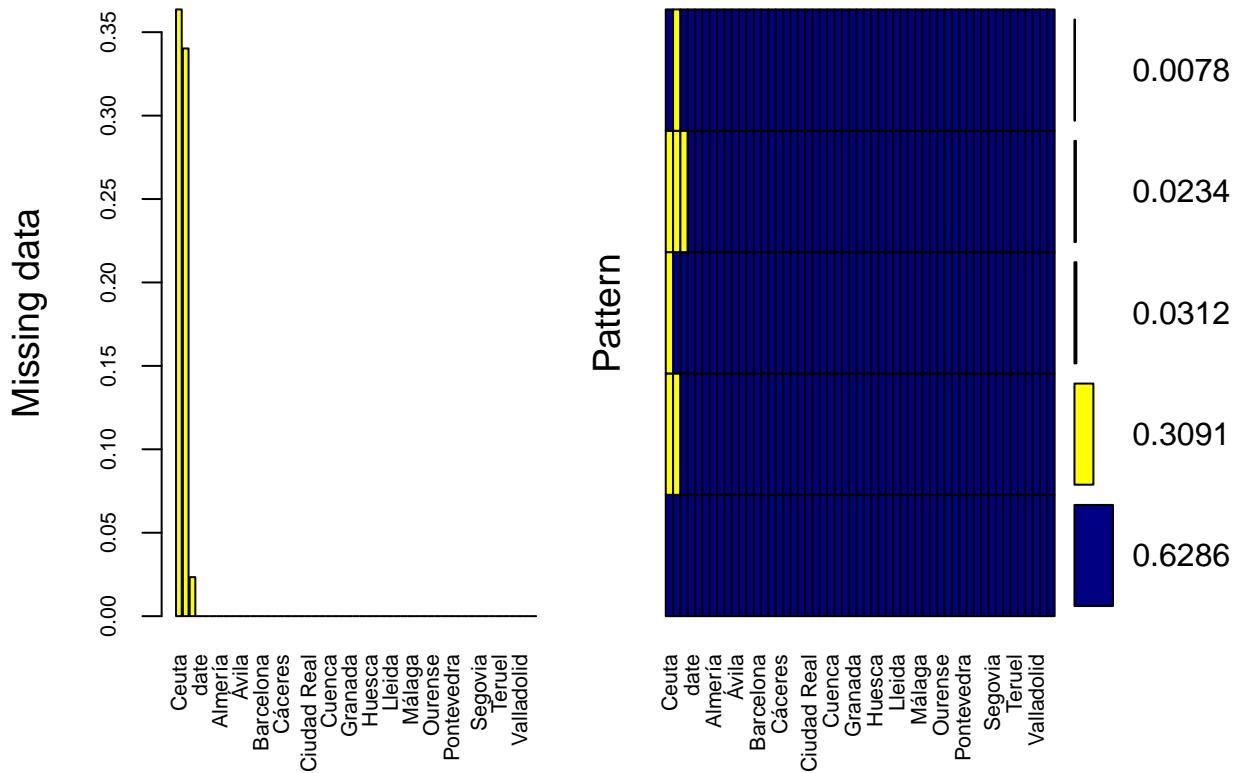


```
# We convert the time series object to a dataframe
Google_workplaces <- ts_df(Google_t_workplaces_ts)

names(Google_workplaces)[names(Google_workplaces) == "id"] <- "sub_region_2"
names(Google_workplaces)[names(Google_workplaces) == "time"] <- "Date"
names(Google_workplaces)[names(Google_workplaces) == "value"] <-
  "workplaces_percent_change_from_baseline"

#####
# Transpose dataframe
Google_residential<-Google[c(2,4,10)]
Google_t_residential<-dcast(Google_residential, date~sub_region_2, fill=NA)

# Visualize missing values
aggr(Google_t_residential, col=c('navyblue','yellow'),
  numbers=TRUE, sortVars=TRUE,
  labels=names(Google_t_residential), cex.axis=.7,
  gap=3, ylab=c("Missing data","Pattern"))
```



```
##  
##  Variables sorted by number of missings:  
##  
##          Variable      Count  
##            Ceuta 0.36363636  
##            Melilla 0.34025974  
##            Soria 0.02337662  
##            date 0.00000000  
##            Albacete 0.00000000  
##            Alicante/Alacant 0.00000000  
##            Almería 0.00000000  
##            Araba/Álava 0.00000000  
##            Asturias 0.00000000  
##            Ávila 0.00000000  
##            Badajoz 0.00000000  
##            Balears, Illes 0.00000000  
##            Barcelona 0.00000000  
##            Bizkaia 0.00000000  
##            Burgos 0.00000000  
##            Cáceres 0.00000000  
##            Cádiz 0.00000000  
##            Cantabria 0.00000000  
##            Castellón/Castelló 0.00000000  
##            Ciudad Real 0.00000000  
##            Córdoba 0.00000000  
##            Coruña, A 0.00000000  
##            Cuenca 0.00000000
```

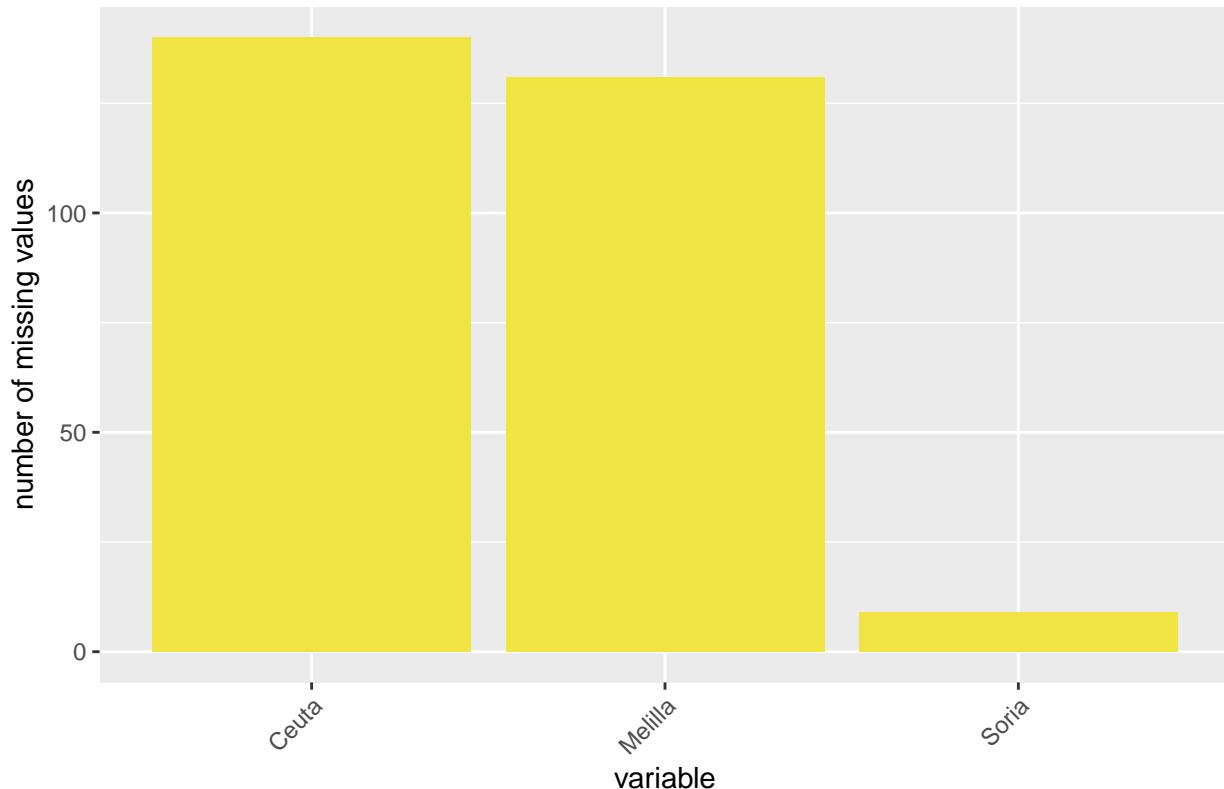
```

##          Gipuzkoa 0.00000000
##          Girona 0.00000000
##          Granada 0.00000000
##          Guadalajara 0.00000000
##          Huelva 0.00000000
##          Huesca 0.00000000
##          Jaén 0.00000000
##          León 0.00000000
##          Lleida 0.00000000
##          Lugo 0.00000000
##          Madrid 0.00000000
##          Málaga 0.00000000
##          Murcia 0.00000000
##          Navarra 0.00000000
##          Ourense 0.00000000
##          Palencia 0.00000000
##          Palmas, Las 0.00000000
##          Pontevedra 0.00000000
##          Rioja, La 0.00000000
##          Salamanca 0.00000000
## Santa Cruz de Tenerife 0.00000000
##          Segovia 0.00000000
##          Sevilla 0.00000000
##          Tarragona 0.00000000
##          Teruel 0.00000000
##          Toledo 0.00000000
## Valencia/València 0.00000000
##          Valladolid 0.00000000
##          Zamora 0.00000000
##          Zaragoza 0.00000000

Google_t_residential %>%
  gather(key = "key", value = "val") %>%
  mutate(is.missing = is.na(val)) %>%
  group_by(key, is.missing) %>%
  summarise(num.missing = n()) %>%
  filter(is.missing==T) %>%
  select(-is.missing) %>%
  arrange(desc(num.missing)) %>%
  ggplot() +
  geom_bar(aes(x=key, y=num.missing), stat = 'identity', fill="#F0E442") +
  labs(x='variable', y="number of missing values",
       title='Number of missing values') +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

```

Number of missing values

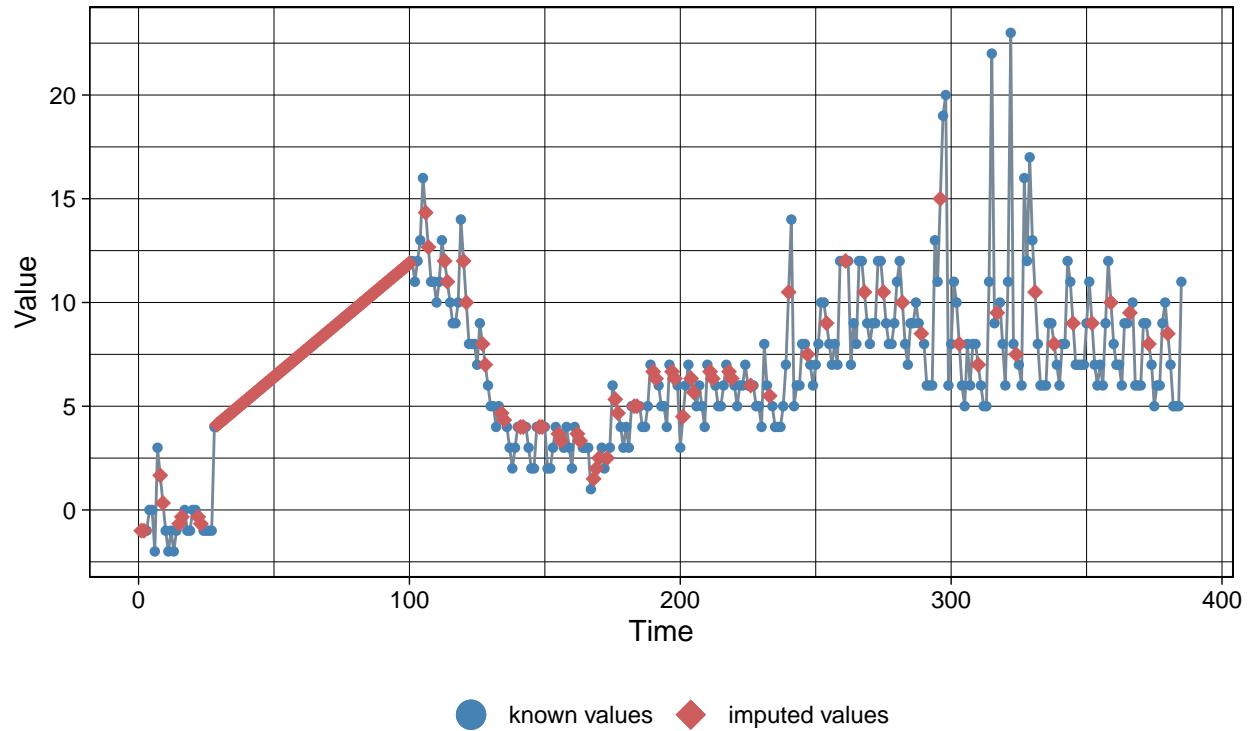


```
# Convert dataframe to ts object
Google_t_residential_ts<-xts(Google_t_residential[-1],Google_t_residential$date)

# Impute the missing values with na_seadec (i.e Ceuta)
imp10 <- na_seadec(Google_t_residential_ts[,16])
ggplot_na_imputations(Google_t_residential_ts[,16], imp10)
```

Imputed Values

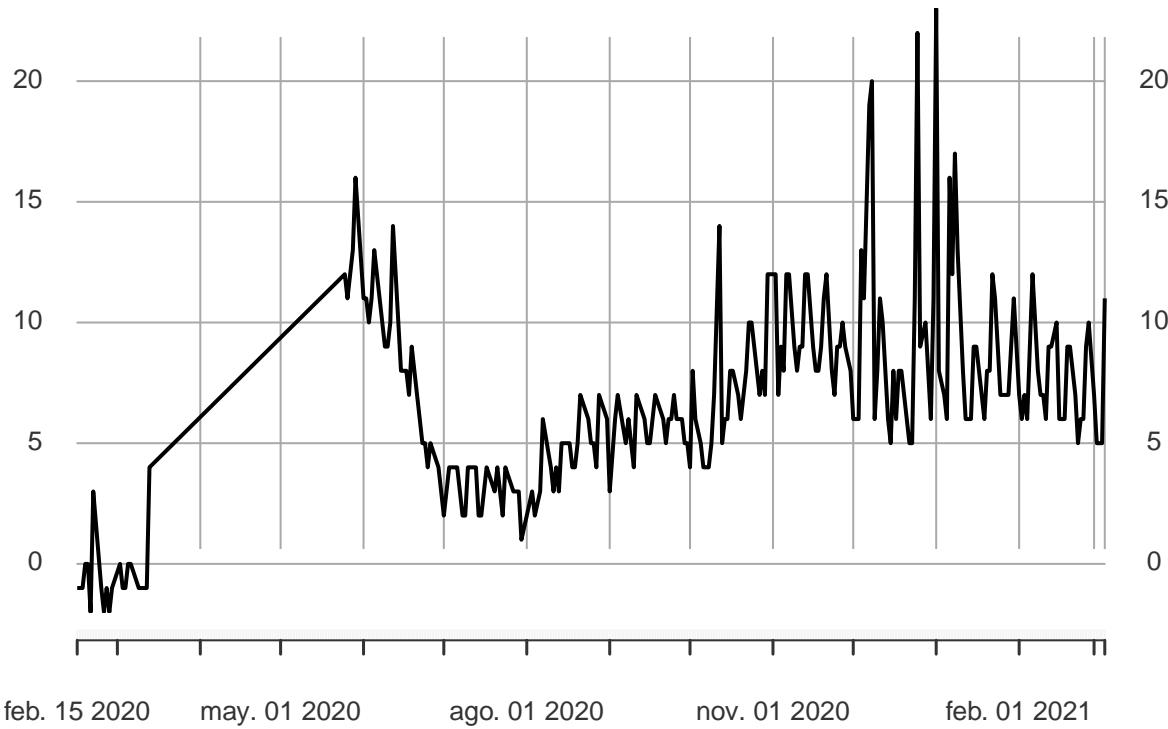
Visualization of missing value replacements



```
# We select na_seadec for the dataset
Google_t_residential_ts <- na_seadec(Google_t_residential_ts)
plot(Google_t_residential_ts[,16])
```

Google_t_residential_ts[, 16]

2020-02-15 / 2021-03-05



```
# We convert the time series object to a dataframe
Google_residential <- ts_df(Google_t_residential_ts)

names(Google_residential)[names(Google_residential) == "id"] <- "sub_region_2"
names(Google_residential)[names(Google_residential) == "time"] <- "Date"
names(Google_residential)[names(Google_residential) == "value"] <-
  "residential_percent_change_from_baseline"
```

Now we merge the previous dataframes into new one with the imputed values and we add the ISO code for the province.

```
# New dataframe Google_b
# This approach assumes that the column names are the same and that there's the same number of rows (the
# Any duplicated columns are automatically eliminated used in the merging process.
Google_b <- merge(Google_retail, Google_grocery) %>%
  merge(Google_parks) %>%
  merge(Google_transit) %>%
  merge(Google_workplaces) %>%
  merge(Google_residential)

# We add the iso code for the province
Google_b$iso_code <- NA
Google_b$iso_code<-Google[match(Google_b$sub_region_2, Google$sub_region_2),3]
rm("Google")
Google<-Google_b
rm("Google_b")
```

```

# Check table
head(Google, 5)

##   sub_region_2      Date retail_and_recreation_percent_change_from_baseline
## 1    Albacete 2020-02-15                               3
## 2    Albacete 2020-02-16                               5
## 3    Albacete 2020-02-17                             -2
## 4    Albacete 2020-02-18                             -3
## 5    Albacete 2020-02-19                               0
##   grocery_and_pharmacy_percent_change_from_baseline
## 1                                         -5
## 2                                         1
## 3                                         3
## 4                                         -1
## 5                                         1
##   parks_percent_change_from_baseline
## 1                               35
## 2                               40
## 3                                7
## 4                               -4
## 5                                7
##   transit_stations_percent_change_from_baseline
## 1                               13
## 2                               18
## 3                               20
## 4                                6
## 5                                9
##   workplaces_percent_change_from_baseline
## 1                                 1
## 2                                 0
## 3                                 5
## 4                                 4
## 5                                 4
##   residential_percent_change_from_baseline iso_code
## 1                               -3      AB
## 2                               -4      AB
## 3                               -1      AB
## 4                               -1      AB
## 5                               -1      AB

table(Google$sub_region_2)

##
##          Albacete      Alicante/Alacant      Almería
## 385                  385                      385
##      Araba/Álava      Asturias                Ávila
## 385                  385                      385
##      Badajoz        Balears, Illes      Barcelona
## 385                  385                      385
##      Bizkaia        Burgos                  Cáceres
## 385                  385                      385
##      Cádiz          Cantabria  Castellón/Castelló
## 385                  385                      385
##      Ceuta         Ciudad Real      Córdoba

```

```

##          385          385          385
## Coruña, A Cuenca Gipuzkoa
##          385          385          385
## Girona Granada Guadalajara
##          385          385          385
## Huelva Huesca Jaén
##          385          385          385
## León Lleida Lugo
##          385          385          385
## Madrid Málaga Melilla
##          385          385          385
## Murcia Navarra Ourense
##          385          385          385
## Palencia Palmas, Las Pontevedra
##          385          385          385
## Rioja, La Salamanca Santa Cruz de Tenerife
##          385          385          385
## Segovia Sevilla Soria
##          385          385          385
## Tarragona Teruel Toledo
##          385          385          385
## Valencia/València Valladolid Zamora
##          385          385          385
## Zaragoza            385
##          385

table(Google$iso_code)

##
##   A  AB  AL  AV  B  BA  BI  BU  C  CA  CC  CE  CO  CR  CS  CU  GC  GI  GR  GU
## 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385
##   H  HU  J  L  LE  LO  LU  M  MA  ML  MU  NA  O  OR  P  PM  PO  S  SA  SE
## 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385
##   SG  SO  SS  T  TE  TF  TO  V  VA  VI  Z  ZA
## 385 385 385 385 385 385 385 385 385 385 385 385

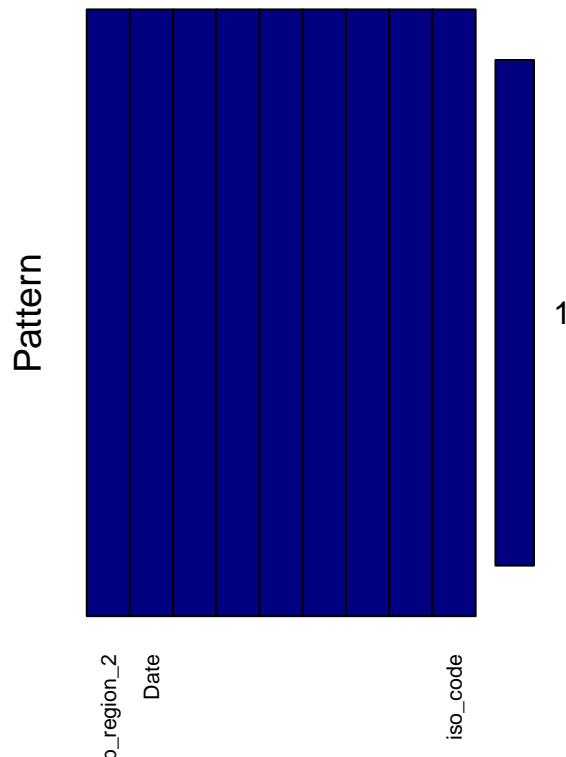
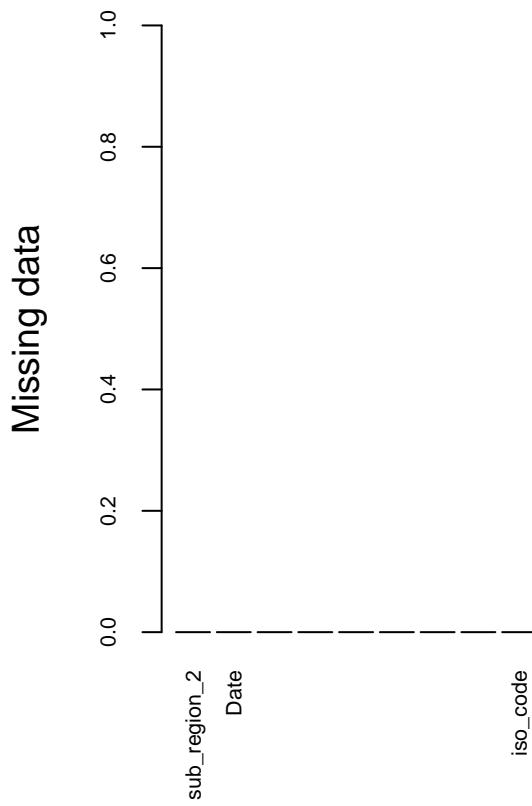
```

We check missing values. We should obtain zero missing values.

```

aggr(Google, col=c('navyblue','yellow'),
      numbers=TRUE, sortVars=TRUE,
      labels=names(Google), cex.axis=.7,
      gap=3, ylab=c("Missing data", "Pattern"))

```



```

## ## Variables sorted by number of missings:
##                                     Variable Count
##                               sub_region_2      0
##                               Date          0
## retail_and_recreation_percent_change_from_baseline      0
## grocery_and_pharmacy_percent_change_from_baseline      0
## parks_percent_change_from_baseline                      0
## transit_stations_percent_change_from_baseline          0
## workplaces_percent_change_from_baseline                 0
## residential_percent_change_from_baseline                0
##                               iso_code      0

Google %>%
  gather(key = "key", value = "val") %>%
  mutate(is.missing = is.na(val)) %>%
  group_by(key, is.missing) %>%
  summarise(num.missing = n()) %>%
  filter(is.missing==T) %>%
  select(-is.missing) %>%
  arrange(desc(num.missing)) %>%
  ggplot() +
  geom_bar(aes(x=key, y=num.missing), stat = 'identity', fill="#F0E442") +
  labs(x='variable', y="number of missing values",
       title='Number of missing values') +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

```

Number of missing values

number of missing values

variable

2.1.9 CNE review

The CSV files are provided per “imputed date” (fecha):

- **cases_technic_province.csv** - Number of cases by diagnostic technique and province (of residence)
- **cases_hosp_uci_def_sexo_edad_provres.csv** - Number of hospitalizations, number of ICU admissions and number of deaths by sex, age and province of residence.

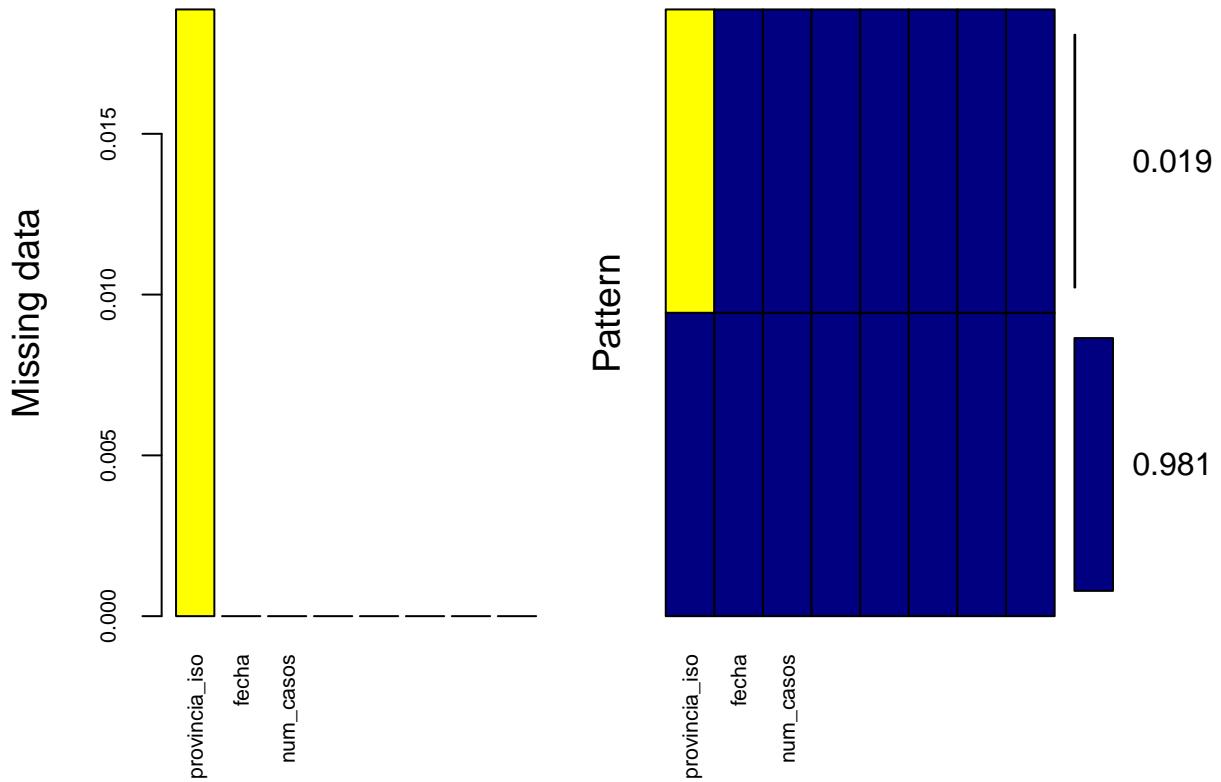
```
summary(CNE_tecnica)
```

```
## provincia_iso         fecha          num_casos      num_casos_prueba_pcr
## Length:23426        Length:23426      Min.   : 0.0    Min.   : 0.0
## Class :character     Class :character    1st Qu.: 2.0    1st Qu.: 2.0
## Mode  :character     Mode  :character    Median : 32.0   Median : 26.0
##                           Mean   : 136.9   Mean   : 109.6
##                           3rd Qu.: 120.0   3rd Qu.: 100.0
##                           Max.  :6972.0   Max.  :6546.0
## num_casos_prueba_test_ac num_casos_prueba_ag num_casos_prueba_elisa
## Min.   : 0.0000       Min.   : 0.00       Min.   : 0.0000
## 1st Qu.: 0.0000       1st Qu.: 0.00       1st Qu.: 0.0000
## Median : 0.0000       Median : 0.00       Median : 0.0000
## Mean   : 0.2037       Mean   : 26.21      Mean   : 0.1602
## 3rd Qu.: 0.0000       3rd Qu.: 9.00       3rd Qu.: 0.0000
## Max.  :32.0000       Max.  :3267.00      Max.  :71.0000
## num_casos_prueba_desconocida
## Min.   : 0.0000
## 1st Qu.: 0.0000
```

2.1.10 CNE review missing values & impute

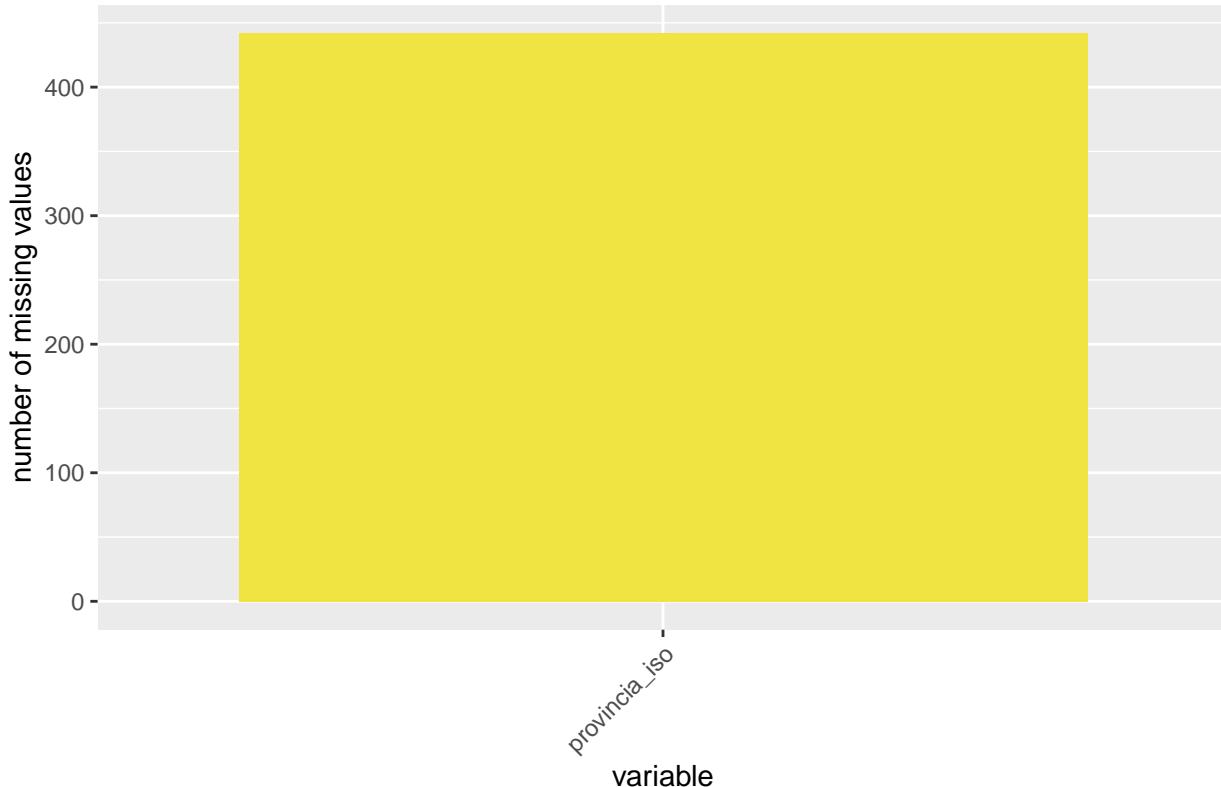
We check missing values for CNE_tecnica. In this case we omit the NA values.

```
agg(CNE_tecnica, col=c('navyblue','yellow'),  
    numbers=TRUE, sortVars=TRUE,  
    labels=names(CNE_tecnica), cex.axis=.7,  
    gap=3, ylab=c("Missing data","Pattern"))
```



```
## ## Variables sorted by number of missings:
##           Variable      Count
##           provincia_iso 0.01886792
##           fecha 0.00000000
##           num_casos 0.00000000
##           num_casos_prueba_pcr 0.00000000
##           num_casos_prueba_test_ac 0.00000000
##           num_casos_prueba_ag 0.00000000
##           num_casos_prueba_elisa 0.00000000
##           num_casos_prueba_desconocida 0.00000000
CNE_tecnica %>%
  gather(key = "key", value = "val") %>%
  mutate(is.missing = is.na(val)) %>%
  group_by(key, is.missing) %>%
  summarise(num.missing = n()) %>%
  filter(is.missing==T) %>%
  select(-is.missing) %>%
  arrange(desc(num.missing)) %>%
  ggplot() +
  geom_bar(aes(x=key, y=num.missing), stat = 'identity', fill="#F0E442") +
  labs(x='variable', y="number of missing values",
       title='Number of missing values') +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

Number of missing values

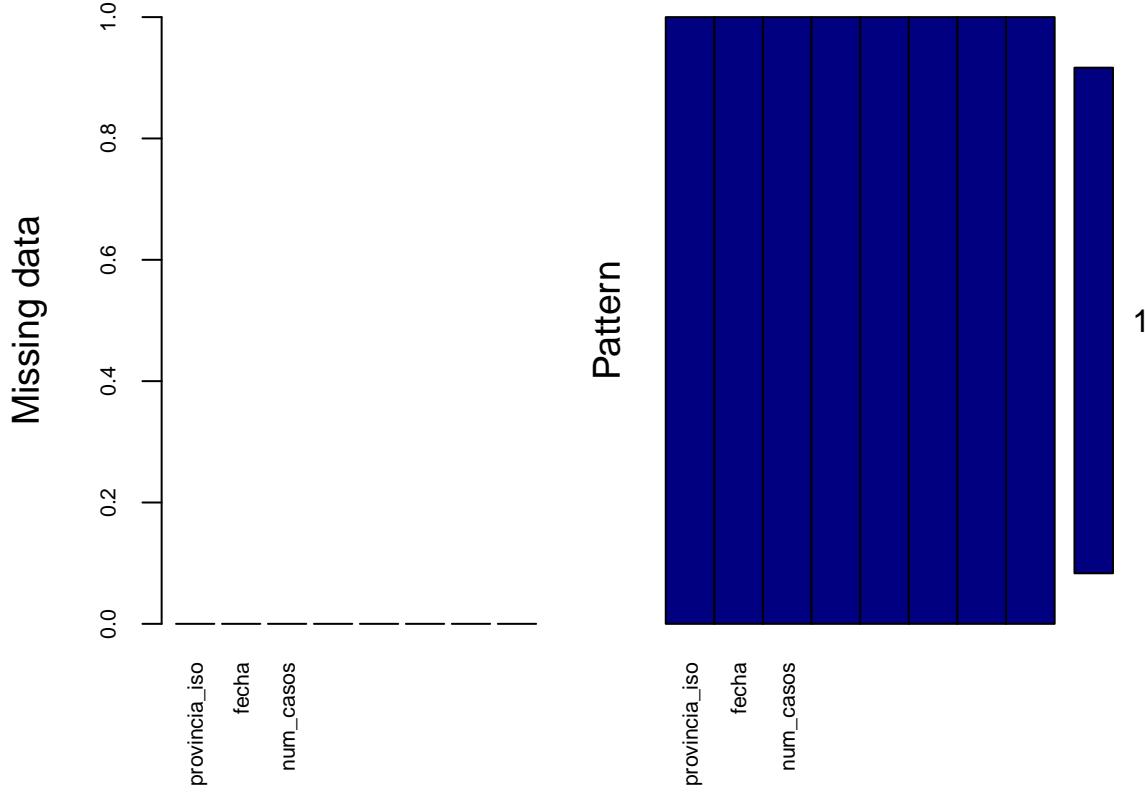


```
theme(axis.text.x = element_text(angle = 45, hjust = 1))

## List of 1
## $ axis.text.x:List of 11
##   ..$ family      : NULL
##   ..$ face        : NULL
##   ..$ colour      : NULL
##   ..$ size        : NULL
##   ..$ hjust       : num 1
##   ..$ vjust       : NULL
##   ..$ angle       : num 45
##   ..$ lineheight  : NULL
##   ..$ margin      : NULL
##   ..$ debug       : NULL
##   ..$ inherit.blank: logi FALSE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##   - attr(*, "class")= chr [1:2] "theme" "gg"
##   - attr(*, "complete")= logi FALSE
##   - attr(*, "validate")= logi TRUE
#####
CNE_tecnica <- na.omit(CNE_tecnica)
#####

aggr(CNE_tecnica, col=c('navyblue','yellow'),
      numbers=TRUE, sortVars=TRUE,
      labels=names(CNE_tecnica), cex.axis=.7,
```

```
gap=3, ylab=c("Missing data","Pattern"))
```



```
##
##  Variables sorted by number of missings:
##          Variable Count
##      provincia_iso      0
##          fecha          0
##      num_casos          0
##      num_casos_prueba_pcr  0
##      num_casos_prueba_test_ac  0
##      num_casos_prueba_ag    0
##      num_casos_prueba_elisa  0
##  num_casos_prueba_desconocida  0

CNE_tecnica %>%
  gather(key = "key", value = "val") %>%
  mutate(is.missing = is.na(val)) %>%
  group_by(key, is.missing) %>%
  summarise(num.missing = n()) %>%
  filter(is.missing==T) %>%
  select(-is.missing) %>%
  arrange(desc(num.missing)) %>%
  ggplot() +
  geom_bar(aes(x=key, y=num.missing), stat = 'identity', fill="#FOE442") +
  labs(x='variable', y="number of missing values",
       title='Number of missing values') +
```

```
theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

Number of missing values

number of missing values

variable

```
summary(CNE_casos)
```

```
##  provincia_iso           sexo          grupo_edad        fecha
##  Length:702780    Length:702780    Length:702780    Length:702780
##  Class :character  Class :character  Class :character  Class :character
##  Mode  :character  Mode  :character  Mode  :character  Mode  :character
##
##  num_casos         num_hosp        num_uci        num_def
##  Min.   : 0.000   Min.   : 0.0000   Min.   : 0.00000   Min.   : 0.0000
##  1st Qu.: 0.000   1st Qu.: 0.0000   1st Qu.: 0.00000   1st Qu.: 0.0000
##  Median : 0.000   Median : 0.0000   Median : 0.00000   Median : 0.0000
##  Mean   : 4.562   Mean   : 0.4611   Mean   : 0.04117   Mean   : 0.1036
##  3rd Qu.: 2.000   3rd Qu.: 0.0000   3rd Qu.: 0.00000   3rd Qu.: 0.0000
##  Max.   :771.000  Max.   :269.0000  Max.   :35.00000   Max.   :100.0000
```

```
head(str(CNE_casos,vec.len=3))
```

```
## 'data.frame': 702780 obs. of 8 variables:
## $ provincia_iso: chr "A" "A" "A" ...
## $ sexo          : chr "H" "H" "H" ...
## $ grupo_edad   : chr "0-9" "10-19" "20-29" ...
## $ fecha         : chr "2020-01-01" "2020-01-01" "2020-01-01" ...
## $ num_casos     : int 0 0 0 0 0 0 0 0 ...
```

```

## $ num_hosp      : int  0 0 0 0 0 0 0 0 ...
## $ num_uci       : int  0 0 0 0 0 0 0 0 ...
## $ num_def       : int  0 0 0 0 0 0 0 0 ...
## NULL
table(CNE_casos$provincia_iso)

##
##   A    AB    AL    AV    B    BA    BI    BU    C    CA    CC    CE    CO
## 13260 13260 13260 13260 13260 13260 13260 13260 13260 13260 13260 13260 13260 13260
##   CR    CS    CU    GC    GI    GR    GU    H    HU    J    L    LE    LO
## 13260 13260 13260 13260 13260 13260 13260 13260 13260 13260 13260 13260 13260 13260
##   LU    M    MA    ML    MU    NC    O    OR    P    PM    PO    S    SA
## 13260 13260 13260 13260 13260 13260 13260 13260 13260 13260 13260 13260 13260 13260
##   SE    SG    SO    SS    T    TE    TF    TO    V    VA    VI    Z    ZA
## 13260 13260 13260 13260 13260 13260 13260 13260 13260 13260 13260 13260 13260 13260

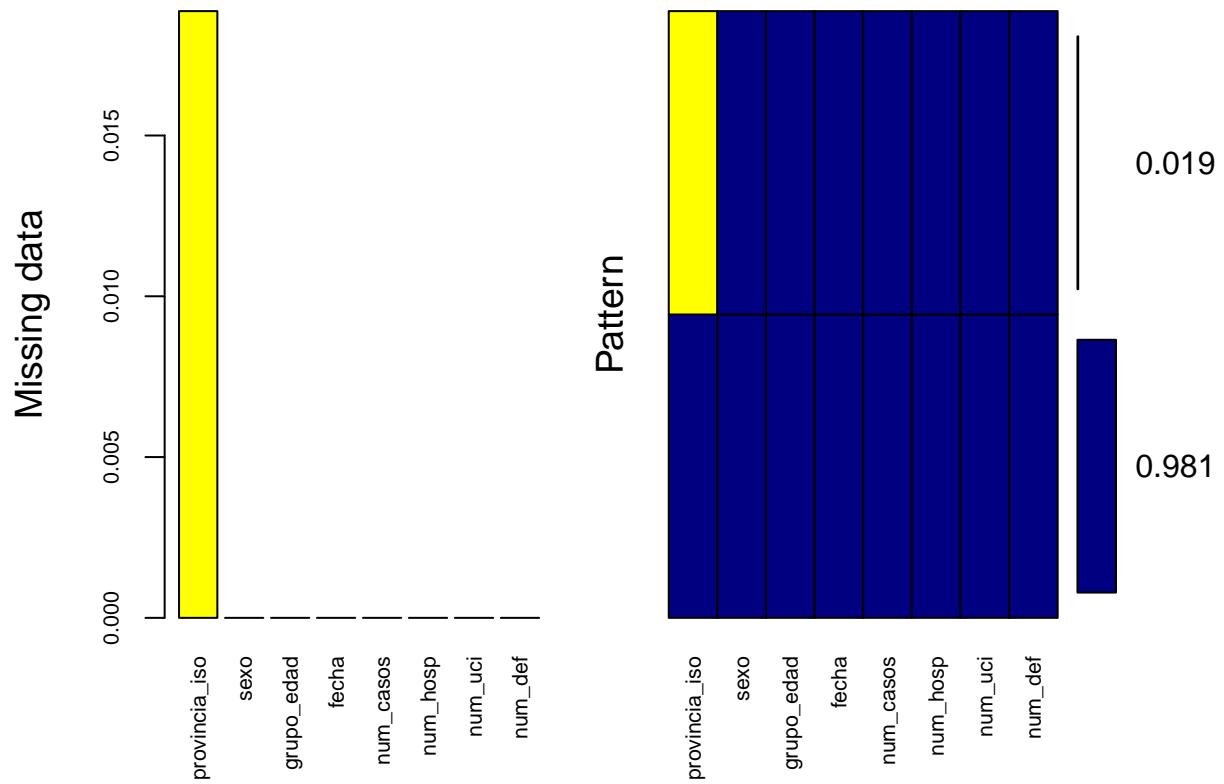
```

We check missing values for CNE_casos. In this case also we omit the NA values.

```

aggr(CNE_casos, col=c('navyblue','yellow'),
      numbers=TRUE, sortVars=TRUE,
      labels=names(CNE_casos), cex.axis=.7,
      gap=3, ylab=c("Missing data","Pattern"))

```



```

##
##  Variables sorted by number of missings:
##          Variable     Count
##  provincia_iso 0.01886792

```

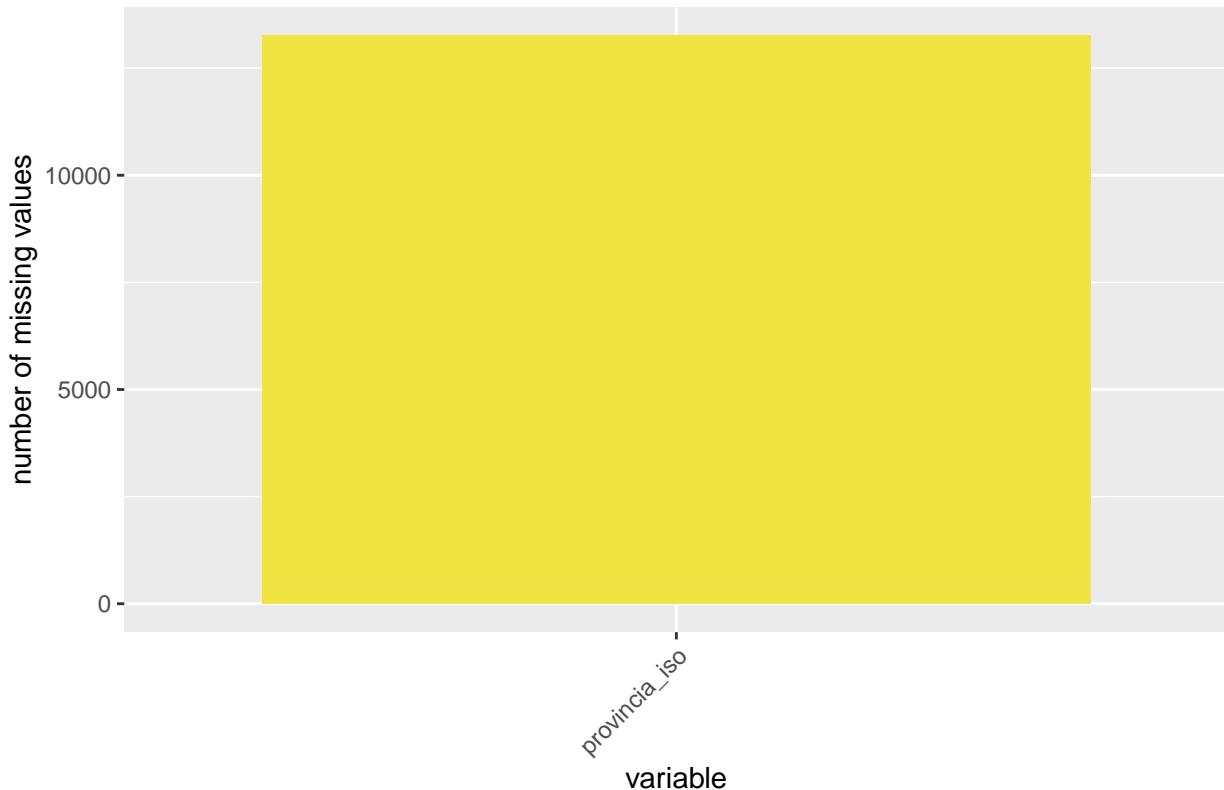
```

##           sexo 0.0000000
##     grupo_edad 0.0000000
##        fecha 0.0000000
##    num_casos 0.0000000
##    num_hosp 0.0000000
##    num_uci 0.0000000
##    num_def 0.0000000

CNE_casos %>%
  gather(key = "key", value = "val") %>%
  mutate(is.missing = is.na(val)) %>%
  group_by(key, is.missing) %>%
  summarise(num.missing = n()) %>%
  filter(is.missing==T) %>%
  select(-is.missing) %>%
  arrange(desc(num.missing)) %>%
  ggplot() +
  geom_bar(aes(x=key, y=num.missing), stat = 'identity', fill="#F0E442") +
  labs(x='variable', y="number of missing values",
       title='Number of missing values') +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

```

Number of missing values



```

#####
CNE_casos <- na.omit(CNE_casos)
#####

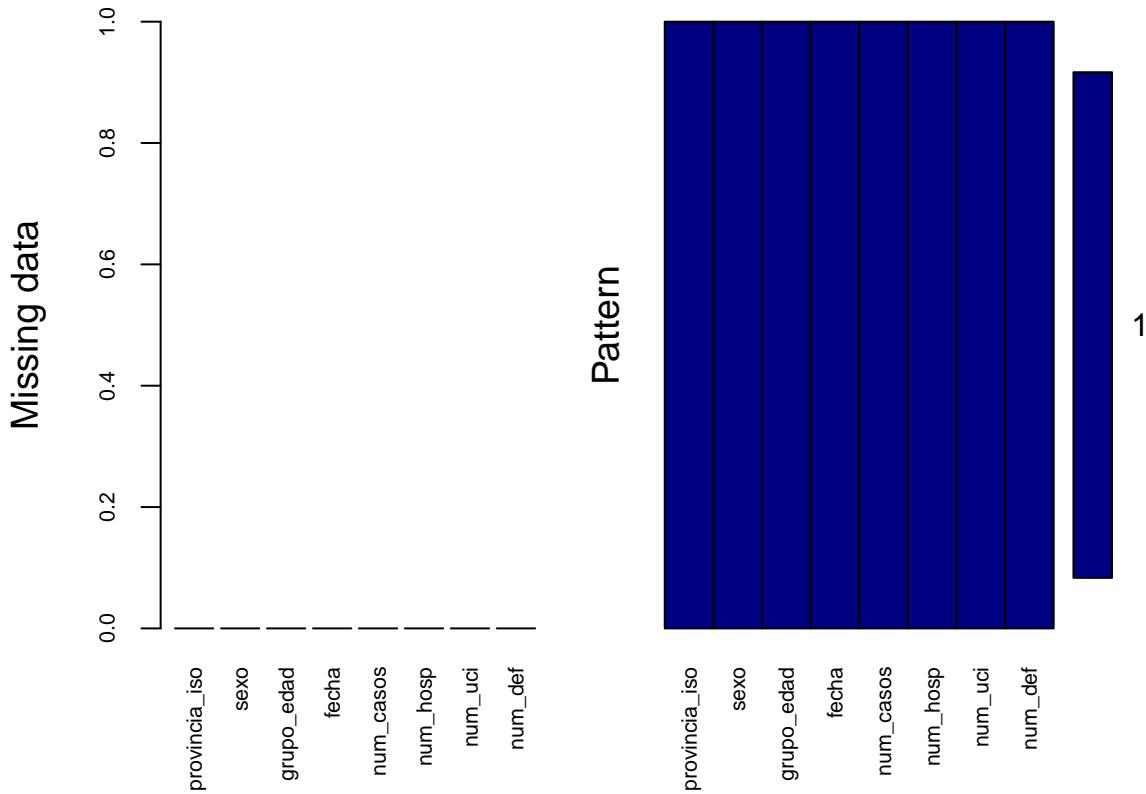
aggr(CNE_casos, col=c('navyblue','yellow'),

```

```

numbers=TRUE, sortVars=TRUE,
labels=names(CNE_casos), cex.axis=.7,
gap=3, ylab=c("Missing data", "Pattern"))

```



```

##
##  Variables sorted by number of missings:
##          Variable Count
##  provincia_iso      0
##          sexo      0
##  grupo_edad      0
##          fecha      0
##  num_casos      0
##  num_hosp      0
##  num_uci      0
##  num_def      0

CNE_casos %>%
  gather(key = "key", value = "val") %>%
  mutate(is.missing = is.na(val)) %>%
  group_by(key, is.missing) %>%
  summarise(num.missing = n()) %>%
  filter(is.missing==T) %>%
  select(-is.missing) %>%
  arrange(desc(num.missing)) %>%
  ggplot() +
  geom_bar(aes(x=key, y=num.missing), stat = 'identity', fill="#F0E442") +

```

```

  labs(x='variable', y="number of missing values",
       title='Number of missing values') +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

```

Number of missing values

number of missing values

variable

2.1.11 CNE data transformation

We are going to **transform / eliminate**:

- A - “Fecha” column is transformed (in both datasets) from “character” to “date.”
- B - “Grupo_edad” and “Sexo” columns are eliminated from dataset “CNE_casos” due to they are not adding value (mobility does not include this variable / dimension).
- C - We change NC iso code to NA (Navarra) in both dataframes.

```

# Transform / eliminate A
CNE_tecnica$fecha <- as.Date(CNE_tecnica$fecha ,format="%Y-%m-%d")
CNE_casos$fecha <- as.Date(CNE_casos$fecha ,format="%Y-%m-%d")

# Transform / eliminate B
CNE_casos<-within(CNE_casos, rm(grupo_edad, sexo))

# Iso code update for Navarra C
CNE_tecnica$provincia_iso[CNE_tecnica$provincia_iso=="NC"] <- "NA"
CNE_casos$provincia_iso[CNE_casos$provincia_iso=="NC"] <- "NA"

# Check table
head(CNE_tecnica,5)

```

```

##    provincia_iso      fecha num_casos num_casos_prueba_pcr
## 1          A 2020-01-01        0            0
## 2         AB 2020-01-01        0            0
## 3         AL 2020-01-01        0            0
## 4         AV 2020-01-01        0            0
## 5          B 2020-01-01        0            0
##    num_casos_prueba_test_ac num_casos_prueba_ag num_casos_prueba_elisa
## 1                      0                      0            0
## 2                      0                      0            0
## 3                      0                      0            0
## 4                      0                      0            0
## 5                      0                      0            0
##    num_casos_prueba_desconocida
## 1                      0
## 2                      0
## 3                      0
## 4                      0
## 5                      0
head(CNE_casos,5)

```

```

##    provincia_iso      fecha num_casos num_hosp num_uci num_def
## 1          A 2020-01-01        0        0        0        0
## 2         AB 2020-01-01        0        0        0        0
## 3         AL 2020-01-01        0        0        0        0
## 4         AV 2020-01-01        0        0        0        0
## 5          B 2020-01-01        0        0        0        0

```

We check both dataframes offers the same total results.

```

# We check both dataframes offers the same total results
CNE_tecnica %>%
  group_by(provincia_iso) %>%
  summarise_at(vars(num_casos), sum)

```

```

## # A tibble: 52 x 2
##   provincia_iso num_casos
##   <chr>           <int>
## 1 A                 143555
## 2 AB                26916
## 3 AL                47032
## 4 AV                11084
## 5 B                 382992
## 6 BA                45886
## 7 BI                80588
## 8 BU                29808
## 9 C                 51272
## 10 CA               70428
## # ... with 42 more rows

```

```

CNE_casos %>%
  group_by(provincia_iso) %>%
  summarise_at(vars(num_casos), sum)

```

```

## # A tibble: 52 x 2
##   provincia_iso num_casos
##   <chr>           <int>
## 1 A                 143555
## 2 AB                26916
## 3 AL                47032
## 4 AV                11084
## 5 B                 382992
## 6 BA                45886
## 7 BI                80588
## 8 BU                29808
## 9 C                 51272
## 10 CA               70428
## # ... with 42 more rows

```

```

## 1 A          143555
## 2 AB         26916
## 3 AL         47032
## 4 AV         11084
## 5 B          382992
## 6 BA         45886
## 7 BI         80588
## 8 BU         29808
## 9 C          51272
## 10 CA        70428
## # ... with 42 more rows

```

2.2 Datasets combinations

We proceed to **combine** the different data sets into one.

2.2.1 CNE_tec_cas

- CNE_casos_g, a grouped dataframe due to the columns eliminated in previous step (grupo_edad, sexo)
- CNE_tec_cas -> CNE_tecnica + CNE_casos_g

Here we merge by columns “provincia_iso,”“fecha.”

```

# CNE_casos_g
CNE_casos_g = CNE_casos %>%
  group_by(provincia_iso, fecha) %>%
  summarise_at(vars(num_casos, num_hosp, num_uci, num_def), sum)
head(CNE_casos_g,5)

## # A tibble: 5 x 6
## # Groups:   provincia_iso [1]
##   provincia_iso fecha      num_casos num_hosp num_uci num_def
##   <chr>       <date>     <int>    <int>   <int>    <int>
## 1 A           2020-01-01     0        1        0        0
## 2 A           2020-01-02     0        0        0        0
## 3 A           2020-01-03     0        0        0        0
## 4 A           2020-01-04     0        0        0        0
## 5 A           2020-01-05     0        1        0        0

# New dataframe CNE_tec_cas
CNE_tec_cas<-merge(CNE_tecnica,
                     CNE_casos_g, by.x=c("provincia_iso","fecha"),
                     by.y=c("provincia_iso","fecha"))

head(CNE_tec_cas,5)

##   provincia_iso     fecha num_casos.x num_casos_prueba_pcr
## 1 A             2020-01-01     0                  0
## 2 A             2020-01-02     0                  0
## 3 A             2020-01-03     0                  0
## 4 A             2020-01-04     0                  0
## 5 A             2020-01-05     0                  0
##   num_casos_prueba_test_ac num_casos_prueba_ag num_casos_prueba_elisa
## 1                      0                      0                      0
## 2                      0                      0                      0
## 3                      0                      0                      0

```

```

## 4          0          0          0
## 5          0          0          0
##   num_casos_prueba_desconocida num_casos.y num_hosp num_uci num_def
## 1          0          0          1          0          0
## 2          0          0          0          0          0
## 3          0          0          0          0          0
## 4          0          0          0          0          0
## 5          0          0          1          0          0
table(CNE_tec_cas$provincia_iso)

##
##   A AB AL AV B BA BI BU C CA CC CE CO CR CS CU GC GI GR GU
## 442 442 442 442 442 442 442 442 442 442 442 442 442 442 442 442 442 442
##   H HU J L LE LO LU M MA ML MU NA O OR P PM PO S SA SE
## 442 442 442 442 442 442 442 442 442 442 442 442 442 442 442 442 442 442
##   SG SO SS T TE TF TO V VA VI Z ZA
## 442 442 442 442 442 442 442 442 442 442 442 442 442 442 442 442 442 442

```

2.2.2 GOG_CNE

- GOG_CNE -> CNE_tec_cas + Google

Here we merge by columns “provincia_iso” / “fecha” and “iso_3166_2_code” / “date.”

```

# New dataframe GOG_CNE
GOG_CNE<-merge(CNE_tec_cas,
                  Google,
                  by.x=c("provincia_iso","fecha"),
                  by.y=c("iso_code","Date"))
head(GOG_CNE,5)

##   provincia_iso      fecha num_casos.x num_casos_prueba_pcr
## 1           A 2020-02-15          1          1
## 2           A 2020-02-16          1          1
## 3           A 2020-02-17          1          1
## 4           A 2020-02-18          1          1
## 5           A 2020-02-19          1          1
##   num_casos_prueba_test_ac num_casos_prueba_ag num_casos_prueba_elisa
## 1              0          0          0
## 2              0          0          0
## 3              0          0          0
## 4              0          0          0
## 5              0          0          0
##   num_casos_prueba_desconocida num_casos.y num_hosp num_uci num_def
## 1          0          0          1          0          0
## 2          0          0          0          0          0
## 3          0          0          1          0          0
## 4          0          0          1          0          0
## 5          0          0          2          1          0
##   sub_region_2 retail_and_recreation_percent_change_from_baseline
## 1 Alicante/Alacant                         3
## 2 Alicante/Alacant                        -2
## 3 Alicante/Alacant                         0
## 4 Alicante/Alacant                        -5
## 5 Alicante/Alacant                         1

```

```

## grocery_and_pharmacy_percent_change_from_baseline
## 1                               -1
## 2                                1
## 3                                2
## 4                               -2
## 5                                1
## parks_percent_change_from_baseline
## 1                                34
## 2                                 8
## 3                                 9
## 4                               -14
## 5                                10
## transit_stations_percent_change_from_baseline
## 1                                 7
## 2                                 5
## 3                                 7
## 4                               -2
## 5                                3
## workplaces_percent_change_from_baseline
## 1                                 0
## 2                               -2
## 3                                 3
## 4                                 2
## 5                                 3
## residential_percent_change_from_baseline
## 1                               -1
## 2                               -1
## 3                                 0
## 4                                 1
## 5                                 0
table(GOG_CNE$provincia_iso)

##
##      A   AB   AL   AV   B   BA   BI   BU   C   CA   CC   CE   CO   CR   CS   CU   GC   GI   GR   GU
## 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385
##      H   HU   J    L   LE   LO   LU   M   MA   ML   MU   NA   O   OR   P   PM   PO   S   SA   SE
## 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385
##      SG  SO   SS   T   TE   TF   TO   V   VA   VI   Z   ZA
## 385 385 385 385 385 385 385 385 385 385 385 385

```

2.2.3 Total GOG_CNE + EM3

- Total -> GOG_CNE + EM3

Here we merge by columns “sub_region_2” / “fecha” and “Zonas.de.movilidad” / “Periodo.” With this dataset we have 17 features for study (plus date, iso code and region name).

```

# New dataframe Total
Total<-merge(GOG_CNE,
              EM3,
              by.x=c("sub_region_2","fecha"),
              by.y=c("Zonas.de.movilidad","Periodo"))

head(Total,5)

```

```

##   sub_region_2      fecha_provincia_iso num_casos.x num_casos_prueba_pcr
## 1 Albacete 2020-03-16          AB        137           132
## 2 Albacete 2020-03-17          AB        128           123
## 3 Albacete 2020-03-18          AB        114           107
## 4 Albacete 2020-03-19          AB        149           133
## 5 Albacete 2020-03-20          AB        131           121
##   num_casos_prueba_test_ac num_casos_prueba_ag num_casos_prueba_elisa
## 1                      5                      0                      0
## 2                      5                      0                      0
## 3                      7                      0                      0
## 4                     16                      0                      0
## 5                     10                      0                      0
##   num_casos_prueba_desconocida num_casos.y num_hosp num_uci num_def
## 1                         0                   65                   43                   3                   7
## 2                         0                   29                   40                   4                   2
## 3                         0                   26                   24                   7                   7
## 4                         0                   22                   40                   5                   7
## 5                         0                   85                   63                   4                   6
##   retail_and_recreation_percent_change_from_baseline
## 1                               -81
## 2                               -84
## 3                               -83
## 4                               -93
## 5                               -87
##   grocery_and_pharmacy_percent_change_from_baseline
## 1                               -32
## 2                               -41
## 3                               -32
## 4                               -92
## 5                               -34
##   parks_percent_change_from_baseline
## 1                               -73
## 2                               -74
## 3                               -70
## 4                               -80
## 5                               -74
##   transit_stations_percent_change_from_baseline
## 1                               -66
## 2                               -72
## 3                               -70
## 4                               -86
## 5                               -76
##   workplaces_percent_change_from_baseline
## 1                               -51
## 2                               -56
## 3                               -58
## 4                               -85
## 5                               -68
##   residential_percent_change_from_baseline Total
## 1                         22 9.900
## 2                         23 9.705
## 3                         23 9.510
## 4                         35 9.130
## 5                         32 8.750

```

```

head(str(Total, vec.len=1))

## 'data.frame': 15080 obs. of 20 variables:
## $ sub_region_2 : chr "Albacete" ...
## $ fecha : Date, format: "2020-03-16" ...
## $ provincia_iso : chr "AB" ...
## $ num_casos.x : int 137 128 ...
## $ num_casos_prueba_pcr : int 132 123 ...
## $ num_casos_prueba_test_ac : int 5 5 ...
## $ num_casos_prueba_ag : int 0 0 ...
## $ num_casos_prueba_elisa : int 0 0 ...
## $ num_casos_prueba_desconocida : int 0 0 ...
## $ num_casos.y : int 65 29 ...
## $ num_hosp : int 43 40 ...
## $ num_uci : int 3 4 ...
## $ num_def : int 7 2 ...
## $ retail_and_recreation_percent_change_from_baseline: num -81 -84 ...
## $ grocery_and_pharmacy_percent_change_from_baseline : num -32 -41 ...
## $ parks_percent_change_from_baseline : num -73 -74 ...
## $ transit_stations_percent_change_from_baseline : num -66 -72 ...
## $ workplaces_percent_change_from_baseline : num -51 -56 ...
## $ residential_percent_change_from_baseline : num 22 23 ...
## $ Total : num 9.9 ...

## NULL

summary(Total)

## sub_region_2      fecha      provincia_iso      num_casos.x
## Length:15080      Min.   :2020-03-16  Length:15080      Min.   : 0
## Class :character  1st Qu.:2020-05-27  Class :character  1st Qu.: 5
## Mode  :character  Median :2020-08-07  Mode  :character  Median : 39
##                   Mean   :2020-08-07          Mean   : 126
##                   3rd Qu.:2020-10-19          3rd Qu.: 120
##                   Max.   :2020-12-30          Max.   :6565
## num_casos_prueba_pcr num_casos_prueba_test_ac num_casos_prueba_ag
## Min.   : 0.0      Min.   : 0.0000      Min.   : 0.00
## 1st Qu.: 5.0      1st Qu.: 0.0000      1st Qu.: 0.00
## Median : 35.0     Median : 0.0000      Median : 0.00
## Mean   : 110.2    Mean   : 0.2832      Mean   : 15.19
## 3rd Qu.: 105.0    3rd Qu.: 0.0000      3rd Qu.: 4.00
## Max.   :6546.0    Max.   :32.0000      Max.   :1465.00
## num_casos_prueba_elisa num_casos_prueba_desconocida num_casos.y
## Min.   : 0.0000      Min.   : 0.0000      Min.   : 0
## 1st Qu.: 0.0000      1st Qu.: 0.0000      1st Qu.: 6
## Median : 0.0000      Median : 0.0000      Median : 37
## Mean   : 0.1989      Mean   : 0.1317      Mean   : 127
## 3rd Qu.: 0.0000      3rd Qu.: 0.0000      3rd Qu.: 117
## Max.   :71.0000      Max.   :65.0000      Max.   :7724
## num_hosp      num_uci      num_def
## Min.   : 0.00      Min.   : 0.000      Min.   : 0.000
## 1st Qu.: 1.00      1st Qu.: 0.000      1st Qu.: 0.000
## Median : 4.00      Median : 0.000      Median : 1.000
## Mean   : 14.86      Mean   : 1.281      Mean   : 3.437
## 3rd Qu.: 12.00      3rd Qu.: 1.000      3rd Qu.: 3.000

```

```

##  Max.    :1930.00   Max.    :135.000   Max.    :334.000
## retail_and_recreation_percent_change_from_baseline
## Min.    :-97.00
## 1st Qu.:-57.00
## Median :-30.00
## Mean   :-37.29
## 3rd Qu.:-17.00
## Max.    : 71.00
## grocery_and_pharmacy_percent_change_from_baseline
## Min.    :-96.00
## 1st Qu.:-24.00
## Median : -6.00
## Mean   :-11.75
## 3rd Qu.:  4.00
## Max.    :194.00
## parks_percent_change_from_baseline
## Min.    :-94.000
## 1st Qu.:-30.000
## Median : -2.000
## Mean   :  5.809
## 3rd Qu.: 30.000
## Max.    :543.000
## transit_stations_percent_change_from_baseline
## Min.    :-100.00
## 1st Qu.:-53.00
## Median : -31.00
## Mean   : -35.19
## 3rd Qu.:-17.00
## Max.    : 74.00
## workplaces_percent_change_from_baseline
## Min.    :-92.00
## 1st Qu.:-43.00
## Median :-26.00
## Mean   :-29.08
## 3rd Qu.:-13.00
## Max.    : 55.00
## residential_percent_change_from_baseline      Total
## Min.    :-10.00                         Min.    : 1.95
## 1st Qu.:  4.00                         1st Qu.:11.36
## Median :  7.00                         Median :14.39
## Mean   : 10.14                        Mean   :14.20
## 3rd Qu.: 14.00                        3rd Qu.:17.11
## Max.    : 48.00                         Max.    :29.00

Total$num_casos.x <- as.numeric(Total$num_casos.x)
Total$num_casos_prueba_pcr <- as.numeric(Total$num_casos_prueba_pcr)
Total$num_casos_prueba_test_ac <- as.numeric(Total$num_casos_prueba_test_ac)
Total$num_casos_prueba_ag <- as.numeric(Total$num_casos_prueba_ag)
Total$num_casos_prueba_elisa <- as.numeric(Total$num_casos_prueba_elisa)
Total$num_casos_prueba_desconocida <- as.numeric(Total$num_casos_prueba_desconocida)
Total$num_casos.y <- as.numeric(Total$num_casos.y)
Total$num_hosp <- as.numeric(Total$num_hosp)
Total$num_uci <- as.numeric(Total$num_uci)
Total$num_def <- as.numeric(Total$num_def)

```

```
table(Total$sub_region_2)
```

```
##          Albacete      Alicante/Alacant      Almería
##          290                  290                  290
##      Araba/Álava      Asturias                  Ávila
##          290                  290                  290
##      Badajoz      Balears, Illes      Barcelona
##          290                  290                  290
##      Bizkaia      Burgos                  Cáceres
##          290                  290                  290
##      Cádiz      Cantabria      Castellón/Castelló
##          290                  290                  290
##      Ceuta      Ciudad Real      Córdoba
##          290                  290                  290
##      Coruña, A      Cuenca      Gipuzkoa
##          290                  290                  290
##      Girona      Granada      Guadalajara
##          290                  290                  290
##      Huelva      Huesca                  Jaén
##          290                  290                  290
##      León      Lleida                  Lugo
##          290                  290                  290
##      Madrid      Málaga      Melilla
##          290                  290                  290
##      Murcia      Navarra      Ourense
##          290                  290                  290
##      Palencia      Palmas, Las      Pontevedra
##          290                  290                  290
##      Rioja, La      Salamanca      Santa Cruz de Tenerife
##          290                  290                  290
##      Segovia      Sevilla                  Soria
##          290                  290                  290
##      Tarragona      Teruel                  Toledo
##          290                  290                  290
##      Valencia/València      Valladolid      Zamora
##          290                  290                  290
##      Zaragoza
##          290
```

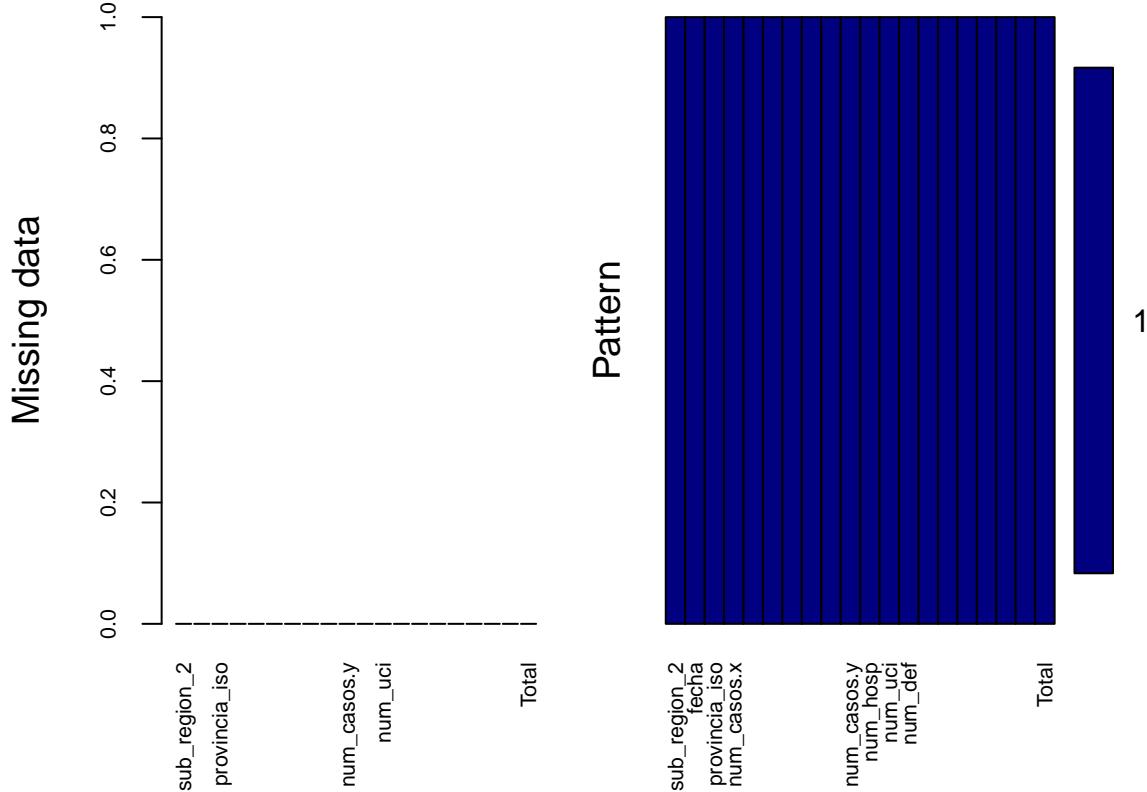
```
table(Total$provincia_iso)
```

```
##          A   AB   AL   AV     B   BA   BI   BU     C   CA   CC   CE   CO   CR   CS   CU   GC   GI   GR   GU
##  290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290
##      H   HU   J    L   LE   LO   LU    M   MA   ML   MU   NA    O   OR    P   PM   PO    S   SA   SE
##  290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290
##      SG   SO   SS    T   TE   TF   TO    V   VA   VI    Z   ZA
##  290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290
```

We check the missing values. We should have zero missing values

```
aggr(Total, col=c('navyblue','yellow'),
      numbers=TRUE, sortVars=TRUE,
      labels=names(Total), cex.axis=.7,
```

```
gap=3, ylab=c("Missing data","Pattern"))
```



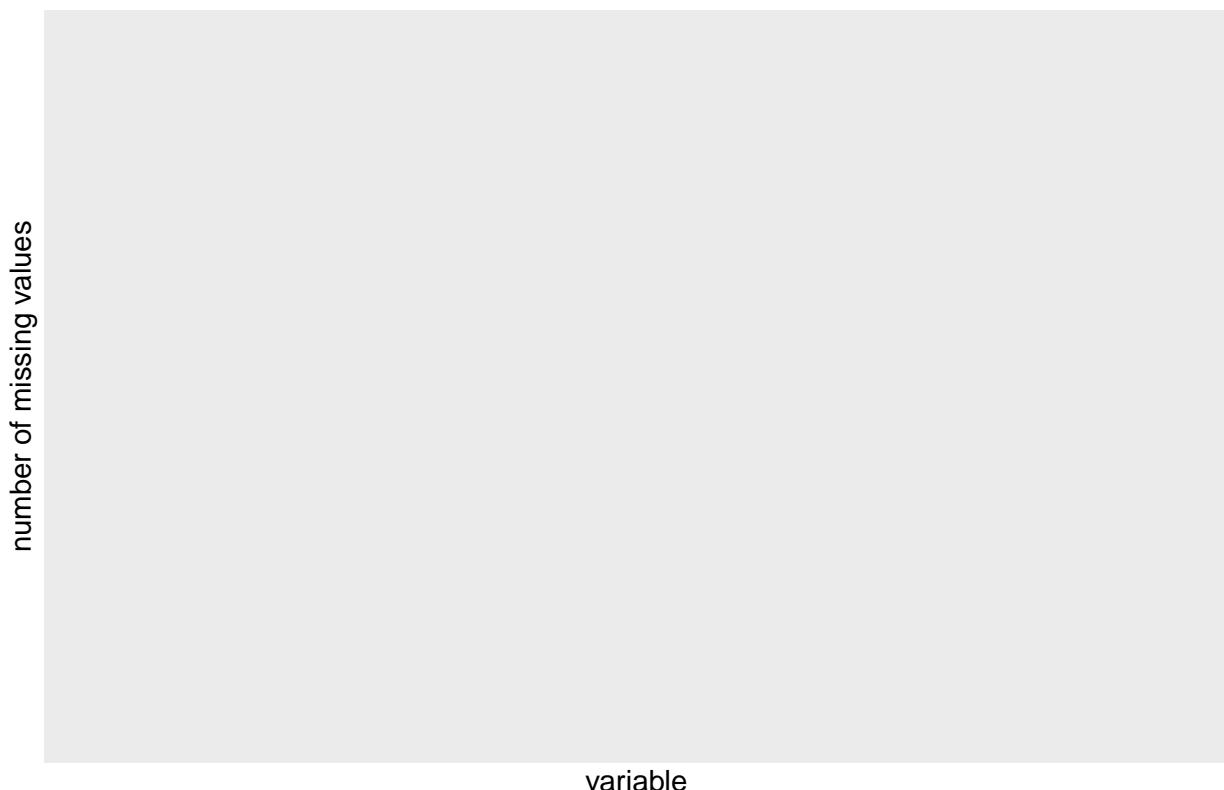
```
##  
##  Variables sorted by number of missings:  
##  
##          Variable Count  
##  sub_region_2          0  
##          fecha          0  
##  provincia_iso         0  
##          num_casos.x         0  
##          num_casos_prueba_pcr 0  
##          num_casos_prueba_test_ac 0  
##          num_casos_prueba_ag 0  
##          num_casos_prueba_elisa 0  
##          num_casos_prueba_desconocida 0  
##          num_casos.y          0  
##          num_hosp          0  
##          num_uci           0  
##          num_def           0  
##  retail_and_recreation_percent_change_from_baseline 0  
##  grocery_and_pharmacy_percent_change_from_baseline 0  
##          parks_percent_change_from_baseline 0  
##          transit_stations_percent_change_from_baseline 0  
##          workplaces_percent_change_from_baseline 0  
##          residential_percent_change_from_baseline 0  
##          Total            0
```

```

Total %>%
  gather(key = "key", value = "val") %>%
  mutate(is.missing = is.na(val)) %>%
  group_by(key, is.missing) %>%
  summarise(num.missing = n()) %>%
  filter(is.missing==T) %>%
  select(-is.missing) %>%
  arrange(desc(num.missing)) %>%
  ggplot() +
  geom_bar(aes(x=key, y=num.missing), stat = 'identity', fill="#F0E442") +
  labs(x='variable', y="number of missing values",
       title='Number of missing values') +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

```

Number of missing values



```

# Review results
# Discrepancies due to different time-frames when merge CNE dataframes (see previous checks)
Total %>%
  group_by(provincia_iso) %>%
  summarise_at(vars(num_casos.x,num_casos.y), sum)

## # A tibble: 52 x 3
##   provincia_iso num_casos.x num_casos.y
##   <chr>           <dbl>        <dbl>
## 1 A                 56493        55068
## 2 AB                16459        16626
## 3 AL                21488        21372

```

```

## 4 AV           6525      6681
## 5 B           257034     261208
## 6 BA          23165      22612
## 7 BI          56867      57728
## 8 BU          22742      22978
## 9 C           25641      25604
## 10 CA         31593      31225
## # ... with 42 more rows
# CSV file generation
head(Total,5)

##   sub_region_2      fecha provincia_iso num_casos.x num_casos_prueba_pcr
## 1    Albacete 2020-03-16          AB        137            132
## 2    Albacete 2020-03-17          AB        128            123
## 3    Albacete 2020-03-18          AB        114            107
## 4    Albacete 2020-03-19          AB        149            133
## 5    Albacete 2020-03-20          AB        131            121
##   num_casos_prueba_test_ac num_casos_prueba_ag num_casos_prueba_elisa
## 1                      5                  0                  0
## 2                      5                  0                  0
## 3                      7                  0                  0
## 4                     16                  0                  0
## 5                     10                  0                  0
##   num_casos_prueba_desconocida num_casos.y num_hosp num_uci num_def
## 1                         0        65       43       3       7
## 2                         0        29       40       4       2
## 3                         0        26       24       7       7
## 4                         0        22       40       5       7
## 5                         0        85       63       4       6
##   retail_and_recreation_percent_change_from_baseline
## 1                                         -81
## 2                                         -84
## 3                                         -83
## 4                                         -93
## 5                                         -87
##   grocery_and_pharmacy_percent_change_from_baseline
## 1                                         -32
## 2                                         -41
## 3                                         -32
## 4                                         -92
## 5                                         -34
##   parks_percent_change_from_baseline
## 1                                         -73
## 2                                         -74
## 3                                         -70
## 4                                         -80
## 5                                         -74
##   transit_stations_percent_change_from_baseline
## 1                                         -66
## 2                                         -72
## 3                                         -70
## 4                                         -86
## 5                                         -76
##   workplaces_percent_change_from_baseline

```

```

## 1 -51
## 2 -56
## 3 -58
## 4 -85
## 5 -68
##   residential_percent_change_from_baseline Total
## 1          22 9.900
## 2          23 9.705
## 3          23 9.510
## 4          35 9.130
## 5          32 8.750

head(str(Total, vec.len=1))

## 'data.frame': 15080 obs. of 20 variables:
## $ sub_region_2 : chr "Albacete" ...
## $ fecha        : Date, format: "2020-03-16" ...
## $ provincia_iso: chr "AB" ...
## $ num_casos.x  : num 137 128 ...
## $ num_casos_pcr: num 132 123 ...
## $ num_casos_prueba_test_ac: num 5 5 ...
## $ num_casos_prueba_ag: num 0 0 ...
## $ num_casos_prueba_elisa: num 0 0 ...
## $ num_casos_prueba_desconocida: num 0 0 ...
## $ num_casos.y  : num 65 29 ...
## $ num_hosp     : num 43 40 ...
## $ num_uci      : num 3 4 ...
## $ num_def      : num 7 2 ...
## $ retail_and_recreation_percent_change_from_baseline: num -81 -84 ...
## $ grocery_and_pharmacy_percent_change_from_baseline : num -32 -41 ...
## $ parks_percent_change_from_baseline                 : num -73 -74 ...
## $ transit_stations_percent_change_from_baseline     : num -66 -72 ...
## $ workplaces_percent_change_from_baseline           : num -51 -56 ...
## $ residential_percent_change_from_baseline          : num 22 23 ...
## $ Total       : num 9.9 ...

## NULL

summary(Total)

## sub_region_2      fecha      provincia_iso      num_casos.x
## Length:15080      Min.   :2020-03-16  Length:15080      Min.   : 0
## Class :character  1st Qu.:2020-05-27  Class :character  1st Qu.: 5
## Mode  :character  Median :2020-08-07  Mode  :character  Median : 39
##                  Mean   :2020-08-07                Mean   : 126
##                  3rd Qu.:2020-10-19                3rd Qu.: 120
##                  Max.   :2020-12-30                Max.   :6565
## num_casos_pcr num_casos_prueba_test_ac num_casos_prueba_ag
## Min.   : 0.0      Min.   : 0.0000      Min.   : 0.00
## 1st Qu.: 5.0      1st Qu.: 0.0000      1st Qu.: 0.00
## Median : 35.0     Median : 0.0000      Median : 0.00
## Mean   : 110.2    Mean   : 0.2832      Mean   : 15.19
## 3rd Qu.: 105.0    3rd Qu.: 0.0000      3rd Qu.: 4.00
## Max.   :6546.0     Max.   :32.0000      Max.   :1465.00
## num_casos_prueba_elisa num_casos_prueba_desconocida num_casos.y
## Min.   : 0.0000      Min.   : 0.0000      Min.   : 0

```

```

## 1st Qu.: 0.0000      1st Qu.: 0.0000      1st Qu.:   6
## Median : 0.0000      Median : 0.0000      Median : 37
## Mean   : 0.1989      Mean   : 0.1317      Mean   : 127
## 3rd Qu.: 0.0000      3rd Qu.: 0.0000      3rd Qu.: 117
## Max.   :71.0000      Max.   :65.0000      Max.   :7724
##      num_hosp          num_uci          num_def
## Min.   : 0.00  Min.   : 0.000  Min.   : 0.000
## 1st Qu.: 1.00  1st Qu.: 0.000  1st Qu.: 0.000
## Median : 4.00  Median : 0.000  Median : 1.000
## Mean   : 14.86  Mean   : 1.281  Mean   : 3.437
## 3rd Qu.: 12.00  3rd Qu.: 1.000  3rd Qu.: 3.000
## Max.   :1930.00  Max.   :135.000 Max.   :334.000
## retail_and_recreation_percent_change_from_baseline
## Min.   :-97.00
## 1st Qu.:-57.00
## Median :-30.00
## Mean   :-37.29
## 3rd Qu.:-17.00
## Max.   : 71.00
## grocery_and_pharmacy_percent_change_from_baseline
## Min.   :-96.00
## 1st Qu.:-24.00
## Median : -6.00
## Mean   : -11.75
## 3rd Qu.:  4.00
## Max.   :194.00
## parks_percent_change_from_baseline
## Min.   :-94.000
## 1st Qu.:-30.000
## Median : -2.000
## Mean   :  5.809
## 3rd Qu.: 30.000
## Max.   :543.000
## transit_stations_percent_change_from_baseline
## Min.   :-100.00
## 1st Qu.:-53.00
## Median : -31.00
## Mean   : -35.19
## 3rd Qu.: -17.00
## Max.   : 74.00
## workplaces_percent_change_from_baseline
## Min.   :-92.00
## 1st Qu.:-43.00
## Median : -26.00
## Mean   : -29.08
## 3rd Qu.:-13.00
## Max.   : 55.00
## residential_percent_change_from_baseline      Total
## Min.   :-10.00                      Min.   : 1.95
## 1st Qu.:  4.00                      1st Qu.:11.36
## Median :  7.00                      Median :14.39
## Mean   : 10.14                      Mean   :14.20
## 3rd Qu.: 14.00                      3rd Qu.:17.11
## Max.   : 48.00                      Max.   :29.00

```

```

table(Total$provincia_iso)

##
##   A   AB   AL   AV   B   BA   BI   BU   C   CA   CC   CE   CO   CR   CS   CU   GC   GI   GR   GU
## 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290
##   H   HU   J   L   LE   LO   LU   M   MA   ML   MU   NA   O   OR   P   PM   PO   S   SA   SE
## 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290
##   SG   SO   SS   T   TE   TF   TO   V   VA   VI   Z   ZA
## 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290

write.csv2(Total, "D:\\UOC Master Data Science\\_ M2.882 - TFM - Área 5\\UOC - Guia - PECS\\Pec3\\Total.csv",
           row.names = FALSE)

```

2.3 Visual analysis

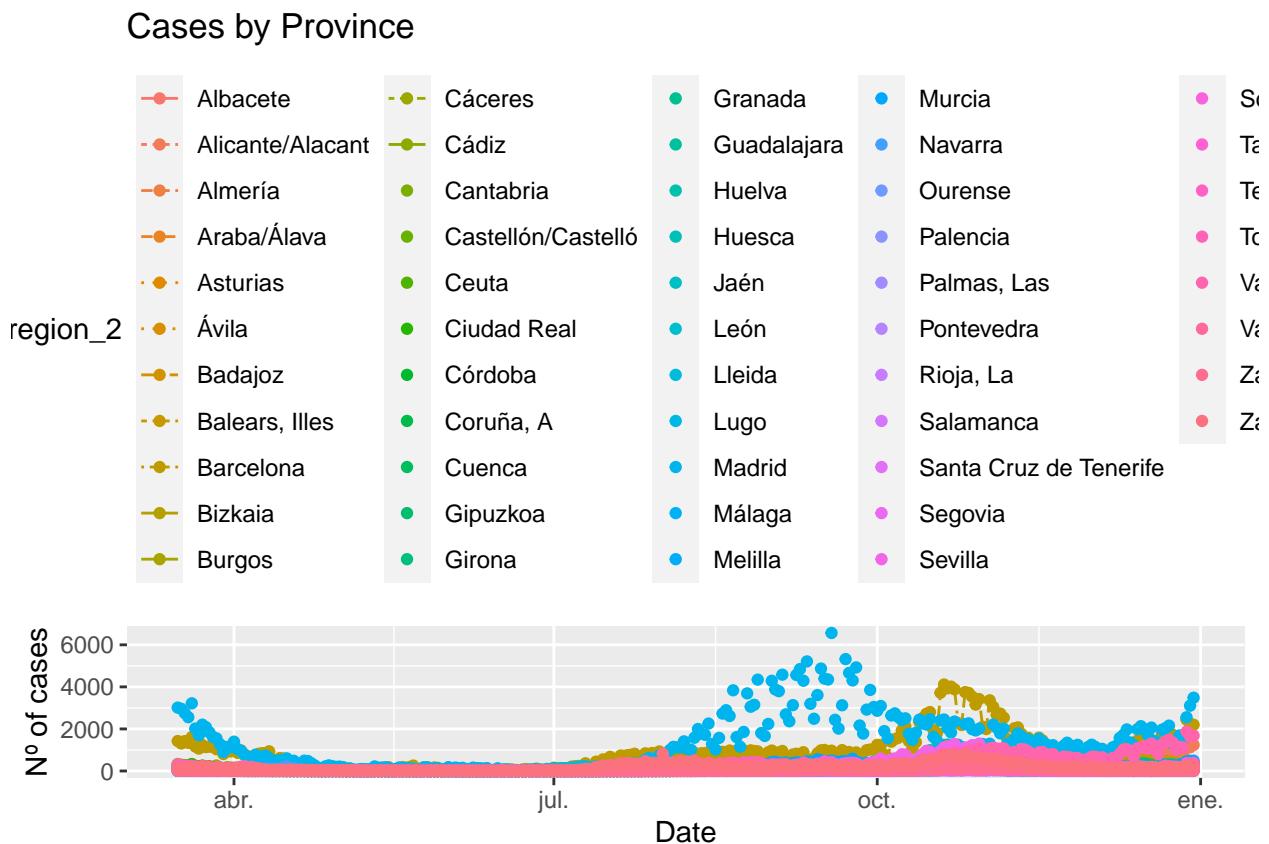
2.3.1 Dataframe plots (Málaga, Sevilla and Cádiz)

We have generated some plots from the `dataframe` object generated.

```

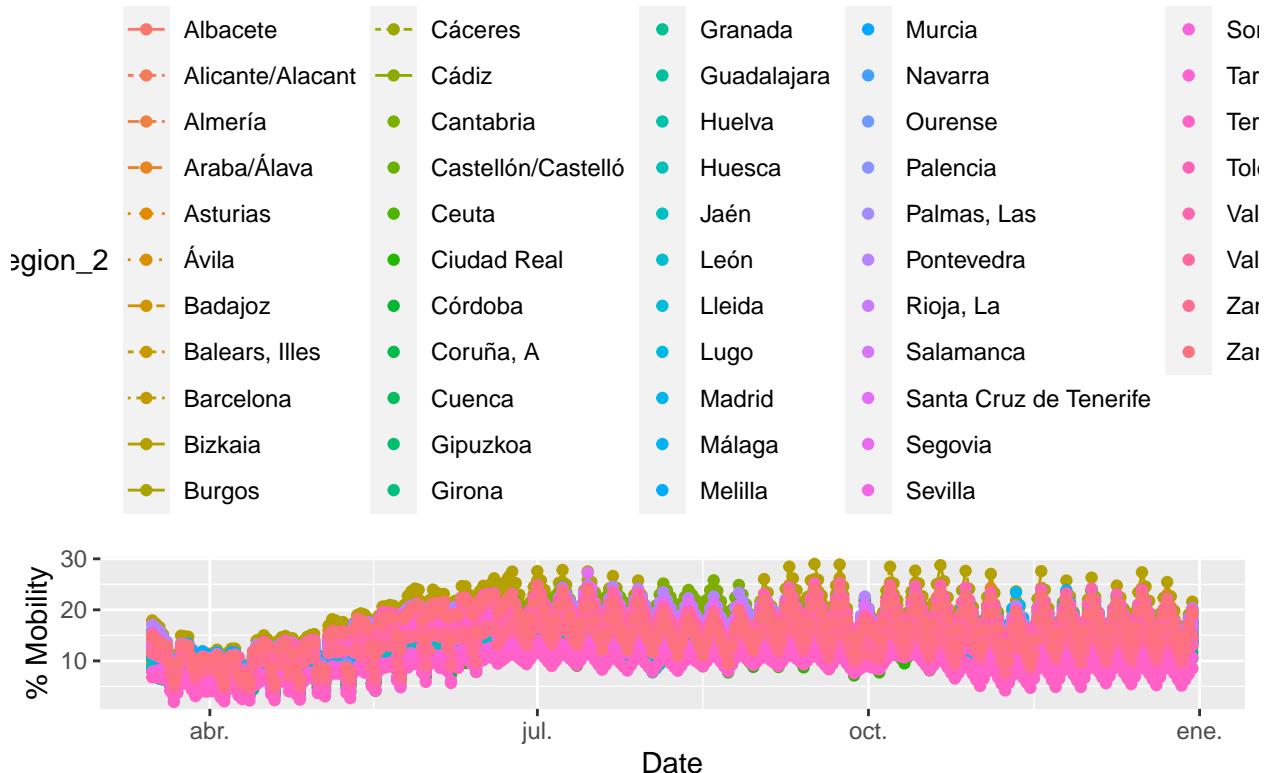
# Line plots
# All num_casos.x
ggplot(Total, aes(x=fecha, y=num_casos.x, group=sub_region_2)) +
  geom_line(aes(linetype=sub_region_2, color=sub_region_2))+
  geom_point(aes(color=sub_region_2))+ 
  theme(legend.position="top") + 
  labs(title="Cases by Province",
       x ="Date", y = "Nº of cases")

```



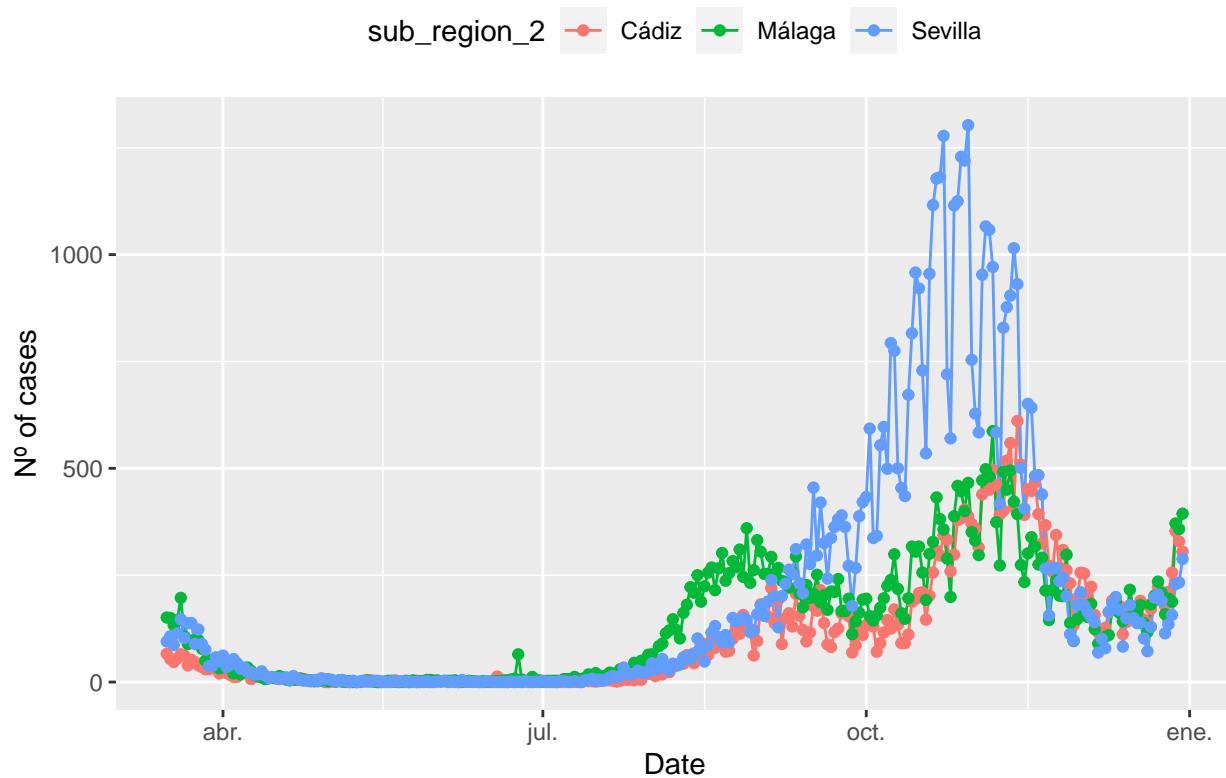
```
# All Total (mobility)
ggplot(Total, aes(x=fecha, y=Total, group=sub_region_2)) +
  geom_line(aes(linetype=sub_region_2, color=sub_region_2))+ 
  geom_point(aes(color=sub_region_2))+ 
  theme(legend.position="top") + 
  labs(title="Mobility Change by Province",
       x ="Date", y = "% Mobility")
```

Mobility Change by Province



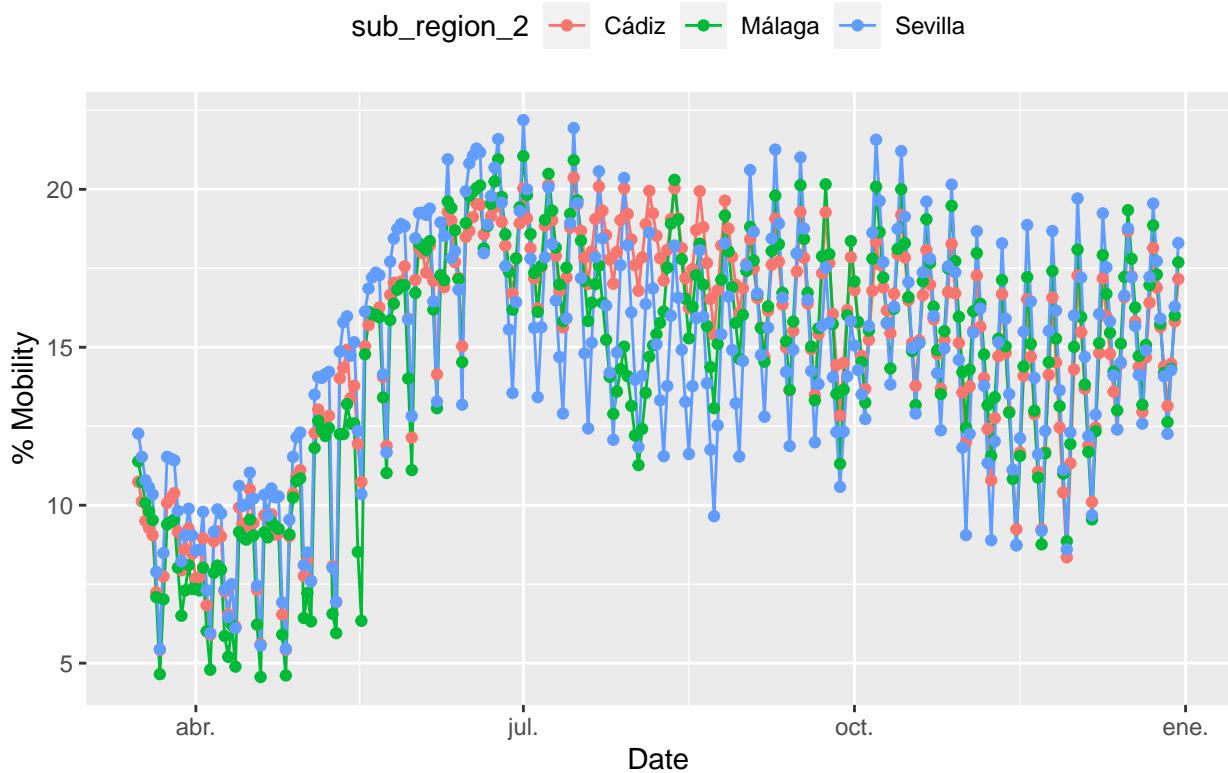
```
# Mal, Sev and Cad - num_casos.x
Total %%%
filter(sub_region_2 == "Málaga" | sub_region_2 == "Cádiz" |
       sub_region_2 == "Sevilla") %>%
ggplot(aes(x=fecha, y=num_casos.x))+
  geom_line(aes(color=sub_region_2))+ 
  geom_point(aes(color=sub_region_2))+ 
  theme(legend.position="top") + 
  labs(title="Cases reported by Province (Málaga, Sevilla and Cádiz)",
       x ="Date", y = "Nº of cases")
```

Cases reported by Province (Málaga, Sevilla and Cádiz)



```
# Mal, Sev and Cad - Total (mobility)
Total %>%
  filter(sub_region_2 == "Málaga" | sub_region_2 == "Cádiz" |
         sub_region_2 == "Sevilla") %>%
  ggplot(aes(x=fecha, y=Total)) +
  geom_line(aes(color=sub_region_2)) +
  geom_point(aes(color=sub_region_2)) +
  theme(legend.position="top") +
  labs(title="Mobility Change by Province (Málaga, Sevilla and Cádiz)",
       x ="Date", y = "% Mobility")
```

Mobility Change by Province (Málaga, Sevilla and Cádiz)



2.3.2 Time-series plots (Barcelona, Madrid, Málaga, Sevilla and Cádiz)

We have generated some plots from the `time-series` object generated. We have used `tsibble()`.

```
# Convert dataframe to ts object
#library(fpp3)
Total_ts <- Total[-3] %>%
  mutate(Dia_c = as_date(fecha)) %>%
  select(-fecha) %>%
  as_tsibble(key = c(sub_region_2),
             index = Dia_c)

# Filter for Bar, Mad, Mal, Cor and, Cad
Total_ts %>% filter(sub_region_2 == "Barcelona" | sub_region_2 == "Madrid" |
                      sub_region_2 == "Málaga" | sub_region_2 == "Sevilla" |
                      sub_region_2 == "Cádiz") -> Total_ts_b

#####
Total_ts

## # A tsibble: 15,080 x 19 [1D]
## # Key:     sub_region_2 [52]
##   sub_region_2 num_casos.x num_casos_prueb~ num_casos_prueb~ num_casos_prueb~
##   <chr>          <dbl>            <dbl>            <dbl>            <dbl>
## 1 Albacete      137             132              5              0
## 2 Albacete      128             123              5              0
## 3 Albacete      114             107              7              0
```

```

## 4 Albacete      149      133      16      0
## 5 Albacete      131      121      10      0
## 6 Albacete      129      120       9      0
## 7 Albacete      125      112      13      0
## 8 Albacete      112      103       9      0
## 9 Albacete      107      91       16      0
## 10 Albacete     78       64       14      0
## # ... with 15,070 more rows, and 14 more variables:
## #   num_casos_prueba_elisa <dbl>, num_casos_prueba_desconocida <dbl>,
## #   num_casos.y <dbl>, num_hosp <dbl>, num_uci <dbl>, num_def <dbl>,
## #   retail_and_recreation_percent_change_from_baseline <dbl>,
## #   grocery_and_pharmacy_percent_change_from_baseline <dbl>,
## #   parks_percent_change_from_baseline <dbl>,
## #   transit_stations_percent_change_from_baseline <dbl>,
## #   workplaces_percent_change_from_baseline <dbl>,
## #   residential_percent_change_from_baseline <dbl>, Total <dbl>, Dia_c <date>
Total_ts_b

## # A tsibble: 1,450 x 19 [1D]
## # Key:     sub_region_2 [5]
##     sub_region_2 num_casos.x num_casos_prueb~ num_casos_prueb~ num_casos_prueb~
##     <chr>          <dbl>        <dbl>        <dbl>        <dbl>
## 1 Barcelona      1424        1351         0         0
## 2 Barcelona      1309        1273         0         0
## 3 Barcelona      1420        1379         0         0
## 4 Barcelona      1338        1289         0         0
## 5 Barcelona      1614        1557         0         0
## 6 Barcelona      1164        1136         0         0
## 7 Barcelona      1065        1032         0         0
## 8 Barcelona      1234        1187         0         0
## 9 Barcelona      1137        1095         0         0
## 10 Barcelona     1159        1131         0         0
## # ... with 1,440 more rows, and 14 more variables:
## #   num_casos_prueba_elisa <dbl>, num_casos_prueba_desconocida <dbl>,
## #   num_casos.y <dbl>, num_hosp <dbl>, num_uci <dbl>, num_def <dbl>,
## #   retail_and_recreation_percent_change_from_baseline <dbl>,
## #   grocery_and_pharmacy_percent_change_from_baseline <dbl>,
## #   parks_percent_change_from_baseline <dbl>,
## #   transit_stations_percent_change_from_baseline <dbl>,
## #   workplaces_percent_change_from_baseline <dbl>,
## #   residential_percent_change_from_baseline <dbl>, Total <dbl>, Dia_c <date>
Total_ts_b %>% distinct(sub_region_2)

## # A tibble: 5 x 1
##   sub_region_2
##   <chr>
## 1 Barcelona
## 2 Cádiz
## 3 Madrid
## 4 Málaga
## 5 Sevilla
#####

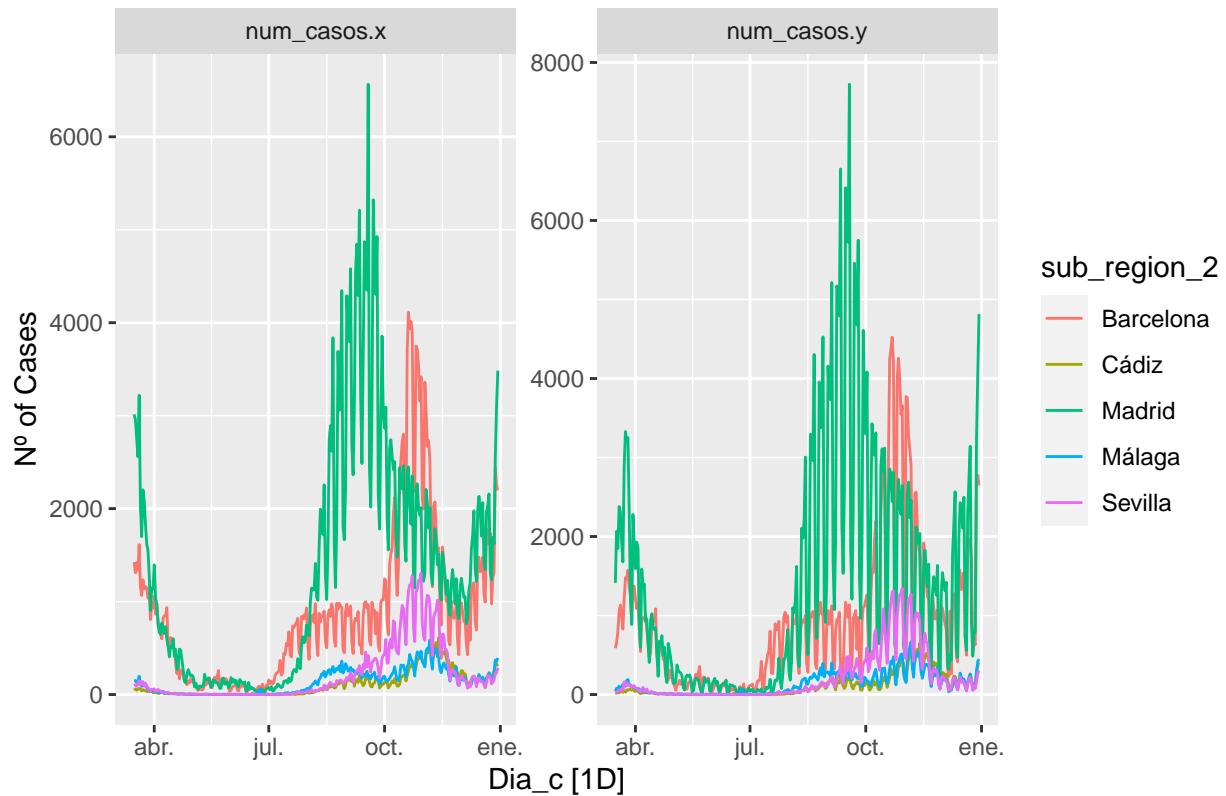
```

```

# Plots
# A num_casos.x,num_casos.y
autoplot(Total_ts_b, vars(num_casos.x,num_casos.y)) +
  labs(y = "Nº of Cases",
       title = "Reported Cases (CNE num_casos.x vs num_casos.y)")

```

Reported Cases (CNE num_casos.x vs num_casos.y)

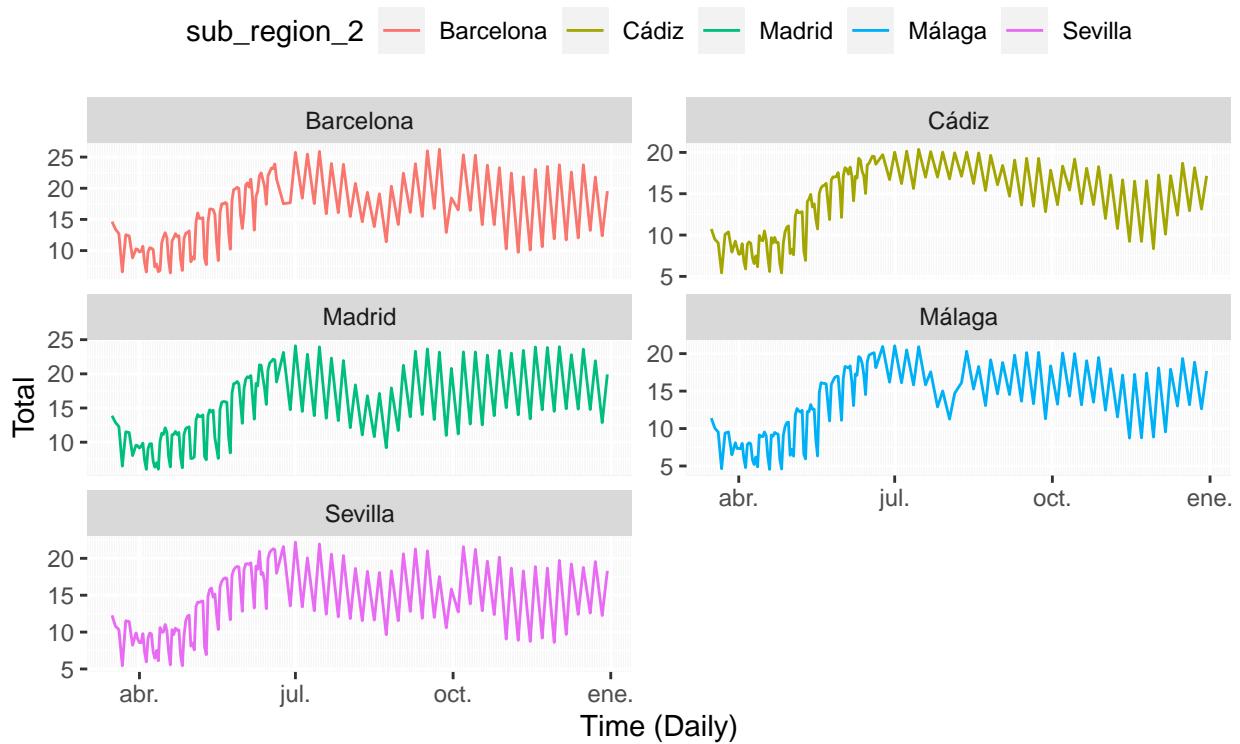


```

# B Total (mobility)
autoplot(Total_ts_b, Total) +
  facet_wrap(~sub_region_2, scales = "free_y", ncol=2) +
  theme(legend.position = "top") +
  scale_x_date(date_minor_breaks = "1 day", name = "Time (Daily)") +
  ggtitle(label = "EM3 - Mobility Change by Province (Barcelona, Madrid, Málaga,
  Córdoba and Cádiz)")

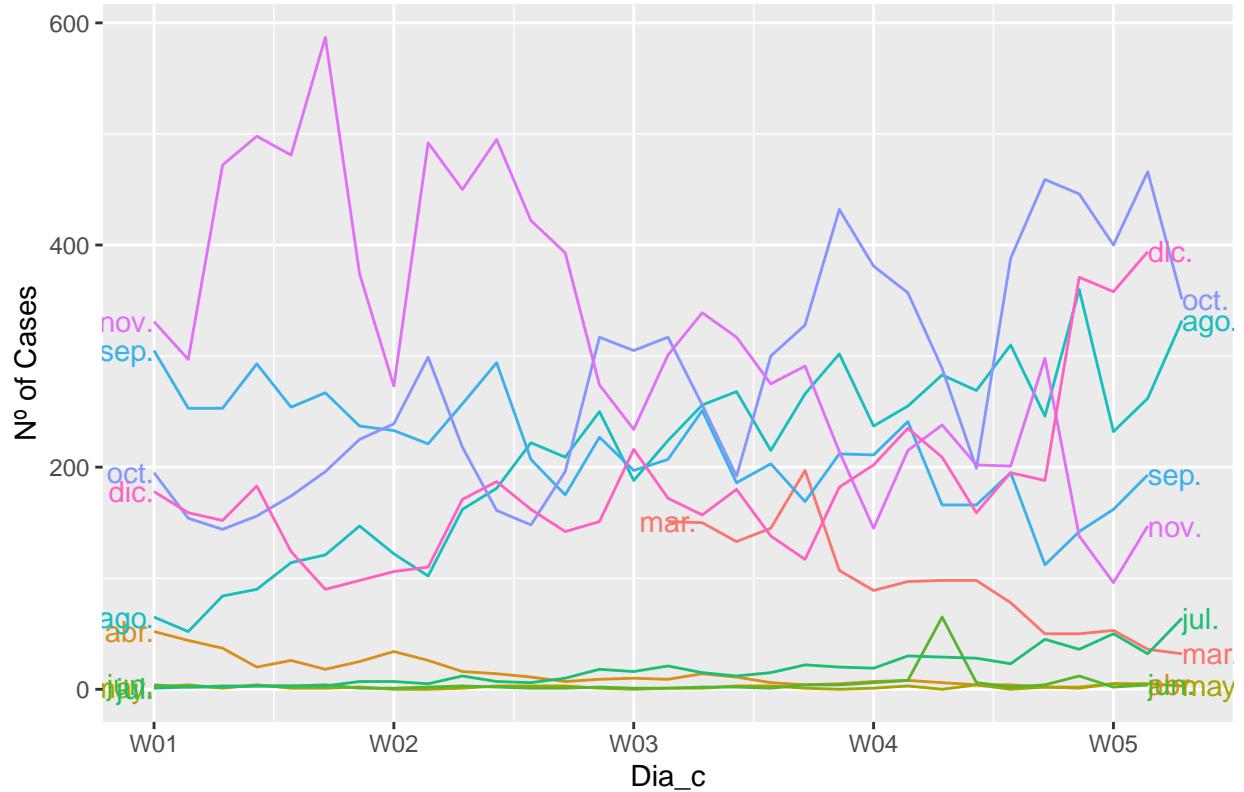
```

EM3 – Mobility Change by Province (Barcelona, Madrid, Málaga, Córdoba and Cádiz)



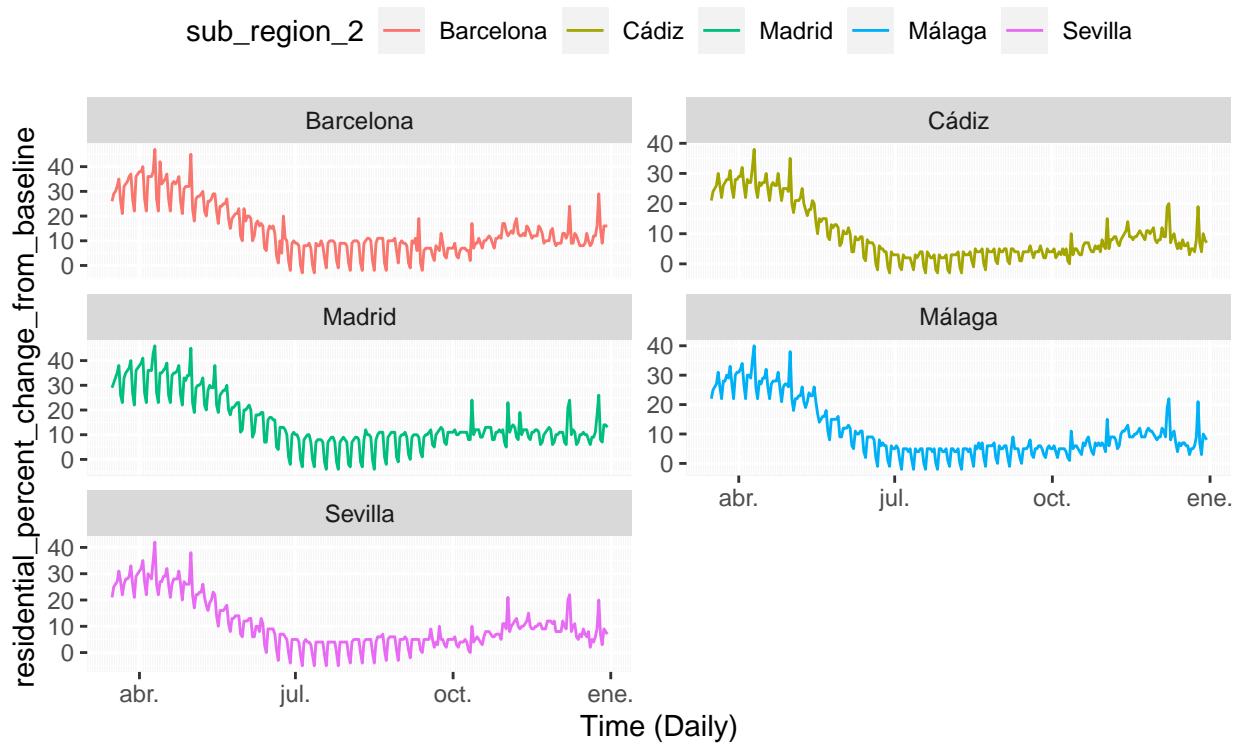
```
# C sub_region_2 == "Barcelona" by month
Total_ts %>% filter(sub_region_2 == "Málaga") %>%
  gg_season(num_casos.x, period = "month", labels = "both") +
  theme(legend.position = "top") +
  labs(y="Nº of Cases", title="Barcelona - Infections by Month")
```

Barcelona – Infections by Month



```
# D Google (% residential change)
autoplot(Total_ts_b, residential_percent_change_from_baseline) +
  facet_wrap(~sub_region_2, scales = "free_y", ncol=2) +
  theme(legend.position = "top") +
  scale_x_date(date_minor_breaks = "1 day", name = "Time (Daily)") +
  ggtitle(label = "Google % residential change (Barcelona, Madrid, Málaga,
  Córdoba and Cádiz)")
```

Google % residential change (Barcelona, Madrid, Málaga, Córdoba and Cádiz)

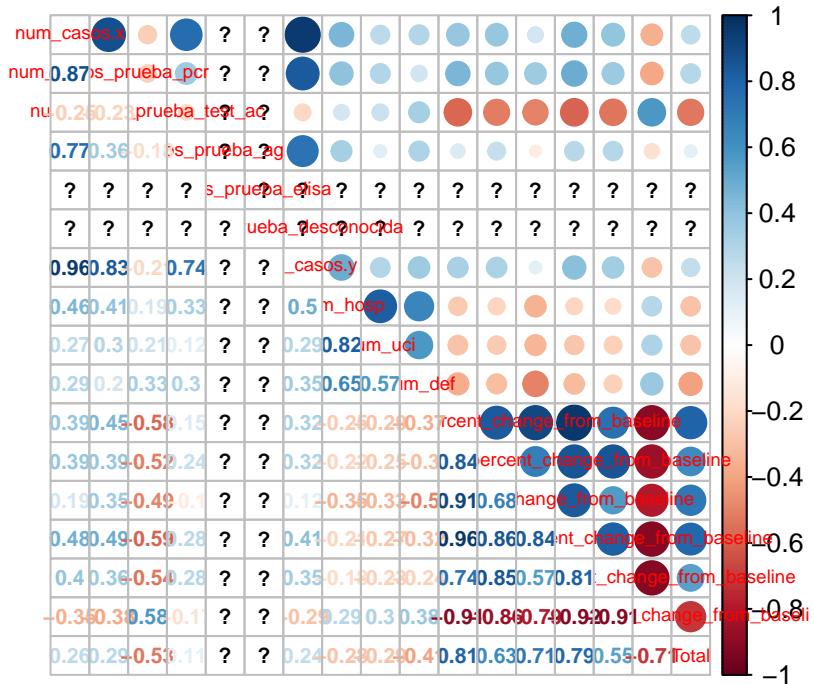


2.3.3 Correlation plots (from dataframe)

```
# Filter to "sub_region_2" == "Barcelona" or "Málaga"
# Character / date columns are eliminated
library(corrplot)

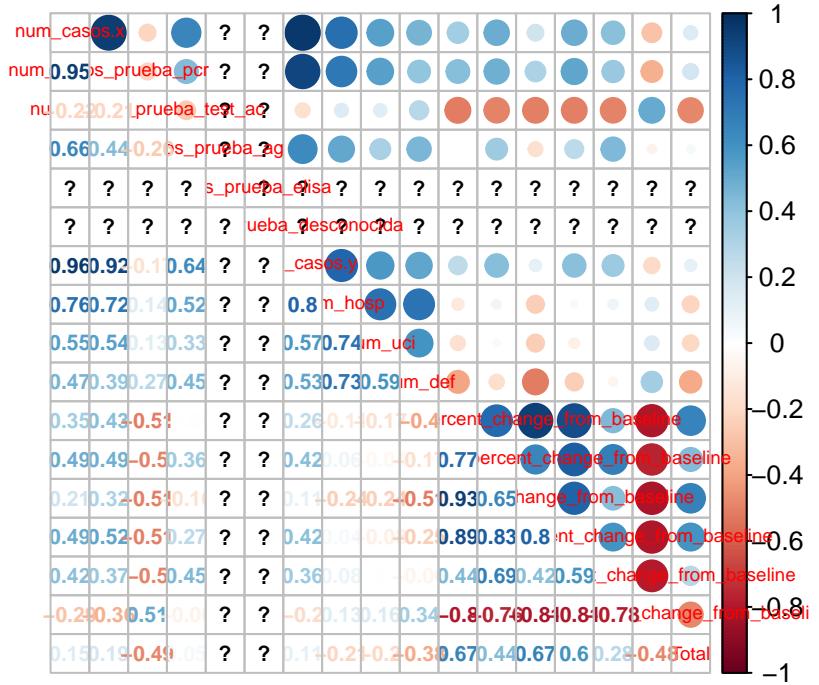
#### Málaga
# pearson
Total.res<-Total %>%
  filter(sub_region_2 == "Málaga")
Total.res<-cor(Total.res[,c(-1,-2,-3)],method="pearson")
corrplot.mixed(Total.res,upper="circle",number.cex=.65,tl.cex=.6, title="Málaga -
  pearson ")
```

pearson



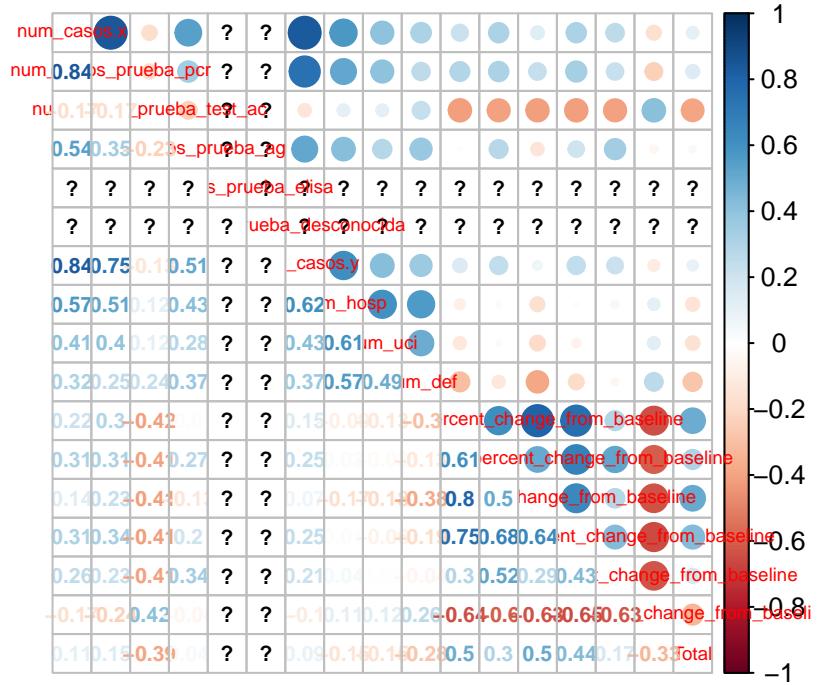
```
# spearman
Total.res<-Total %>%
  filter(sub_region_2 == "Málaga")
Total.res<-cor(Total.res[,c(-1,-2,-3)],method="spearman")
corrplot.mixed(Total.res,upper="circle",number.cex=.65,tl.cex=.6, title="Málaga -
spearman ")
```

spearman



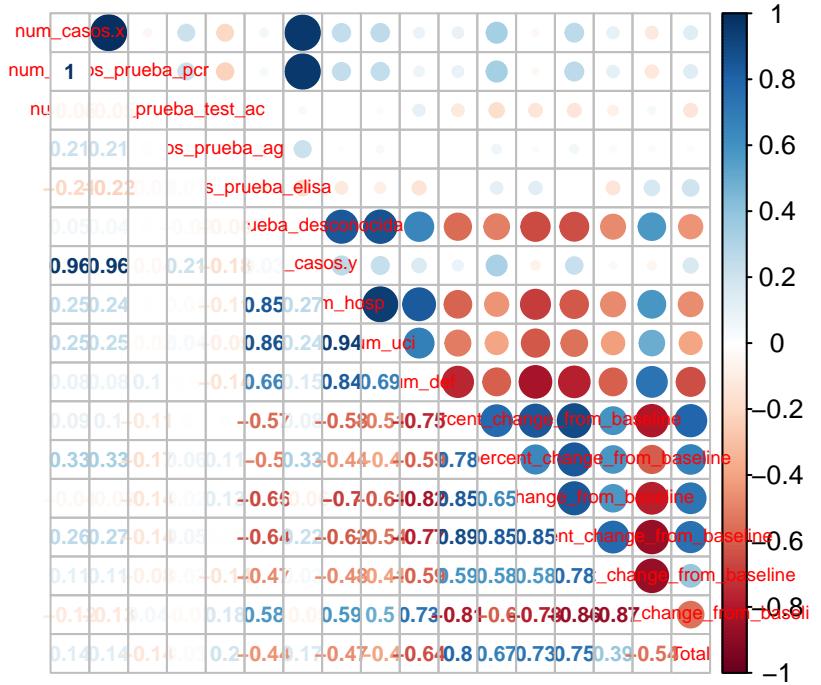
```
# kendall
Total.res<-Total %>%
  filter(sub_region_2 == "Málaga")
Total.res<-cor(Total.res[,c(-1,-2,-3)],method="kendall")
corrplot.mixed(Total.res,upper="circle",number.cex=.65,tl.cex=.6, title="Málaga -
  kendall ")
```

kendall



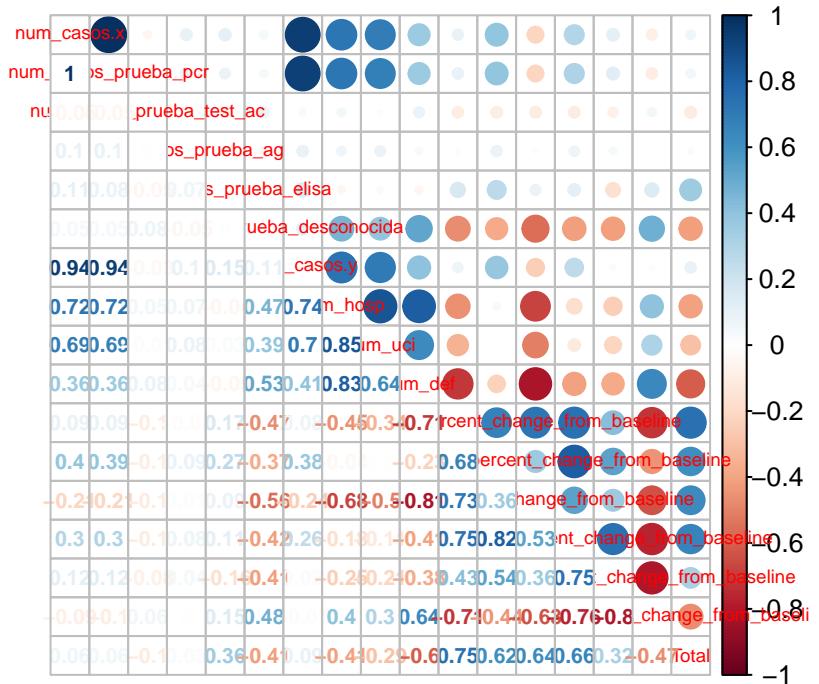
```
#### Barcelona
# pearson
Total.res<-Total %>%
  filter(sub_region_2 == "Barcelona")
Total.res<-cor(Total.res[,c(-1,-2,-3)],method="pearson")
corrplot.mixed(Total.res,upper="circle",number.cex=.65,tl.cex=.6, title="Barcelona -
  pearson ")
```

pearson



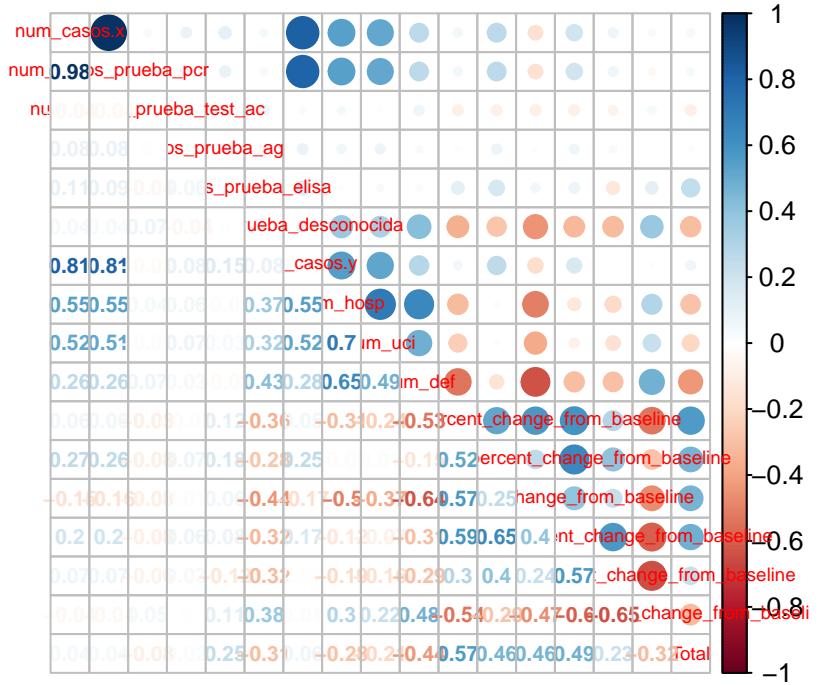
```
# spearman
Total.res<-Total %>%
  filter(sub_region_2 == "Barcelona")
Total.res<-cor(Total.res[,c(-1,-2,-3)],method="spearman")
corrplot.mixed(Total.res,upper="circle",number.cex=.65,tl.cex=.6, title="Barcelona - spearman")
```

spearman



```
# kendall
Total.res<-Total %>%
  filter(sub_region_2 == "Barcelona")
Total.res<-cor(Total.res[,c(-1,-2,-3)],method="kendall")
corrplot.mixed(Total.res,upper="circle",number.cex=.65,tl.cex=.6, title="Barcelona - kendall ")
```

kendall



2.3.4 PCA (Barcelona)

```
pca <- prcomp(Total.res, scale = T)
summary(pca)

## Importance of components:
##                PC1       PC2       PC3       PC4       PC5       PC6       PC7
## Standard deviation   3.0590  1.8914  1.12142  1.02409  0.83504  0.57825  0.46266
## Proportion of Variance 0.5504  0.2104  0.07398  0.06169  0.04102  0.01967  0.01259
## Cumulative Proportion 0.5504  0.7609  0.83485  0.89655  0.93756  0.95723  0.96982
##                  PC8       PC9       PC10      PC11      PC12      PC13      PC14
## Standard deviation   0.41006  0.33620  0.28445  0.21986  0.17734  0.16911  0.15372
## Proportion of Variance 0.00989  0.00665  0.00476  0.00284  0.00185  0.00168  0.00139
## Cumulative Proportion 0.97971  0.98636  0.99112  0.99397  0.99582  0.99750  0.99889
##                  PC15      PC16      PC17
## Standard deviation   0.1369  0.01274  1.826e-17
## Proportion of Variance 0.0011  0.00001  0.000e+00
## Cumulative Proportion 1.0000  1.00000  1.000e+00

pca$rotation

##                                     PC1          PC2
## num_casos.x      -0.1354826901 -0.469400471
## num_casos_pcr    -0.1356873829 -0.469237915
## num_casos_prueba_ac -0.0782958432  0.237502756
## num_casos_prueba_ag  0.0005323686  0.008790618
```

```

## num_casos_prueba_elisa          0.0495912937  0.004385553
## num_casos_prueba_desconocida   -0.2907430443  0.113686282
## num_casos.y                     -0.1546876637 -0.450080493
## num_hosp                        -0.2956764064 -0.199498016
## num_uci                          -0.2788863801 -0.230508957
## num_def                         -0.3153112111 -0.030904089
## retail_and_recreation_percent_change_from_baseline 0.3077977291 -0.116093786
## grocery_and_pharmacy_percent_change_from_baseline   0.2553316609 -0.261732823
## parks_percent_change_from_baseline      0.3128404870  0.012776948
## transit_stations_percent_change_from_baseline 0.2855716515 -0.218931952
## workplaces_percent_change_from_baseline   0.2622358364 -0.170838668
## residential_percent_change_from_baseline    -0.2895648649  0.166332144
## Total                            0.2993175108 -0.081611685
##                                         PC3          PC4
## num_casos.x                      -0.002875288 -0.028445983
## num_casos_prueba_pcr            -0.012292151 -0.026795983
## num_casos_prueba_test_ac        -0.392712472 -0.268237834
## num_casos_prueba_ag             -0.034460138  0.950253303
## num_casos_prueba_elisa          0.811470962 -0.060011278
## num_casos_prueba_desconocida   0.019674417 -0.078105637
## num_casos.y                     0.041668217 -0.036946432
## num_hosp                        -0.051419434 -0.010413059
## num_uci                          -0.018126448  0.006323566
## num_def                         -0.031133681  0.009747098
## retail_and_recreation_percent_change_from_baseline 0.033200491 -0.075398568
## grocery_and_pharmacy_percent_change_from_baseline   0.027494995 -0.030995413
## parks_percent_change_from_baseline      0.029299658 -0.011018427
## transit_stations_percent_change_from_baseline  -0.099387689 -0.018021579
## workplaces_percent_change_from_baseline   -0.303553292  0.028981871
## residential_percent_change_from_baseline    0.211975965  0.003369575
## Total                            0.173800838 -0.065471155
##                                         PC5          PC6
## num_casos.x                      -0.15653475  0.16549669
## num_casos_prueba_pcr            -0.15256567  0.17028866
## num_casos_prueba_test_ac        -0.81534671 -0.10365823
## num_casos_prueba_ag             -0.25969423 -0.01360833
## num_casos_prueba_elisa          -0.34805371 -0.37185910
## num_casos_prueba_desconocida   0.20580692 -0.11429059
## num_casos.y                     -0.15827091  0.15297764
## num_hosp                        0.03855580 -0.11688342
## num_uci                          0.06172358 -0.06173530
## num_def                         0.08791553 -0.26888200
## retail_and_recreation_percent_change_from_baseline -0.02731712  0.18700422
## grocery_and_pharmacy_percent_change_from_baseline   0.01445198 -0.34534708
## parks_percent_change_from_baseline      -0.01705528  0.39173313
## transit_stations_percent_change_from_baseline  0.03520761 -0.26003250
## workplaces_percent_change_from_baseline   0.10943614 -0.48832683
## residential_percent_change_from_baseline    0.03104587  0.13040337
## Total                            -0.05468582  0.19337305
##                                         PC7          PC8
## num_casos.x                      0.11538168  0.07515692
## num_casos_prueba_pcr            0.11048811  0.07713501
## num_casos_prueba_test_ac        -0.08510797 -0.12117026
## num_casos_prueba_ag             -0.03926649 -0.14832473

```

## num_casos_prueba_elisa	0.24859208	0.01374512
## num_casos_prueba_desconocida	0.19845835	-0.75630827
## num_casos.y	0.04117476	-0.02860659
## num_hosp	-0.11367668	-0.07611704
## num_uci	-0.02723992	-0.20583241
## num_def	-0.20890547	0.15436921
## retail_and_recreation_percent_change_from_baseline	-0.10130686	-0.31700087
## grocery_and_pharmacy_percent_change_from_baseline	-0.63305922	-0.04093499
## parks_percent_change_from_baseline	0.15076652	0.04293438
## transit_stations_percent_change_from_baseline	-0.06852096	-0.19335324
## workplaces_percent_change_from_baseline	0.37588850	0.19906444
## residential_percent_change_from_baseline	-0.39494991	0.23079899
## Total	-0.26415699	-0.26912855
##		PC9 PC10
## num_casos.x	-0.204798939	-0.059620970
## num_casos_prueba_pcr	-0.206074966	-0.057649643
## num_casos_prueba_test_ac	0.031408747	0.010328003
## num_casos_prueba_ag	-0.027397924	-0.004813823
## num_casos_prueba_elisa	0.085159377	-0.047189805
## num_casos_prueba_desconocida	-0.349092565	-0.125363275
## num_casos.y	-0.130984317	0.031574368
## num_hosp	0.283065654	0.067094187
## num_uci	0.717067401	0.169083317
## num_def	-0.009119052	0.013475201
## retail_and_recreation_percent_change_from_baseline	0.099531350	-0.317145228
## grocery_and_pharmacy_percent_change_from_baseline	-0.110976209	-0.296856620
## parks_percent_change_from_baseline	0.283874181	-0.240873233
## transit_stations_percent_change_from_baseline	0.094114251	-0.074265913
## workplaces_percent_change_from_baseline	-0.125028631	0.240735868
## residential_percent_change_from_baseline	-0.178700468	0.013386663
## Total	-0.118124812	0.793981499
##		PC11 PC12
## num_casos.x	-0.049617465	-0.0988357734
## num_casos_prueba_pcr	-0.044726847	-0.1058640000
## num_casos_prueba_test_ac	-0.042562355	-0.0389190655
## num_casos_prueba_ag	0.003758095	-0.0008550931
## num_casos_prueba_elisa	0.023430627	-0.0013867741
## num_casos_prueba_desconocida	-0.154872738	-0.1615965585
## num_casos.y	0.033429358	0.1359821524
## num_hosp	0.302911164	0.0671255962
## num_uci	-0.400459562	0.0477114671
## num_def	0.553043315	-0.1950704924
## retail_and_recreation_percent_change_from_baseline	0.286708424	0.6275693247
## grocery_and_pharmacy_percent_change_from_baseline	-0.348092535	-0.0594956464
## parks_percent_change_from_baseline	-0.082227368	-0.4958947989
## transit_stations_percent_change_from_baseline	0.264287632	-0.4433548371
## workplaces_percent_change_from_baseline	-0.216784094	0.2008296543
## residential_percent_change_from_baseline	-0.278596900	0.0313367951
## Total	0.074087201	-0.0830357636
##		PC13 PC14
## num_casos.x	0.3503737964	-0.03447444
## num_casos_prueba_pcr	0.3691954013	-0.03383423
## num_casos_prueba_test_ac	-0.0023721487	-0.01431289
## num_casos_prueba_ag	0.0009495141	-0.01895981

```

## num_casos_prueba_elisa          0.0189278939 -0.03470038
## num_casos_prueba_desconocida   -0.0341231415 -0.11380732
## num_casos.y                     -0.8087585769  0.15136854
## num_hosp                        -0.0093409355 -0.32920173
## num_uci                          0.0905105752  0.12476243
## num_def                         -0.0179131471 -0.20425404
## retail_and_recreation_percent_change_from_baseline 0.1591640063 0.05248551
## grocery_and_pharmacy_percent_change_from_baseline   -0.0689717137 -0.30356466
## parks_percent_change_from_baseline      -0.1967768342 -0.33947866
## transit_stations_percent_change_from_baseline 0.0373896719 0.66046288
## workplaces_percent_change_from_baseline   -0.0493969622 -0.09490414
## residential_percent_change_from_baseline    0.0240600499 0.34670086
## Total                            0.0590314137 -0.13712447
##                                         PC15          PC16
## num_casos.x                      0.021886709 -0.710827710
## num_casos_prueba_pcr            0.020989875 0.696614910
## num_casos_prueba_test_ac        0.002128322 -0.009045443
## num_casos_prueba_ag             -0.001031653 -0.006169593
## num_casos_prueba_elisa          -0.019875868 0.004055133
## num_casos_prueba_desconocida   -0.000227830 -0.014131490
## num_casos.y                     0.065935929 0.012850260
## num_hosp                        -0.738022542 -0.005532761
## num_uci                          0.291185664 -0.005062923
## num_def                          0.517898894 -0.032682612
## retail_and_recreation_percent_change_from_baseline 0.103782290 -0.031882076
## grocery_and_pharmacy_percent_change_from_baseline   0.033876519 0.014999252
## parks_percent_change_from_baseline      -0.032945065 -0.039812604
## transit_stations_percent_change_from_baseline  -0.169052448 -0.008281975
## workplaces_percent_change_from_baseline   -0.059237981 -0.040905347
## residential_percent_change_from_baseline    -0.223342981 -0.056238440
## Total                            0.040900018 -0.009746369
##                                         PC17
## num_casos.x                      -0.04694046
## num_casos_prueba_pcr            0.08912178
## num_casos_prueba_test_ac        0.09544863
## num_casos_prueba_ag             0.05828430
## num_casos_prueba_elisa          0.04498527
## num_casos_prueba_desconocida   0.16508413
## num_casos.y                     -0.01242631
## num_hosp                        0.04584112
## num_uci                          0.06322805
## num_def                          0.30912666
## retail_and_recreation_percent_change_from_baseline 0.33484091
## grocery_and_pharmacy_percent_change_from_baseline   -0.13871900
## parks_percent_change_from_baseline      0.41803506
## transit_stations_percent_change_from_baseline 0.04834211
## workplaces_percent_change_from_baseline   0.44739326
## residential_percent_change_from_baseline    0.57600490
## Total                            0.07657033

if(!require(FactoMineR)){
  install.packages('FactoMineR', repos='http://cran.us.r-project.org')
  library(FactoMineR)}
if(!require(factoextra)) {

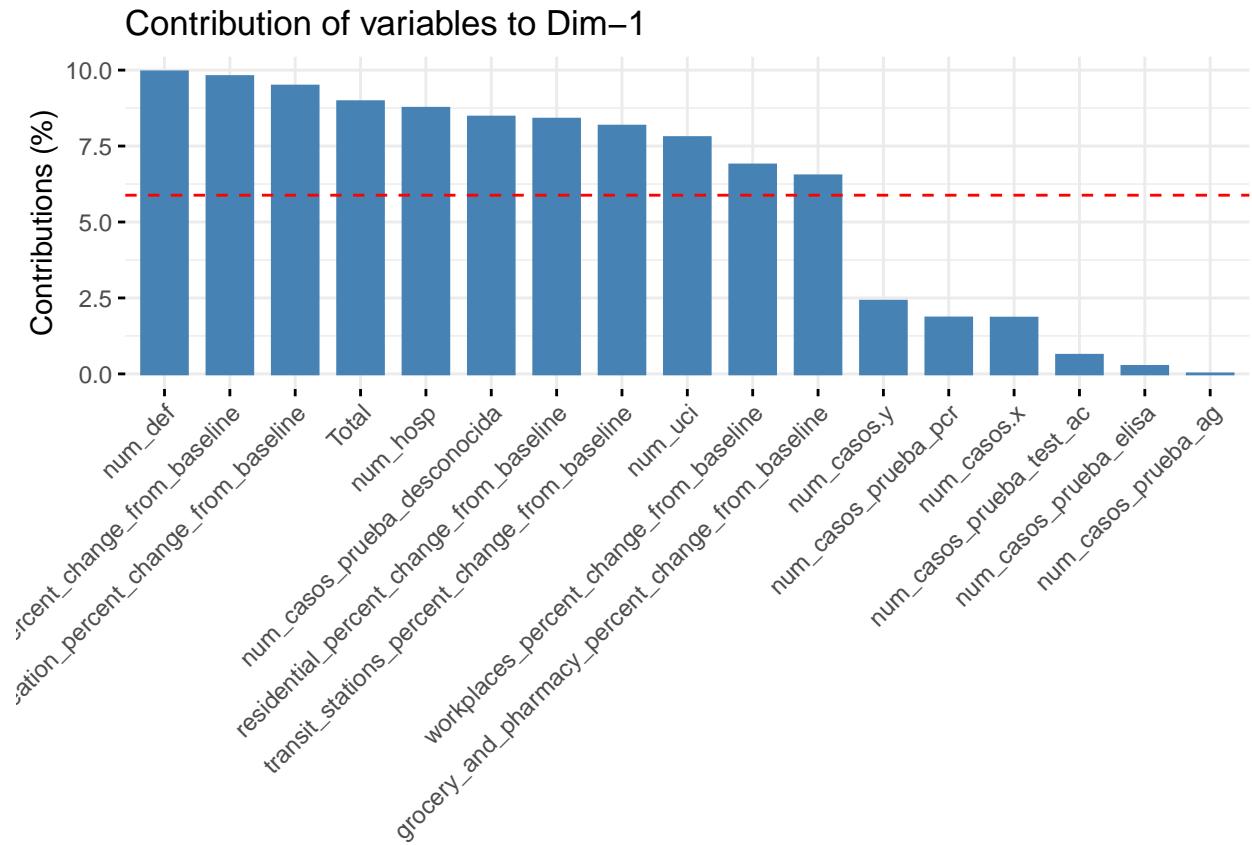
```

```

install.packages('factoextra', repos='http://cran.us.r-project.org')
library(factoextra)

# Var contribution for PC1-PC5
fviz_contrib(pca, choice = "var", axes = 1)

```

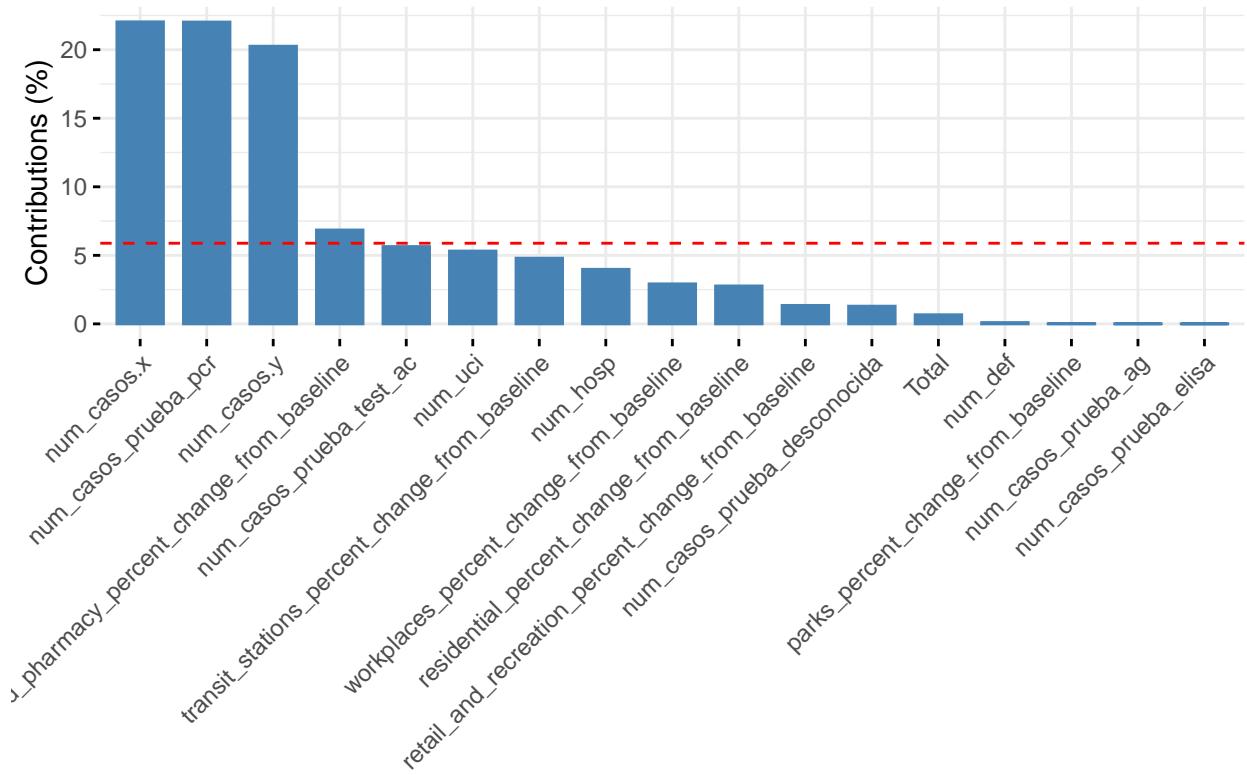


```

fviz_contrib(pca, choice = "var", axes = 2)

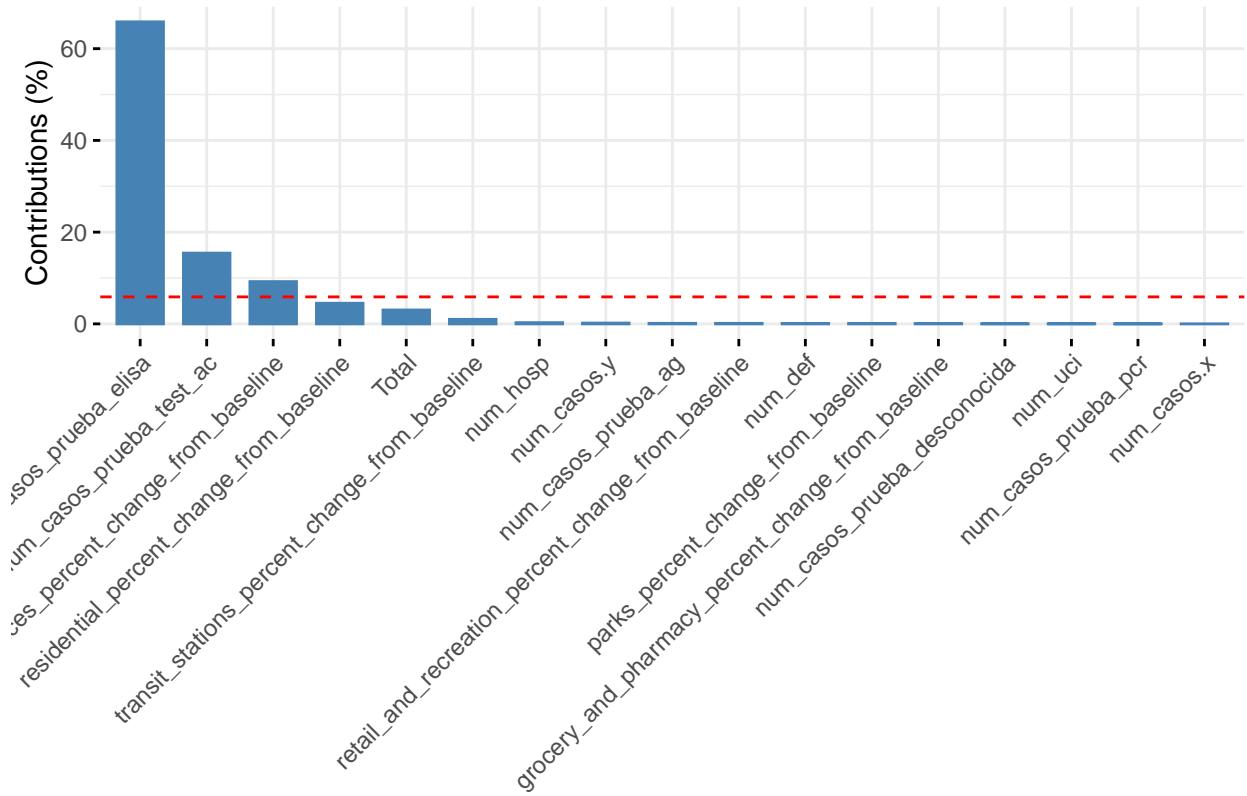
```

Contribution of variables to Dim-2



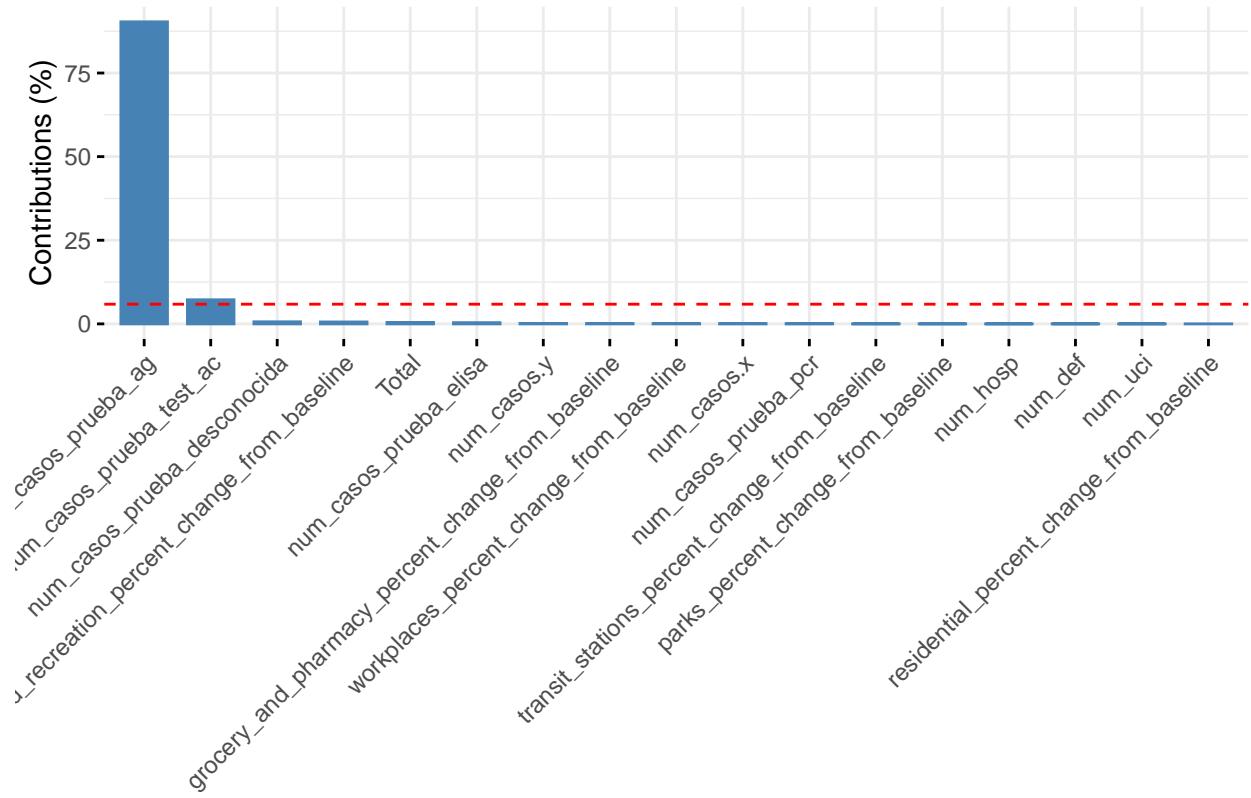
```
fviz_contrib(pca, choice = "var", axes = 3)
```

Contribution of variables to Dim-3



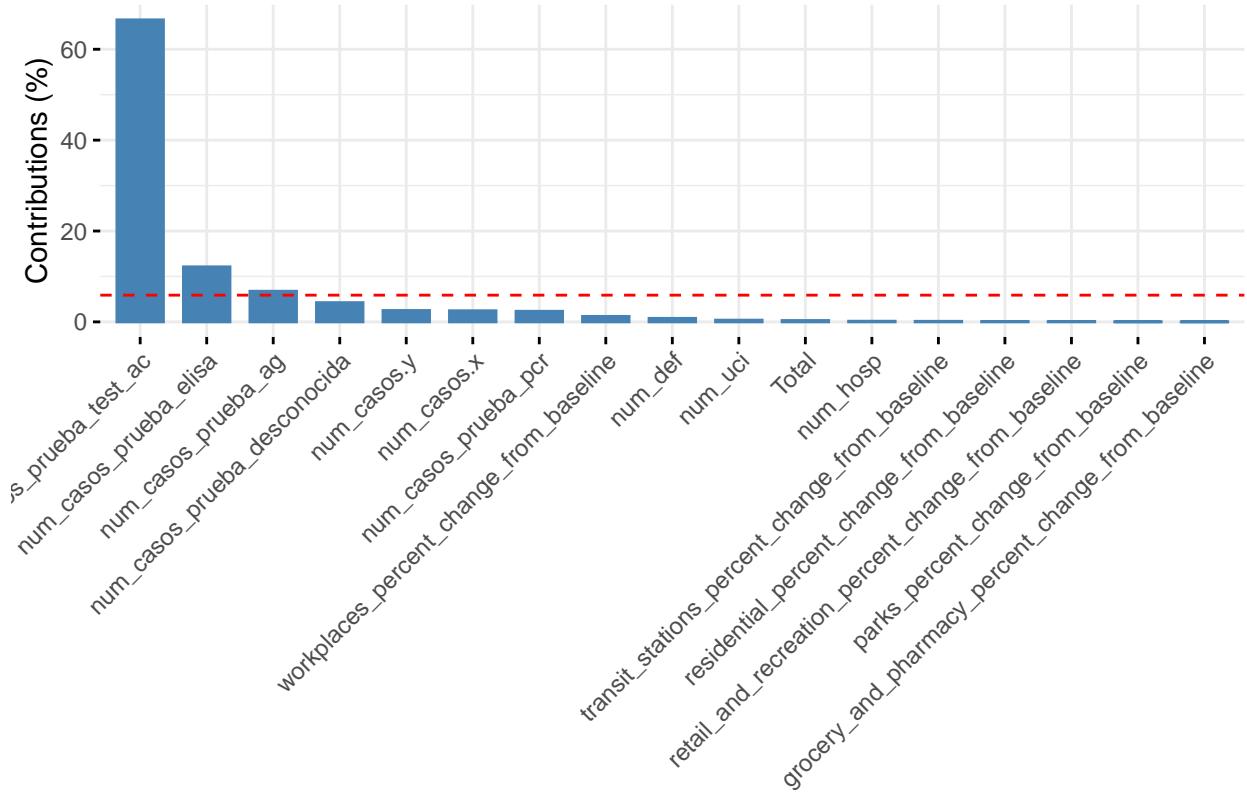
```
fviz_contrib(pca, choice = "var", axes = 4)
```

Contribution of variables to Dim-4



```
fviz_contrib(pca, choice = "var", axes = 5)
```

Contribution of variables to Dim-5



2.3.5 Review normality (Barcelona)

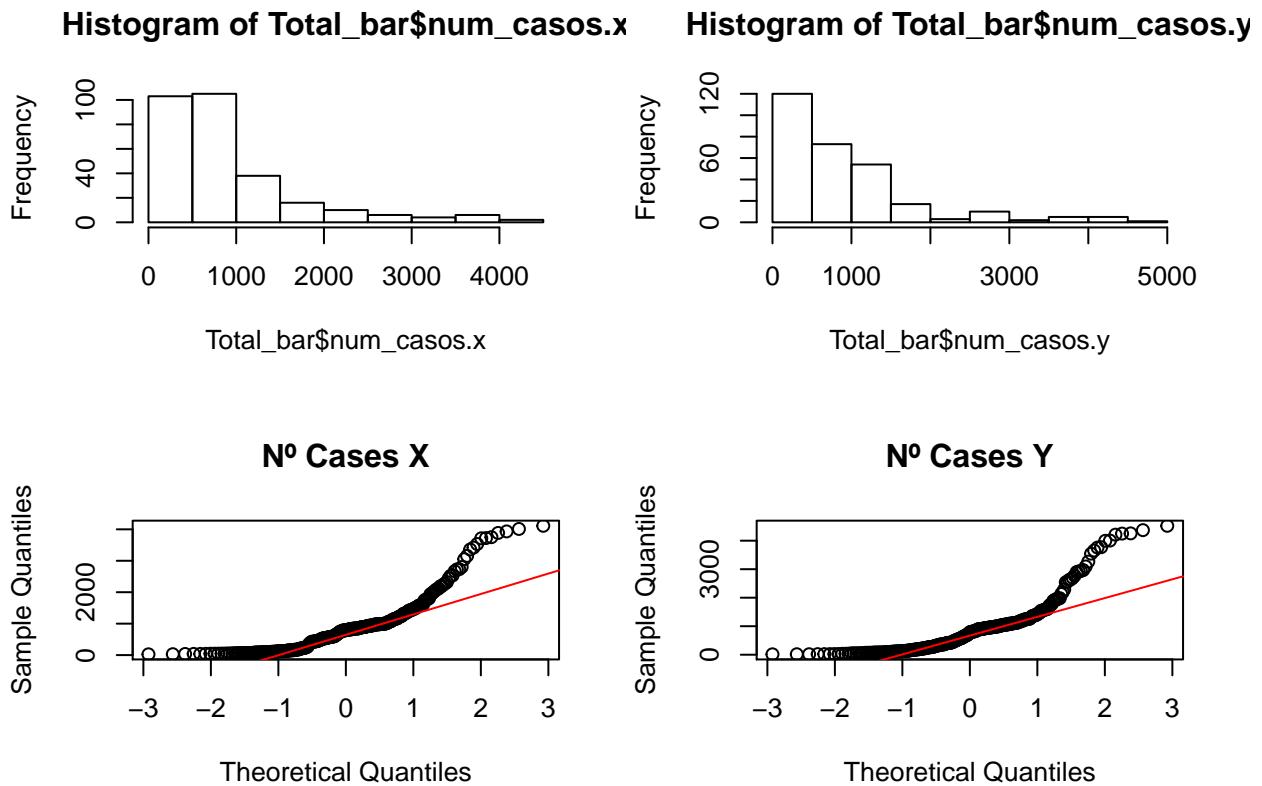
```
# Check for Barcelona
# Raw
Total %>%
  filter(sub_region_2 == "Barcelona") -> Total_bar

par(mfrow=c(2,2))

hist(Total_bar$num_casos.x)
hist(Total_bar$num_casos.y)

qqnorm(Total_bar$num_casos.x, main="Nº Cases X")
qqline(Total_bar$num_casos.x,col=2)

qqnorm(Total_bar$num_casos.y, main="Nº Cases Y")
qqline(Total_bar$num_casos.y,col=2)
```

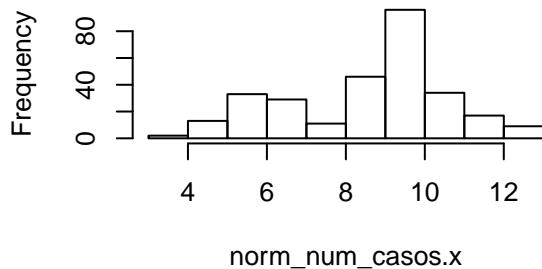
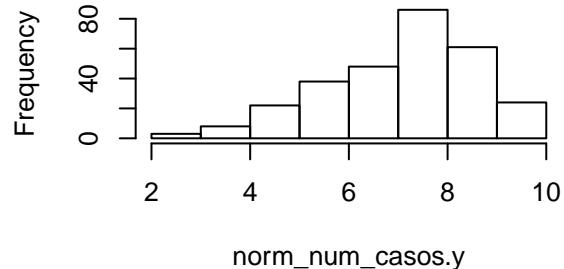
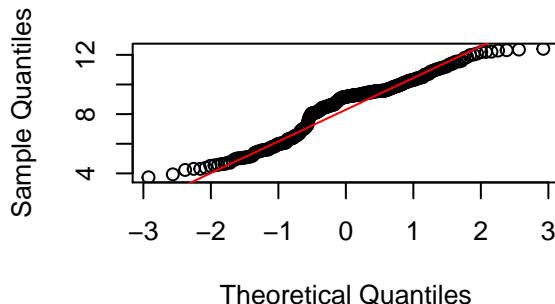
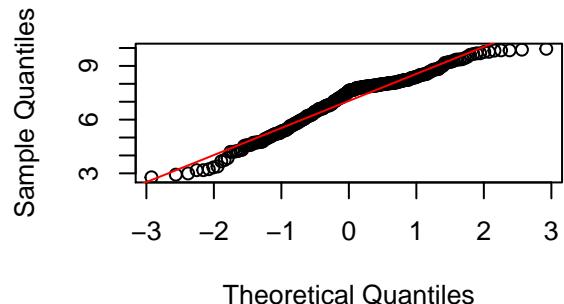


```
# Normalize
library(DescTools)
norm_num_casos.x <- BoxCox(Total_bar$num_casos.x, lambda =
                           BoxCoxLambda(Total_bar$num_casos.x))
norm_num_casos.y <- BoxCox(Total_bar$num_casos.y, lambda =
                           BoxCoxLambda(Total_bar$num_casos.y))

hist(norm_num_casos.x)
hist(norm_num_casos.y)

qqnorm(norm_num_casos.x, main="Nº Cases X")
qqline(norm_num_casos.x,col=2)

qqnorm(norm_num_casos.y, main="Nº Cases Y")
qqline(norm_num_casos.y,col=2)
```

Histogram of norm_num_casos.x**Histogram of norm_num_casos.y****Nº Cases X****Nº Cases Y**

```
# Columns removal according PCA results and SME knowledge
Total <- Total[c(-3,-6,-7,-8,-10)]
Total_bar <- Total_bar[c(-3,-6,-7,-8,-10)]
Total_ts <- Total_ts[c(-4,-5,-6,-8)]
Total_ts_b <- Total_ts_b[c(-4,-5,-6,-8)]
table(Total_ts$sub_region_2)
```

##	Albacete	Alicante/Alacant	Almería
##	290	290	290
##	Araba/Álava	Asturias	Ávila
##	290	290	290
##	Badajoz	Balears, Illes	Barcelona
##	290	290	290
##	Bizkaia	Burgos	Cáceres
##	290	290	290
##	Cádiz	Cantabria	Castellón/Castelló
##	290	290	290
##	Ceuta	Ciudad Real	Córdoba
##	290	290	290
##	Coruña, A	Cuenca	Gipuzkoa
##	290	290	290
##	Girona	Granada	Guadalajara
##	290	290	290
##	Huelva	Huesca	Jaén
##	290	290	290

```

##          León           Lleida          Lugo
##          290            290            290
##          Madrid         Málaga         Melilla
##          290            290            290
##          Murcia        Navarra        Ourense
##          290            290            290
##          Palencia      Palmas, Las Pontevedra
##          290            290            290
##          Rioja, La    Salamanca Santa Cruz de Tenerife
##          290            290            290
##          Segovia       Sevilla        Soria
##          290            290            290
##          Tarragona     Teruel        Toledo
##          290            290            290
##          Valencia/València Valladolid Zamora
##          290            290            290
##          Zaragoza
##          290

#str(Total_ts)
summary(Total_ts)

##   sub_region_2      num_casos.x  num_casos_prueba_pcr
## Length:15080      Min. : 0      Min. : 0.0
## Class :character  1st Qu.: 5    1st Qu.: 5.0
## Mode  :character Median : 39   Median : 35.0
##                  Mean  : 126  Mean  : 110.2
##                  3rd Qu.: 120  3rd Qu.: 105.0
##                  Max. :6565  Max. :6546.0
##   num_casos_prueba_desconocida  num_hosp      num_uci
##   Min. : 0.0000      Min. : 0.00  Min. : 0.000
##   1st Qu.: 0.0000      1st Qu.: 1.00  1st Qu.: 0.000
##   Median : 0.0000      Median : 4.00  Median : 0.000
##   Mean   : 0.1317      Mean   : 14.86  Mean   : 1.281
##   3rd Qu.: 0.0000      3rd Qu.: 12.00  3rd Qu.: 1.000
##   Max.  :65.0000      Max.  :1930.00  Max.  :135.000
##   num_def      retail_and_recreation_percent_change_from_baseline
##   Min. : 0.000  Min. :-97.00
##   1st Qu.: 0.000  1st Qu.:-57.00
##   Median : 1.000  Median :-30.00
##   Mean   : 3.437  Mean  :-37.29
##   3rd Qu.: 3.000  3rd Qu.:-17.00
##   Max.  :334.000  Max.  : 71.00
##   grocery_and_pharmacy_percent_change_from_baseline
##   Min.  :-96.00
##   1st Qu.:-24.00
##   Median : -6.00
##   Mean   : -11.75
##   3rd Qu.:  4.00
##   Max.   :194.00
##   parks_percent_change_from_baseline
##   Min.  :-94.000
##   1st Qu.:-30.000
##   Median : -2.000
##   Mean   :  5.809

```

```

## 3rd Qu.: 30.000
## Max.    :543.000
## transit_stations_percent_change_from_baseline
## Min.    :-100.00
## 1st Qu.: -53.00
## Median  : -31.00
## Mean    : -35.19
## 3rd Qu.: -17.00
## Max.    : 74.00
## workplaces_percent_change_from_baseline
## Min.    :-92.00
## 1st Qu.: -43.00
## Median : -26.00
## Mean   : -29.08
## 3rd Qu.: -13.00
## Max.    : 55.00
## residential_percent_change_from_baseline      Total          Dia_c
## Min.    :-10.00                                Min.   : 1.95  Min.   :2020-03-16
## 1st Qu.:  4.00                                1st Qu.:11.36  1st Qu.:2020-05-27
## Median  :  7.00                                Median :14.39  Median :2020-08-07
## Mean    : 10.14                                Mean   :14.20  Mean   :2020-08-07
## 3rd Qu.: 14.00                                3rd Qu.:17.11  3rd Qu.:2020-10-19
## Max.    : 48.00                                Max.   :29.00  Max.   :2020-12-30

```

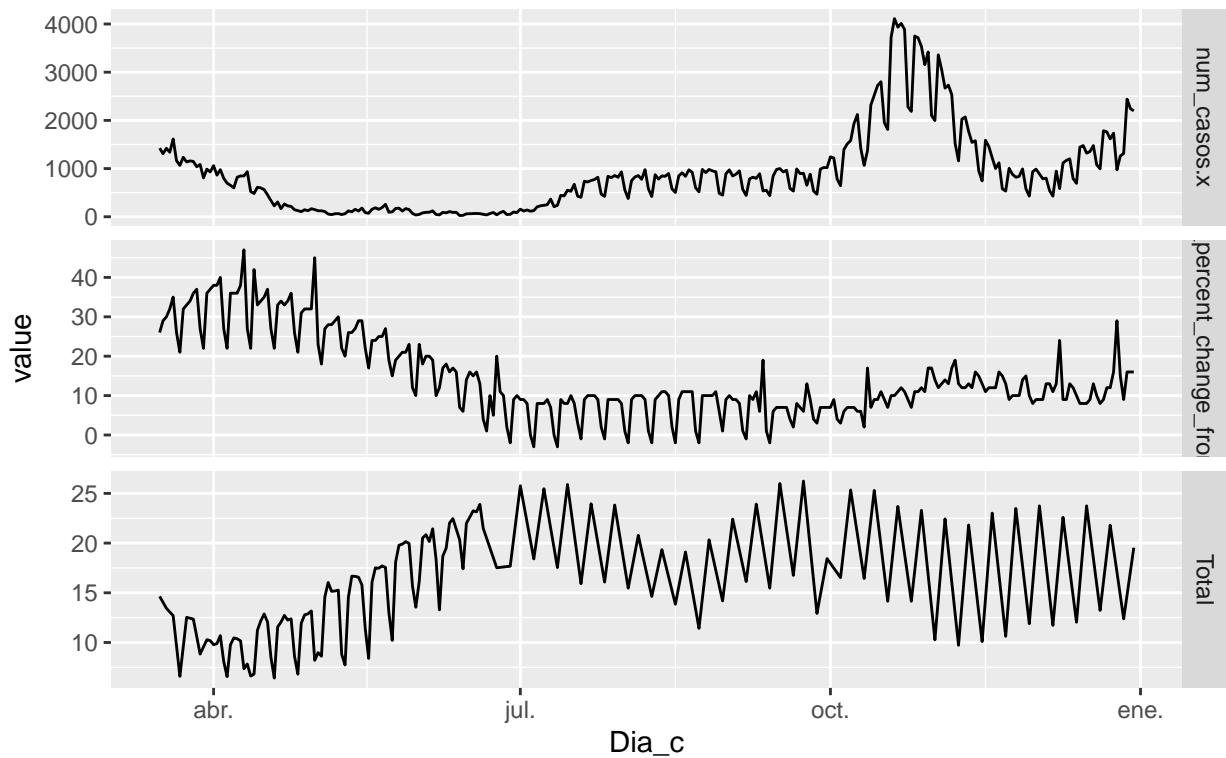
2.3.6 Final plots (Barcelona and others)

```

# Total_ts plots for the selected variables
# % of mobility reported by INE and Google (EM3 study)
Total_ts_b %>%
  filter(sub_region_2 == "Barcelona") %>%
  pivot_longer(c(2,13,14)) %>%
  ggplot(aes(x = Dia_c, y = value)) +
  geom_line() +
  facet_grid(vars(name), scales = "free_y")+
  labs(title = "Bar - N° cases (CNE) vs % Residentail (Google) and Tot (INE)
               mobility change")

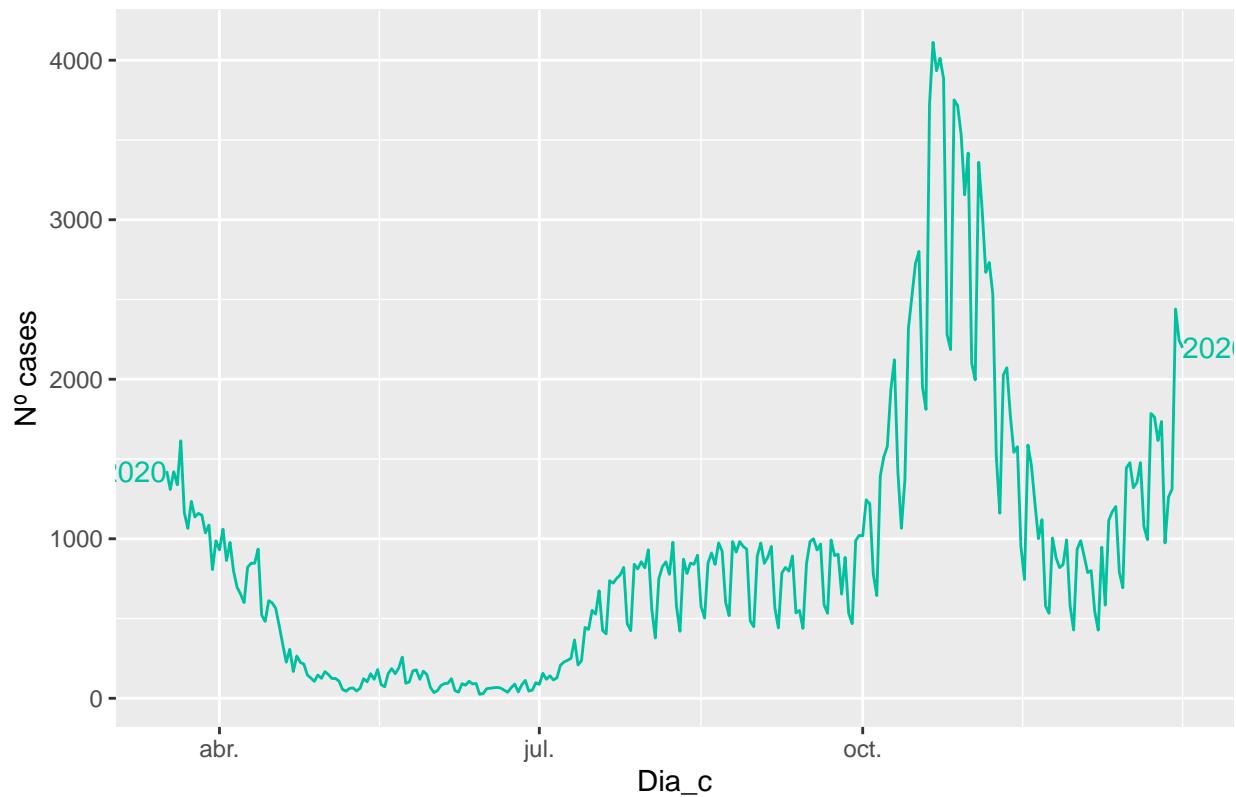
```

Bar – N° cases (CNE) vs % Residential (Google) and Tot (INE)
mobility change



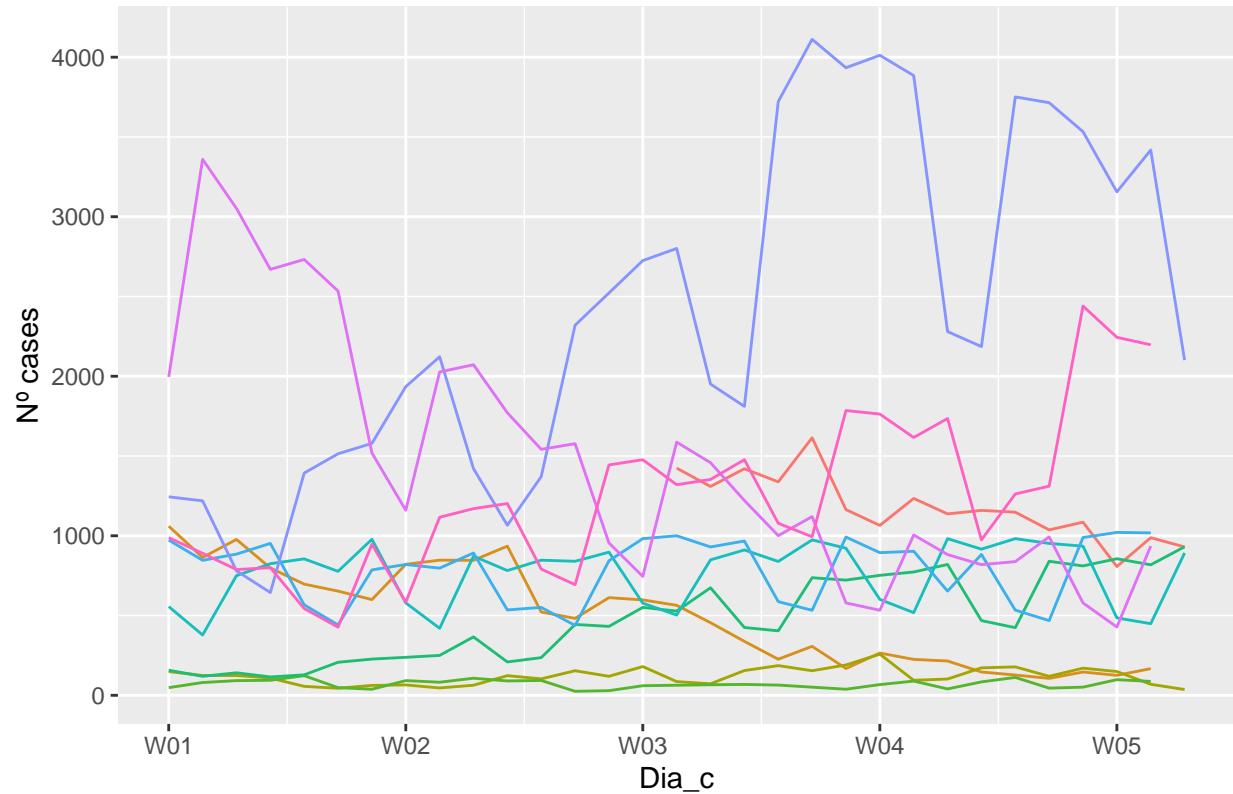
```
# Barcelona Seasonal plot: N° cases
Total_ts_b %>%
  filter(sub_region_2 == "Barcelona") %>%
  gg_season(num_casos.x, labels = "both") +
  labs(y = "Nº cases",
       title = "Barcelona Seasonal plot: Nº cases")
```

Barcelona Seasonal plot: N° cases



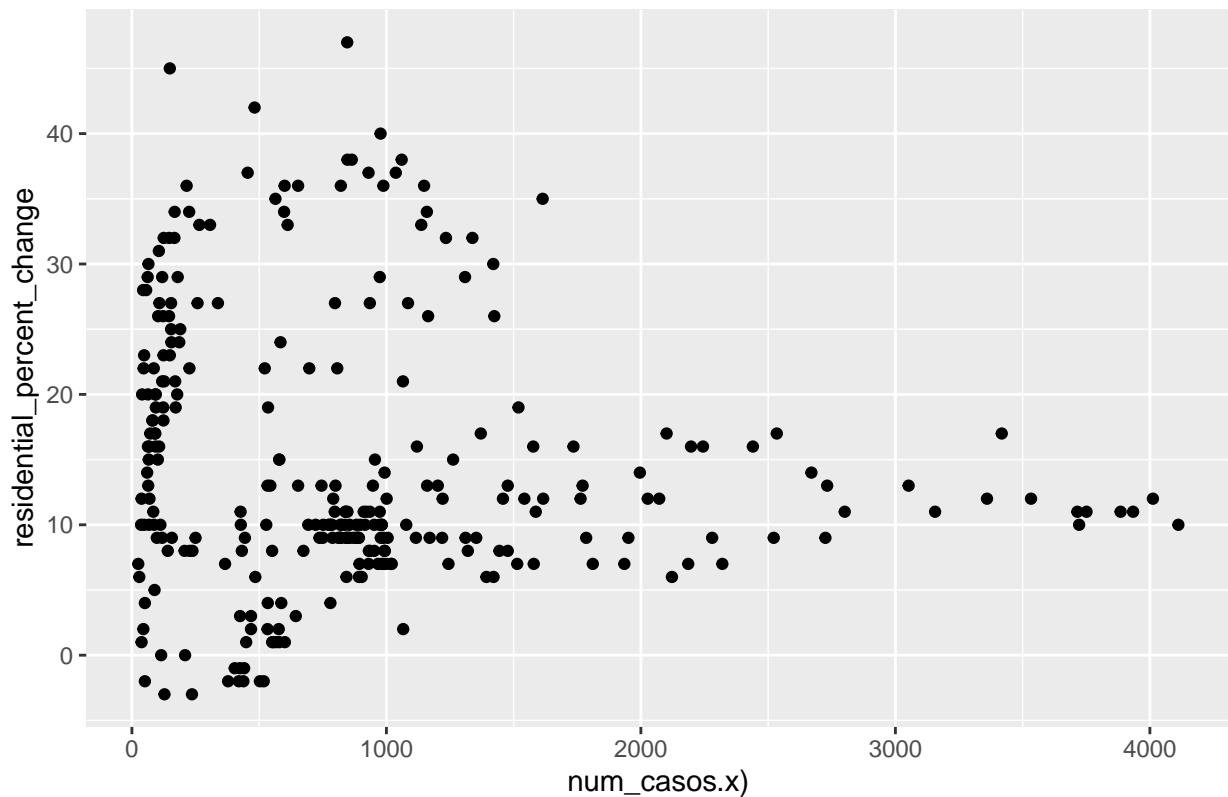
```
# Barcelona Seasonal plot: N° cases by month
Total_ts_b %>%
  filter(sub_region_2 == "Barcelona") %>%
  gg_season(num_casos.x, period = "month") +
  theme(legend.position = "none") +
  labs(y="Nº cases", title="Barcelona Seasonal plot: N° cases - month")
```

Barcelona Seasonal plot: N° cases – month



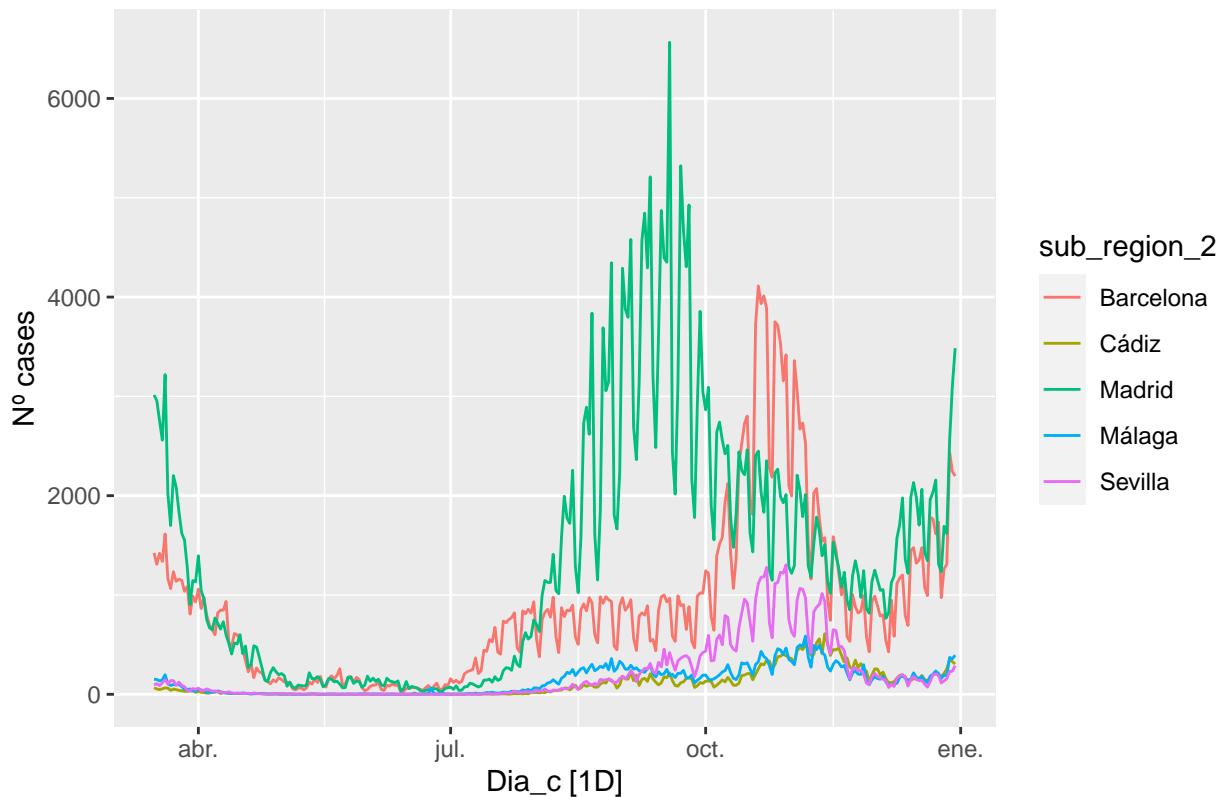
```
# Barcelona scatter plot N° cases vs residential_percent_change
Total_ts_b %>%
  filter(sub_region_2 == "Barcelona") %>%
  ggplot(aes(x = num_casos.x, y = residential_percent_change_from_baseline )) +
  geom_point() +
  labs(x = "num_casos.x",
       y = "residential_percent_change",
       title="Barcelona scatter plot N° cases vs residential_percent_change")
```

Barcelona scatter plot N° cases vs residential_percent_change

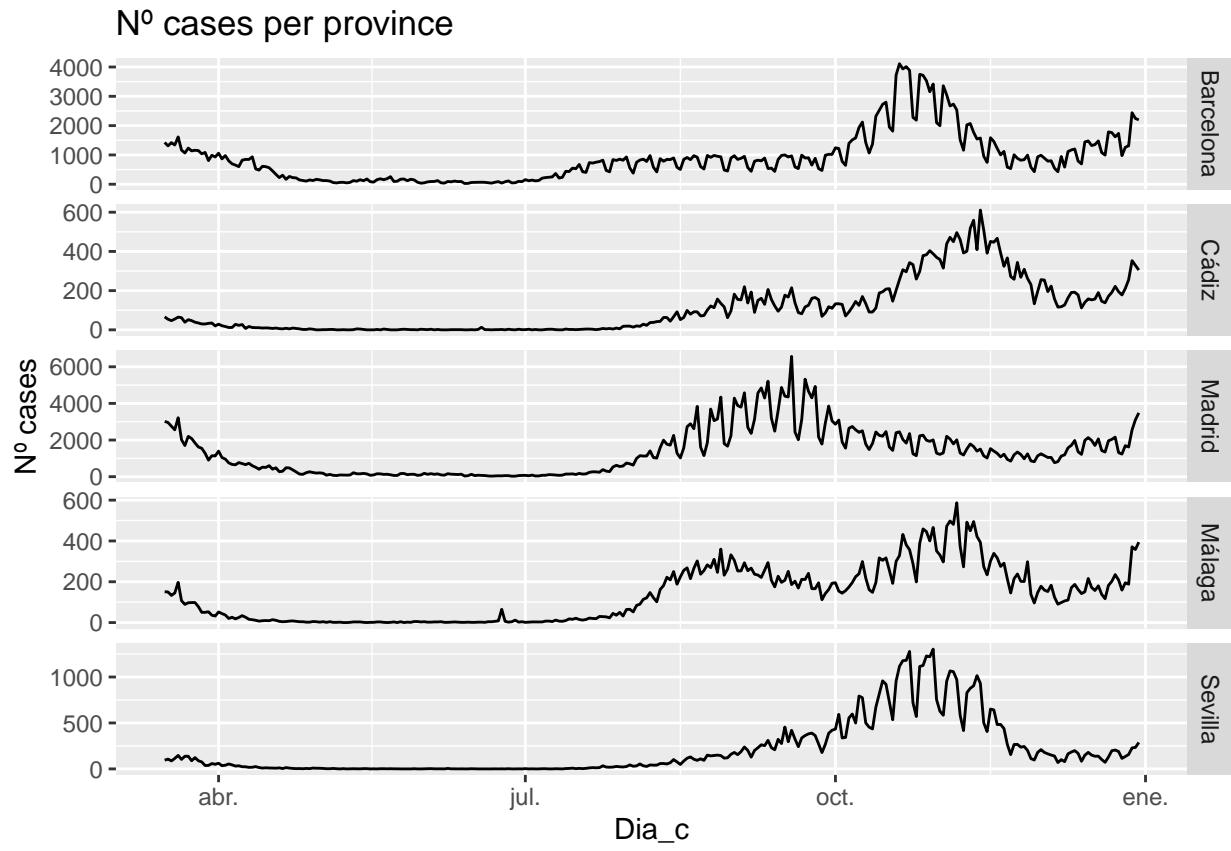


```
#####
# A - N° cases per province (Barcelona, Madrid, Málaga, Córdoba and Cádiz)
autoplot(Total_ts_b, num_casos.x) +
  labs(y = "N° cases",
       title = "N° cases per province")
```

Nº cases per province

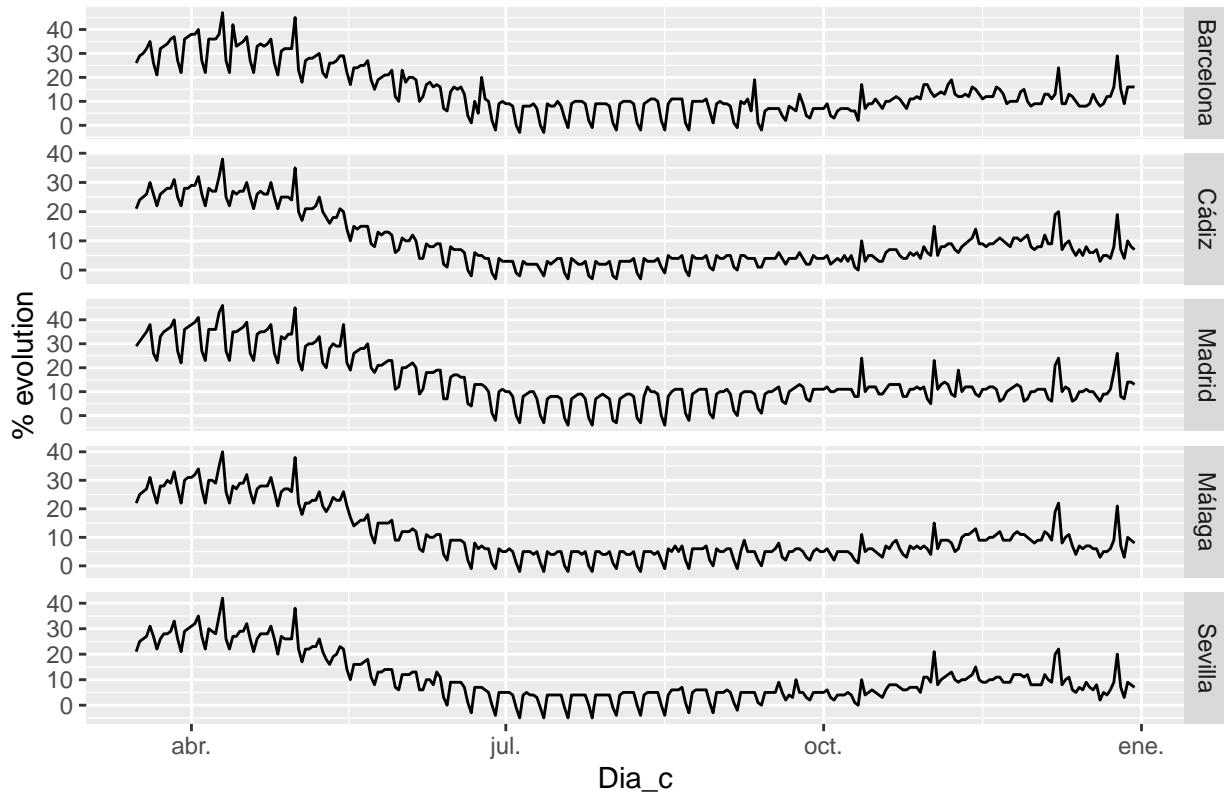


```
# B - Nº cases per province (Barcelona, Madrid, Málaga, Córdoba and Cádiz)
Total_ts_b %>%
  group_by(sub_region_2) %>%
  summarise(CASOS = sum(num_casos.x))%>%
  ggplot(aes(x = Dia_c, y = CASOS)) +
  geom_line() +
  facet_grid(vars(sub_region_2), scales = "free_y") +
  labs(title = "Nº cases per province", y= "Nº cases")
```



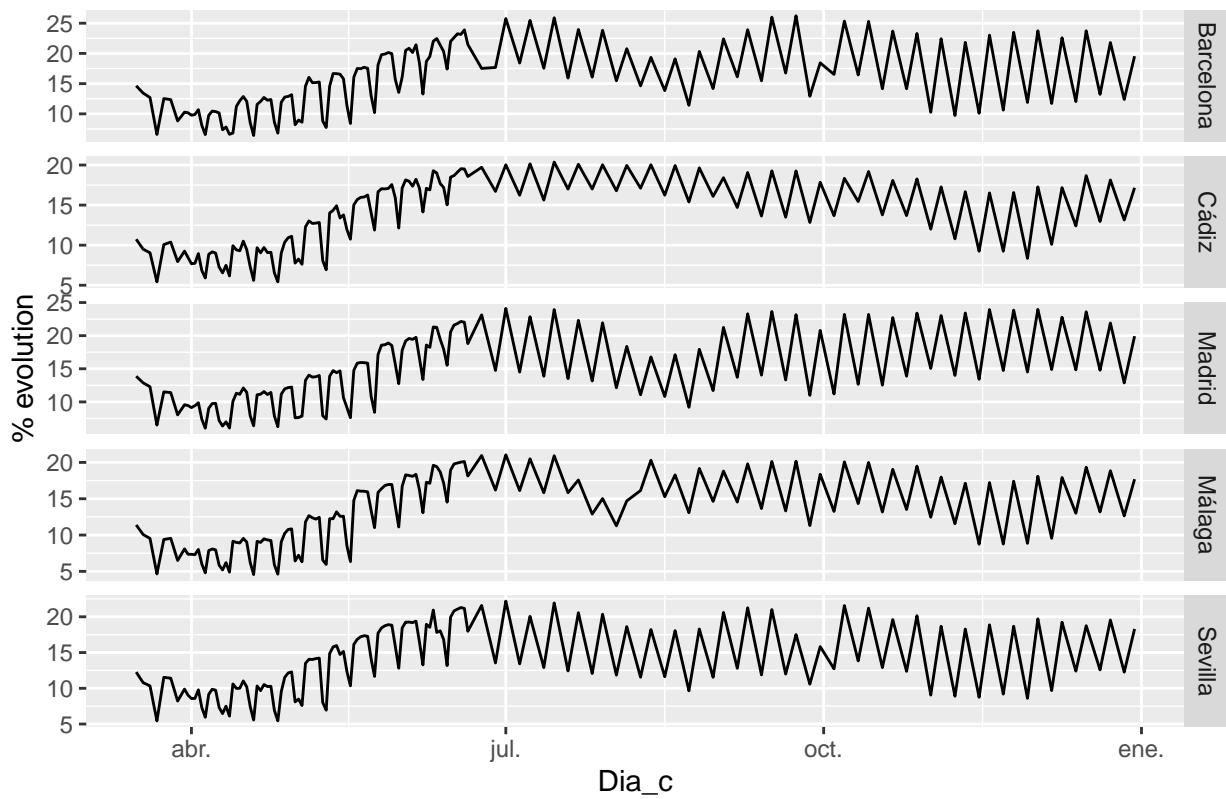
```
# B.b - Google % change residential mobility per province (Barcelona, Madrid, Málaga, Cádiz and Sevilla)
Total_ts_b %>%
  group_by(sub_region_2) %>%
  summarise(per_c = (residential_percent_change_from_baseline))%>%
  ggplot(aes(x = Dia_c, y = per_c)) +
  geom_line() +
  facet_grid(vars(sub_region_2), scales = "free_y") +
  labs(title = "Google % change residential mobility per province", y= "% evolution")
```

Google % change residential mobility per province



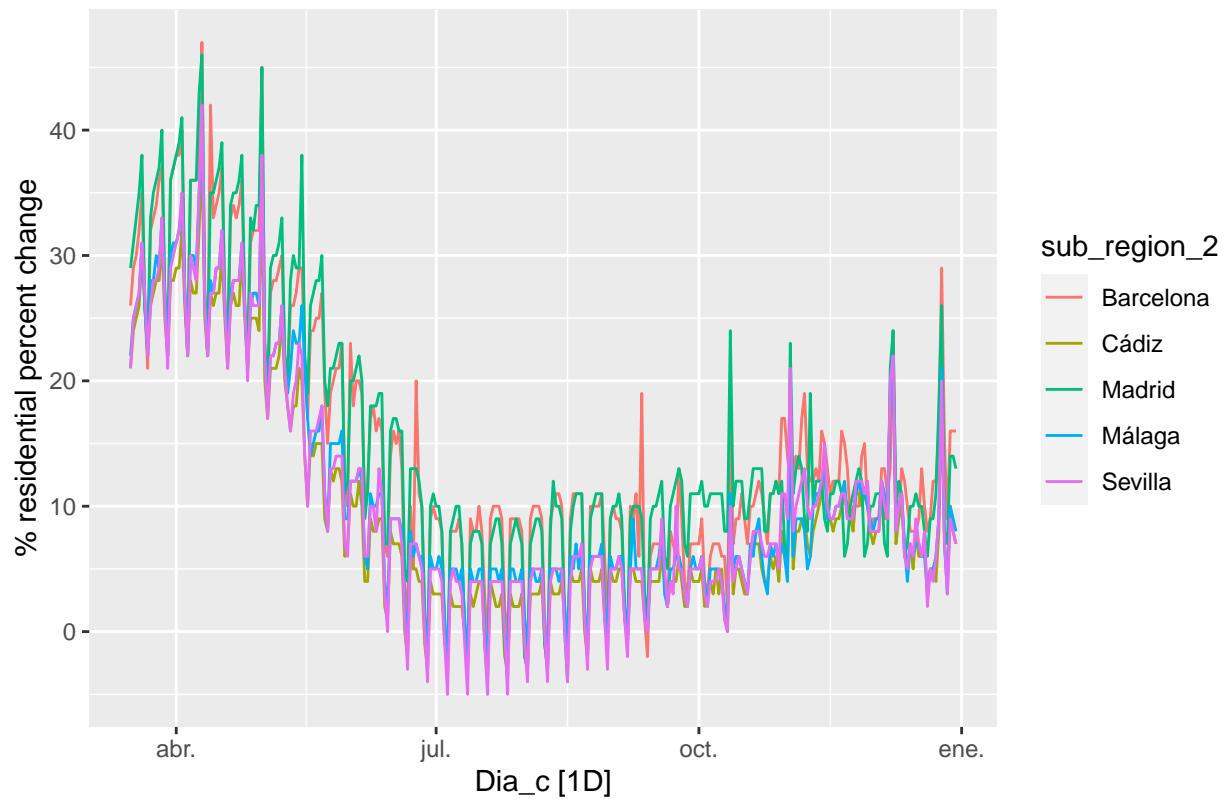
```
# B.c - EM3 % change residential mobility per province (Barcelona, Madrid, Málaga, Cádiz and Sevilla)
Total_ts_b %>%
  group_by(sub_region_2) %>%
  summarise(per_c = (Total))%>%
  ggplot(aes(x = Dia_c, y = per_c)) +
  geom_line() +
  facet_grid(vars(sub_region_2), scales = "free_y") +
  labs(title = "EM3 % change residential mobility per province", y= "% evolution")
```

EM3 % change residential mobility per province



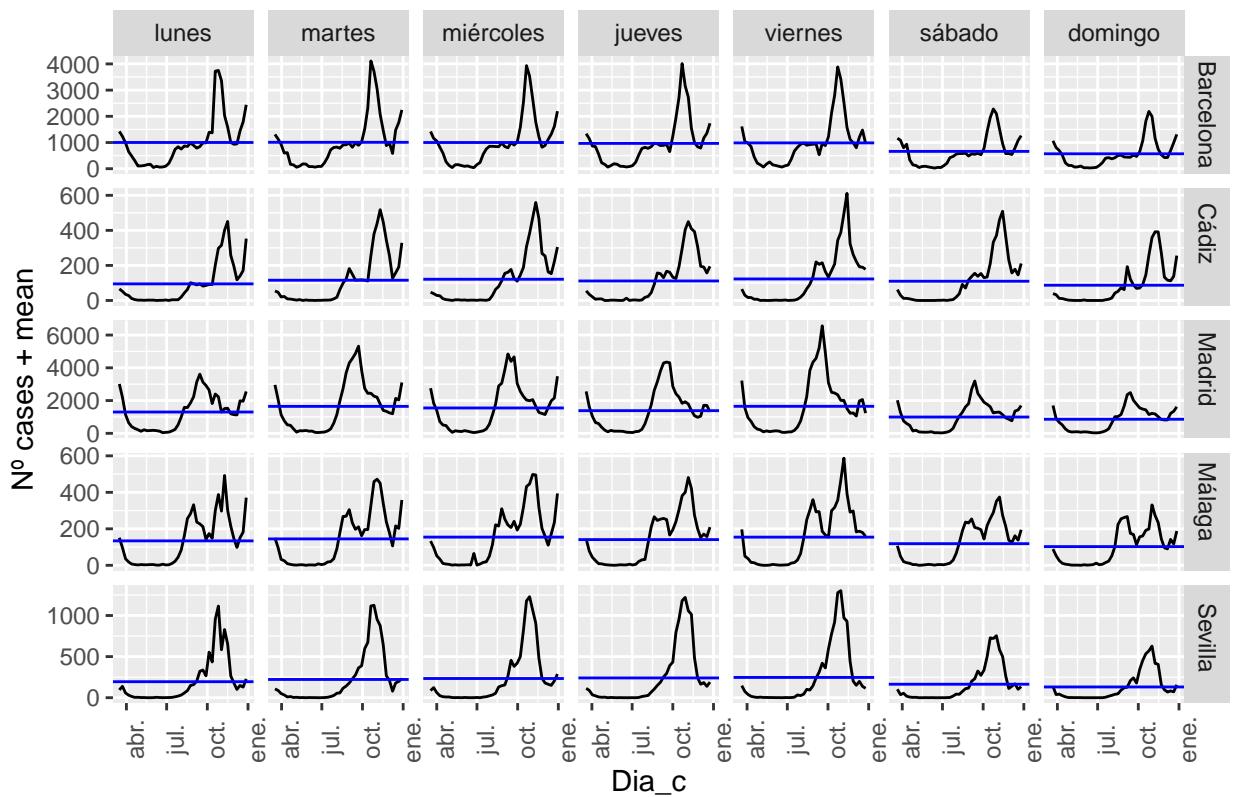
```
# % residential percent change (Barcelona, Madrid, Málaga, Cádiz and Sevilla)
autoplot(Total_ts_b, residential_percent_change_from_baseline ) +
  labs(y = "% residential percent change",
       title = "Residential percent change")
```

Residential percent change

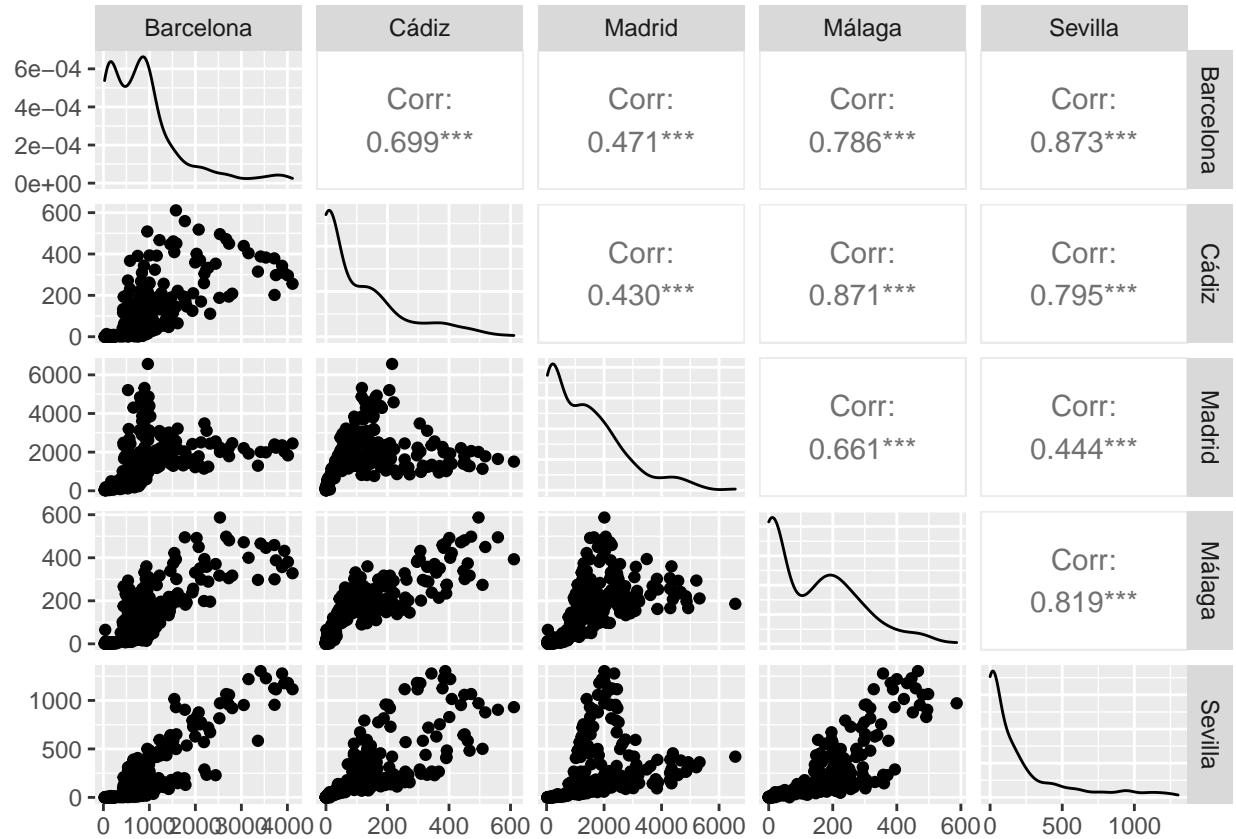


```
# N° cases per province and day of week + mean
Total_ts_b %>%
  gg_subseries(num_casos.x, period = "week") +
  labs(y = "N° cases + mean",
       title = "N° cases per province and day of week + mean")
```

Nº cases per province and day of week + mean

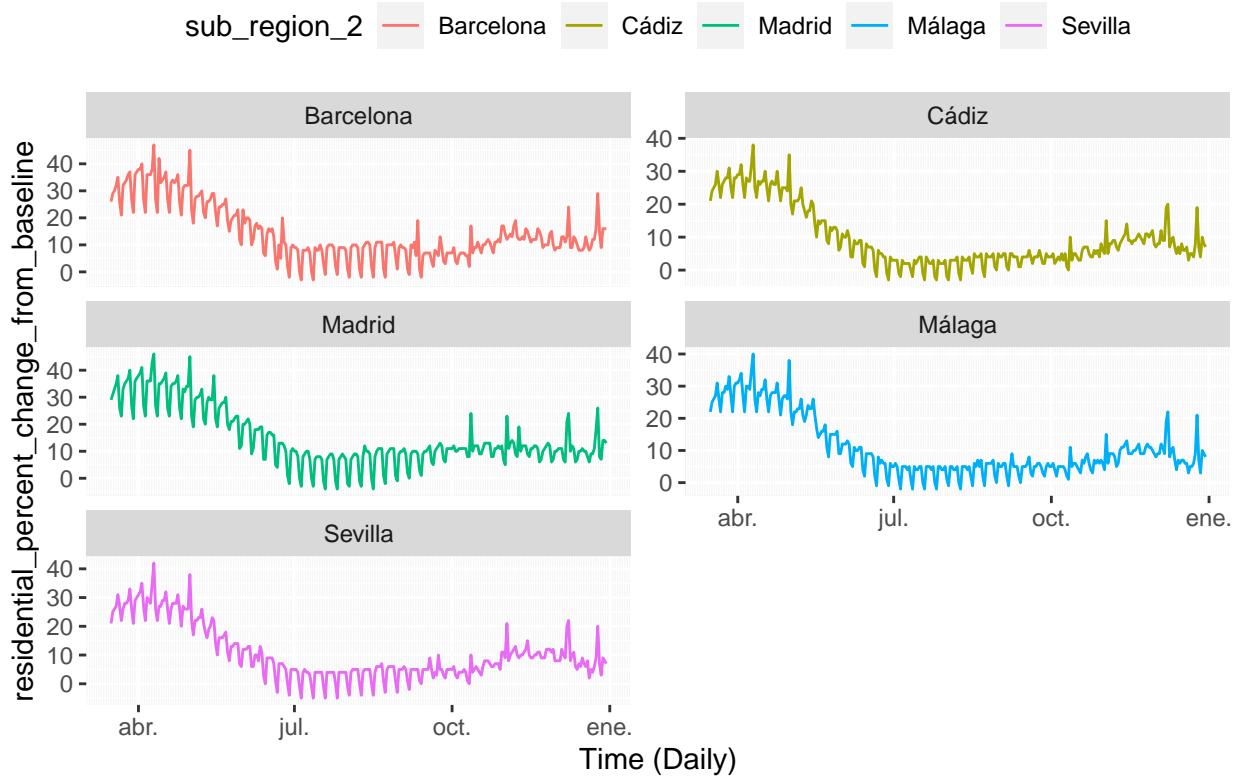


```
# Correlation plot / nº cases by province
Total_ts_b %>%
  group_by(sub_region_2) %>%
  summarise(CASOS = sum(num_casos.x))%>%
  pivot_wider(values_from=CASOS, names_from=sub_region_2) %>%
  GGally::ggpairs(2:6)
```



```
# % of mobility reported by Google - residential_percent_change
autoplot(Total_ts_b, residential_percent_change_from_baseline) +
  facet_wrap(~sub_region_2, scales = "free_y", ncol=2) +
  theme(legend.position = "top") +
  scale_x_date(date_minor_breaks = "1 day", name = "Time (Daily)") +
  ggtitle(label = "% of mobility reported by Google - home (Barcelona, Madrid, Málaga, Cádiz and Sevilla)
```

% of mobility reported by Google – home (Barcelona, Madrid, Málaga, Cádiz, Sevilla)



3 ARIMA - fpp3 library

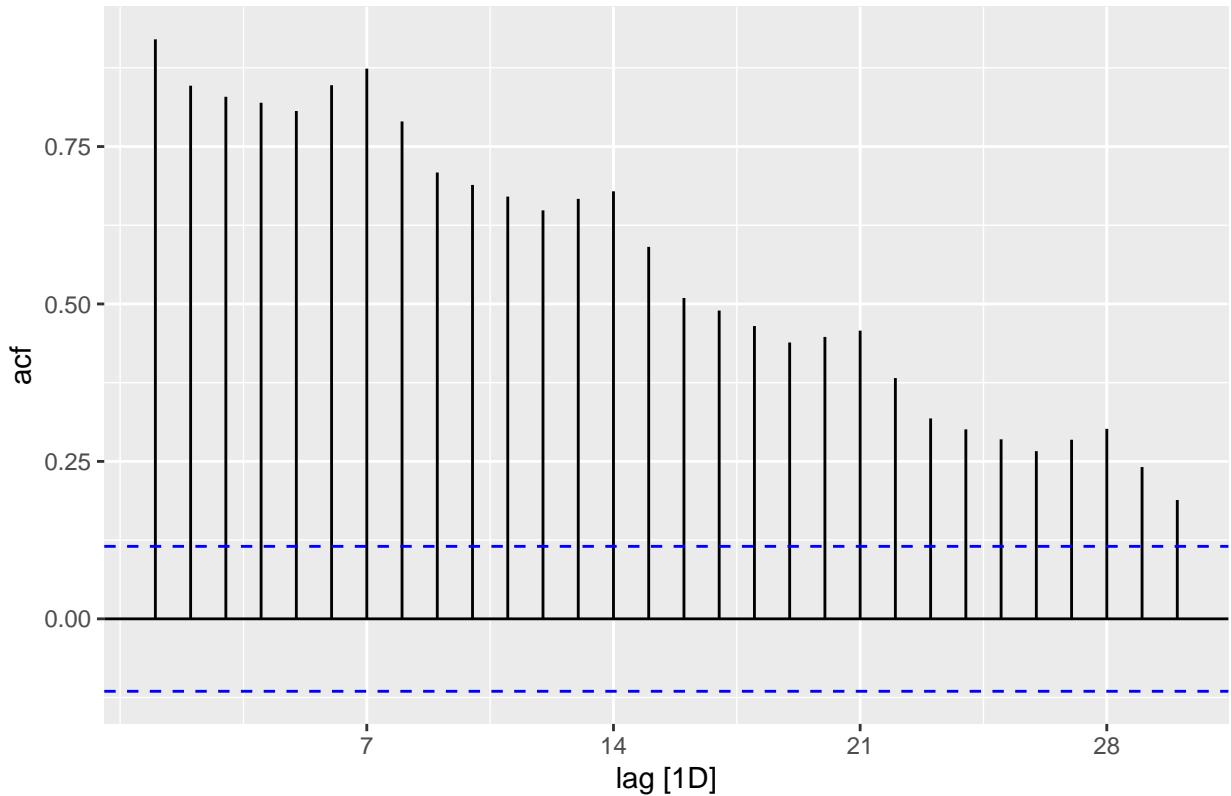
3.1 ACF and PACF (Barcelona, Madrid, Málaga, Córdoba and Cádiz)

As stated by (Hyndman and Athanasopoulos 2021)... “ACF plot is also useful for identifying non-stationary time series. For a stationary time series, the ACF will drop to zero relatively quickly, while the ACF of non-stationary data decreases slowly. Also, for non-stationary data, the value of r_1 is often large and positive... PACF partial autocorrelations. These measure the relationship between y_t and y_{t-k} after removing the effects of lags $1, 2, 3, \dots, k-1$.”

As stated by (Hyndman and Athanasopoulos 2021)... “When data have a trend, the autocorrelations for small lags tend to be large and positive because observations nearby in time are also nearby in value. So the ACF of a trended time series tends to have positive values that slowly decrease as the lags increase. When data are seasonal, the autocorrelations will be larger for the seasonal lags (at multiples of the seasonal period) than for other lags. When data are both trended and seasonal, you see a combination of these effects...”

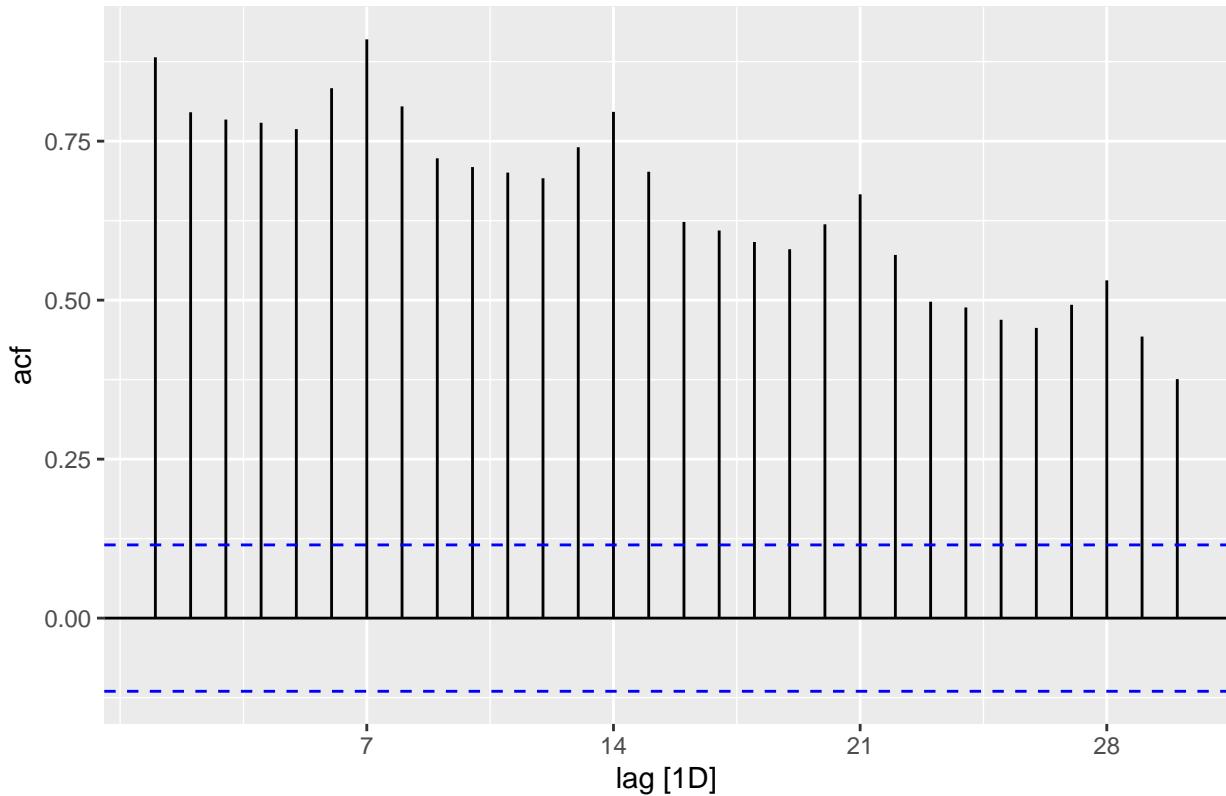
```
# ACF
Total_ts %>%
  filter(sub_region_2 == "Barcelona") %>%
  ACF(num_casos.x, lag_max = 30) %>%
  autoplot() +
  labs(title=" Barcelona - ACF N° Cases")
```

Barcelona – ACF Nº Cases



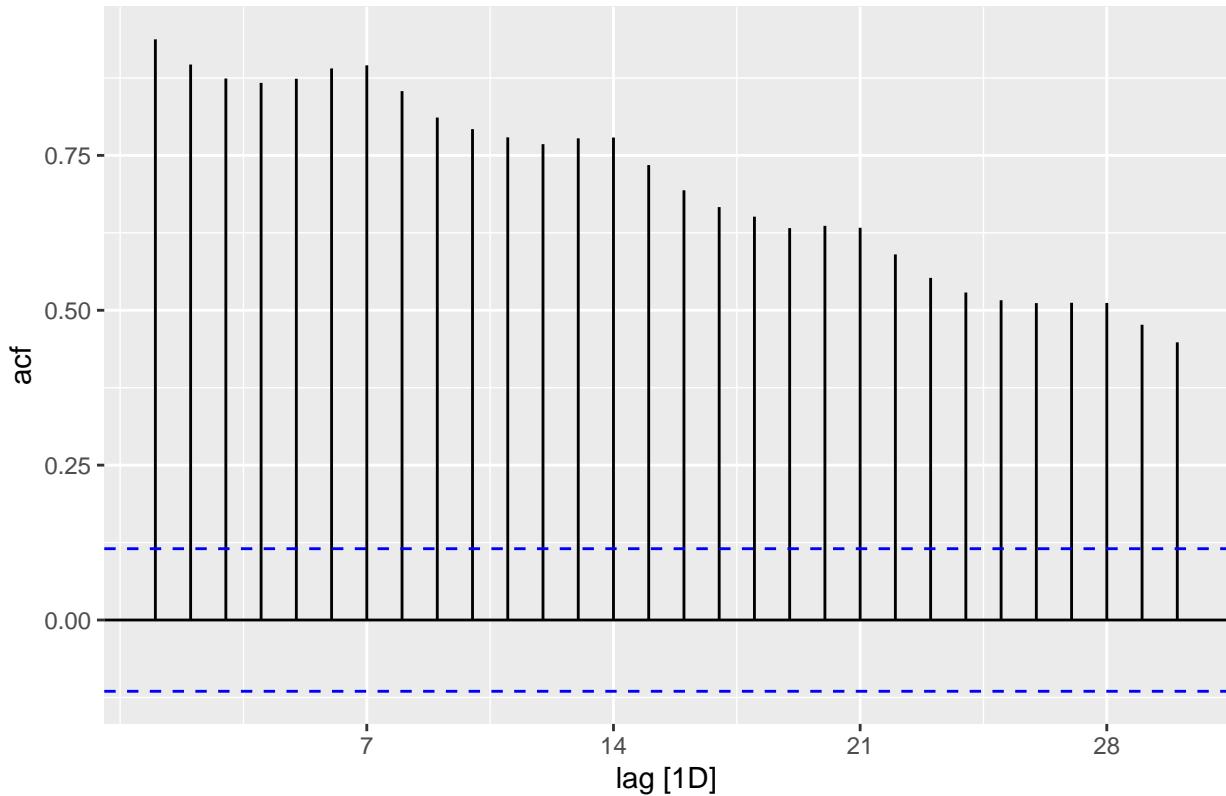
```
Total_ts %>%
  filter(sub_region_2 == "Madrid") %>%
  ACF(num_casos.x, lag_max = 30) %>%
  autoplot() +
  labs(title=" Madrid – ACF Nº Cases")
```

Madrid – ACF N° Cases



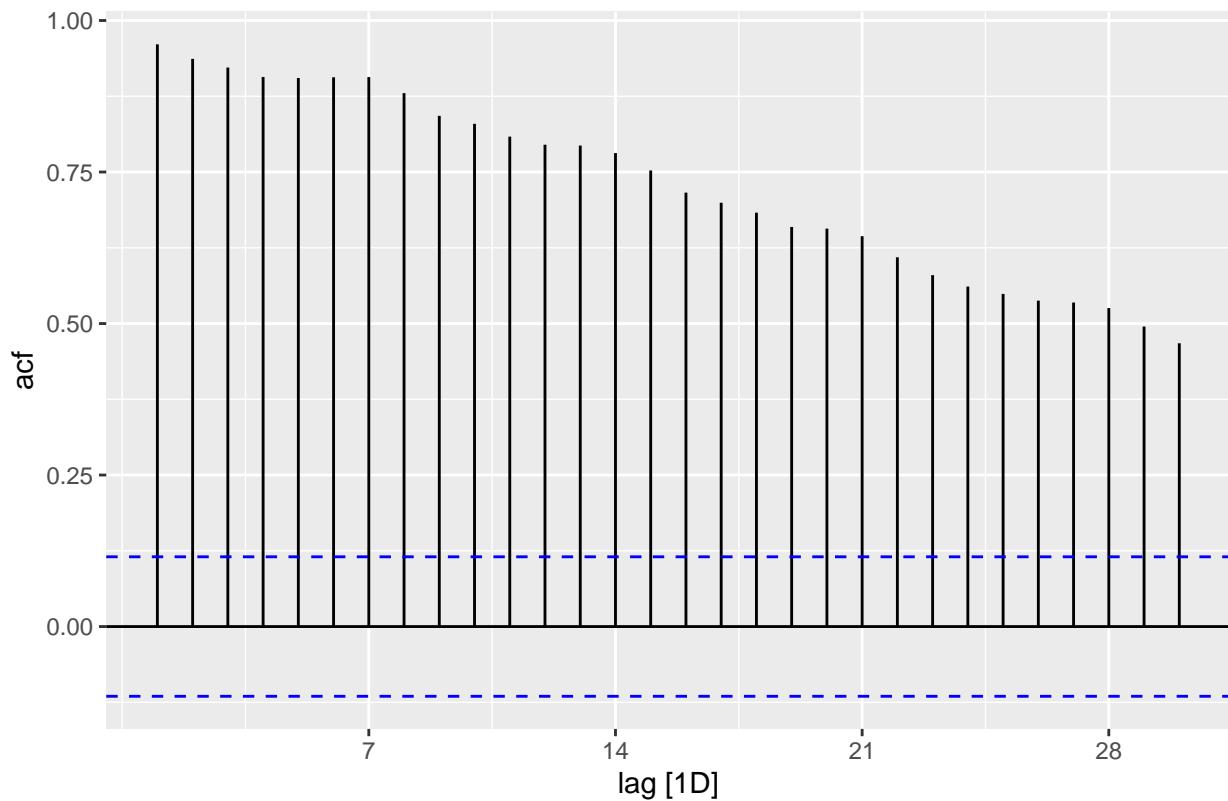
```
Total_ts %>%
  filter(sub_region_2 == "Málaga") %>%
  ACF(num_casos.x, lag_max = 30) %>%
  autoplot() +
  labs(title=" Málaga – ACF N° Cases")
```

Málaga – ACF Nº Cases



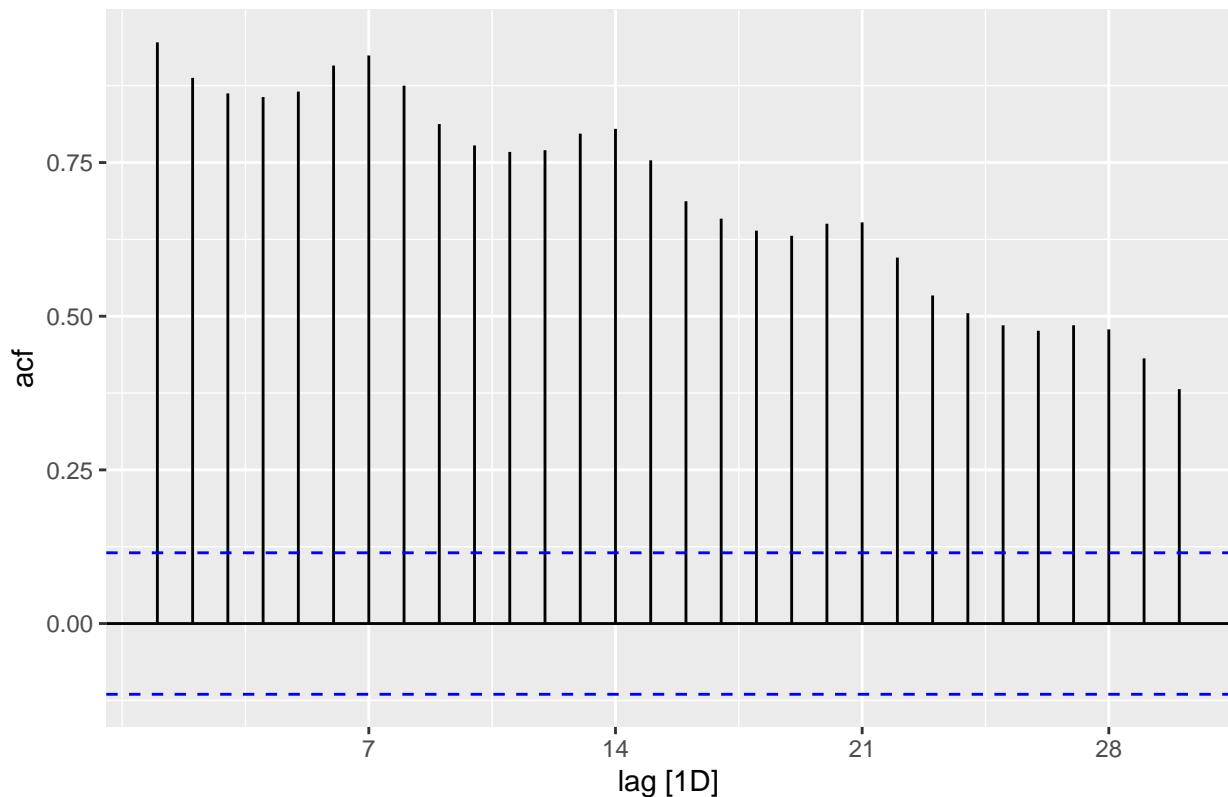
```
Total_ts %>%
  filter(sub_region_2 == "Cádiz") %>%
  ACF(num_casos.x, lag_max = 30) %>%
  autoplot() +
  labs(title=" Cádiz – ACF Nº Cases")
```

Cádiz – ACF N° Cases



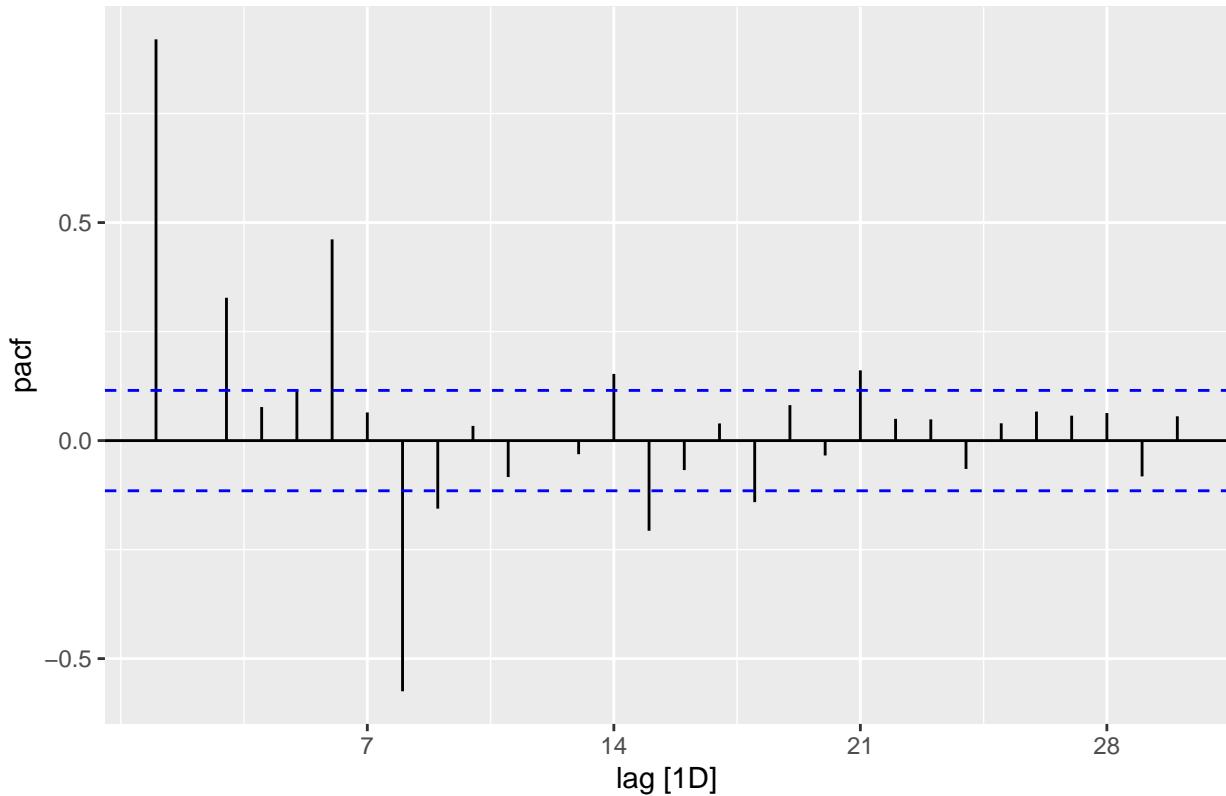
```
Total_ts %>%
  filter(sub_region_2 == "Sevilla") %>%
  ACF(num_casos.x, lag_max = 30) %>%
  autoplot() +
  labs(title=" Sevilla – ACF N° Cases")
```

Sevilla – ACF Nº Cases



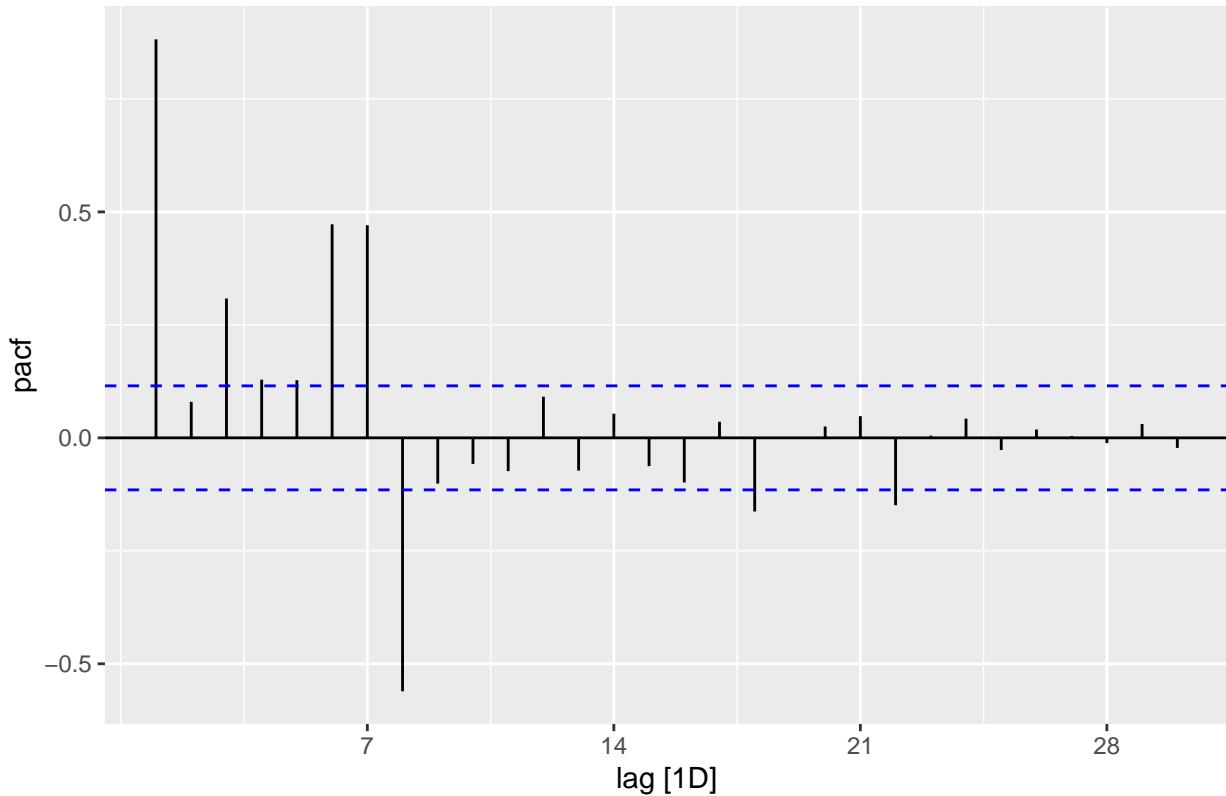
```
# PACF
Total_ts %>%
  filter(sub_region_2 == "Barcelona") %>%
  PACF(num_casos.x, lag_max = 30) %>%
  autoplot() +
  labs(title=" Barcelona – ACF Nº Cases")
```

Barcelona – ACF Nº Cases



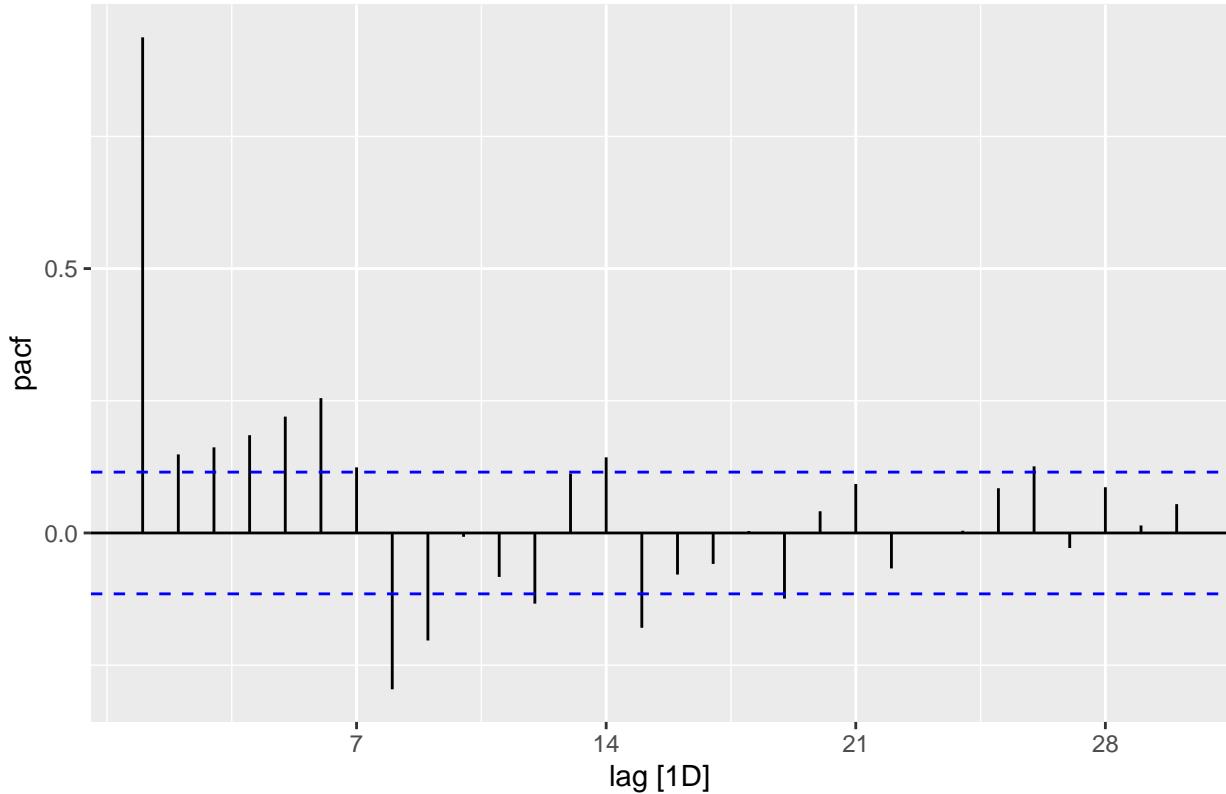
```
Total_ts %>%
  filter(sub_region_2 == "Madrid") %>%
  PACF(num_casos.x, lag_max = 30) %>%
  autoplot() +
  labs(title=" Madrid – ACF Nº Cases")
```

Madrid – ACF N° Cases



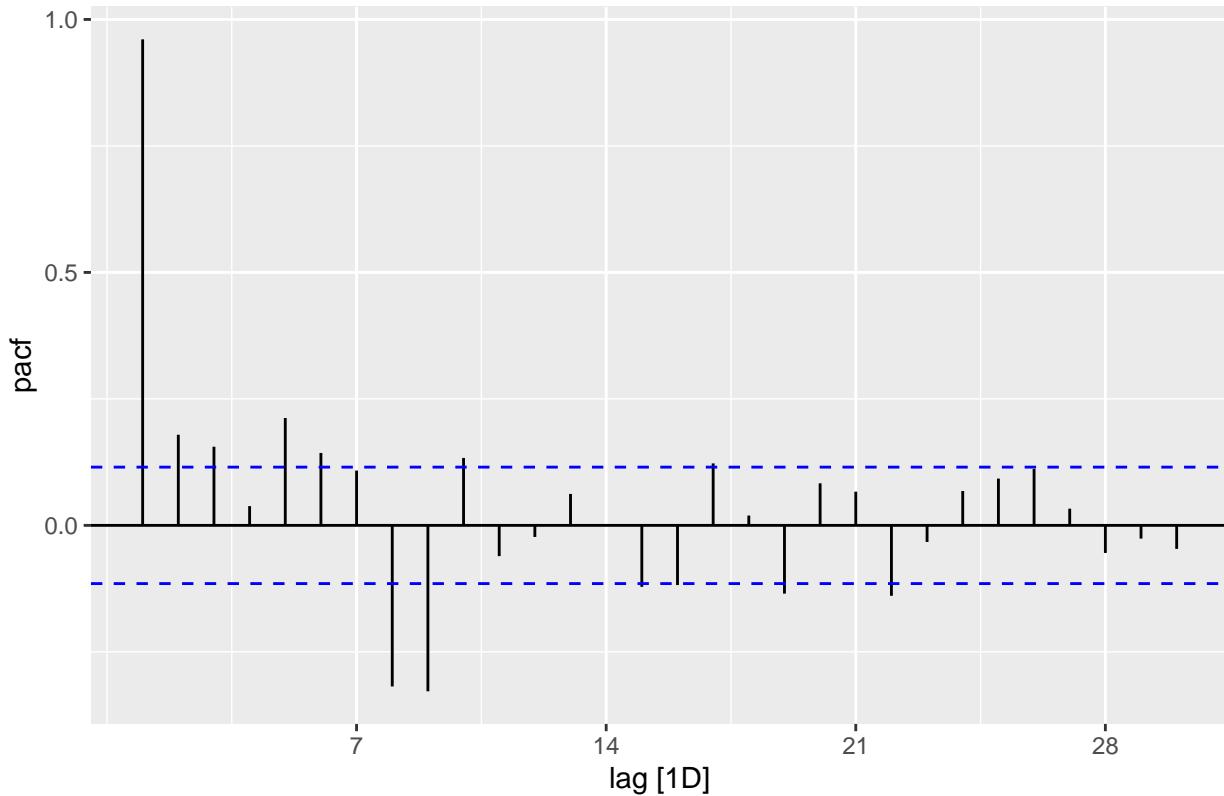
```
Total_ts %>%
  filter(sub_region_2 == "Málaga") %>%
  PACF(num_casos.x, lag_max = 30) %>%
  autoplot() +
  labs(title=" Málaga – ACF N° Cases")
```

Málaga – ACF Nº Cases

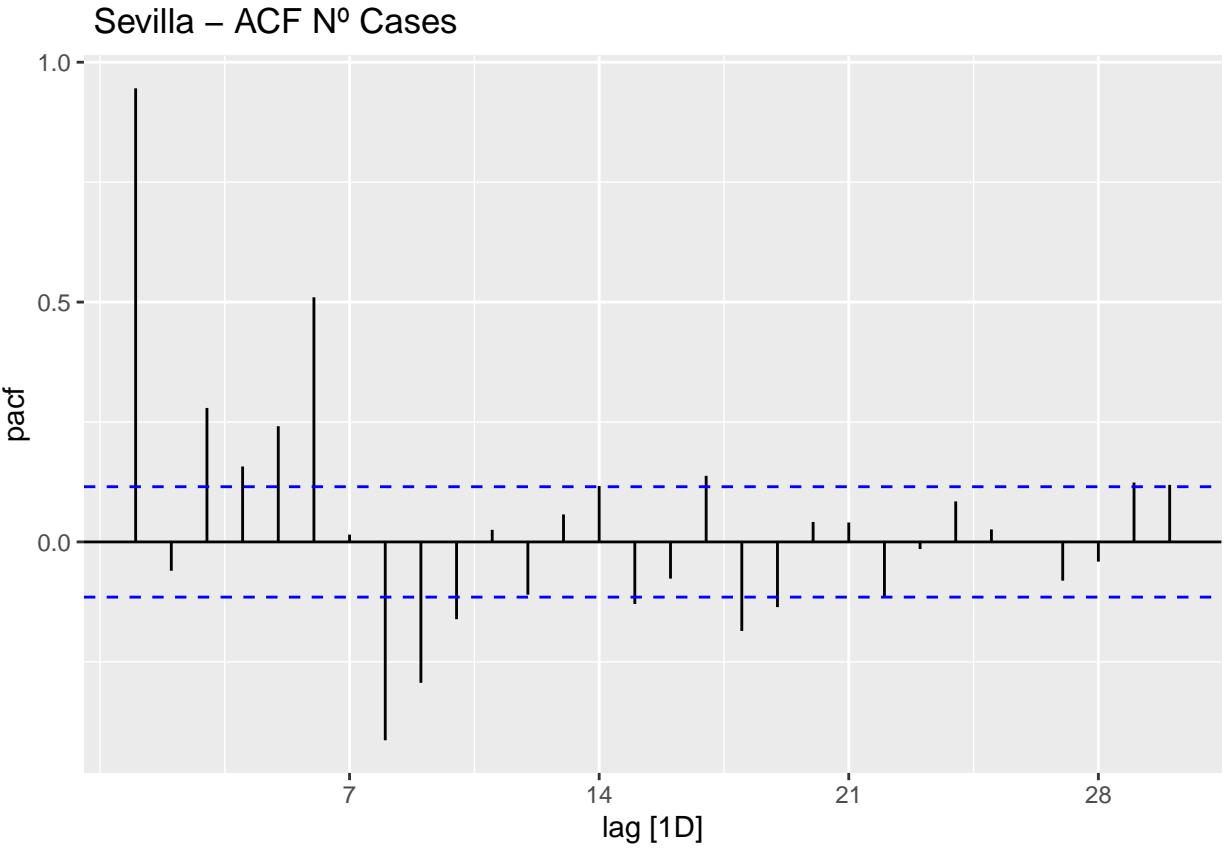


```
Total_ts %>%
  filter(sub_region_2 == "Cádiz") %>%
  PACF(num_casos.x, lag_max = 30) %>%
  autoplot() +
  labs(title=" Cádiz – ACF Nº Cases")
```

Cádiz – ACF N° Cases



```
Total_ts %>%
  filter(sub_region_2 == "Sevilla") %>%
  PACF(num_casos.x, lag_max = 30) %>%
  autoplot() +
  labs(title=" Sevilla – ACF N° Cases")
```



3.2 STL (Seasonal and Trend decomposition using Loess - Barcelona, Madrid, Málaga, Cádiz and Sevilla)

As stated by (Hyndman and Athanasopoulos 2021)... "STL has several advantages over classical decomposition, and the SEATS and X-11 methods:

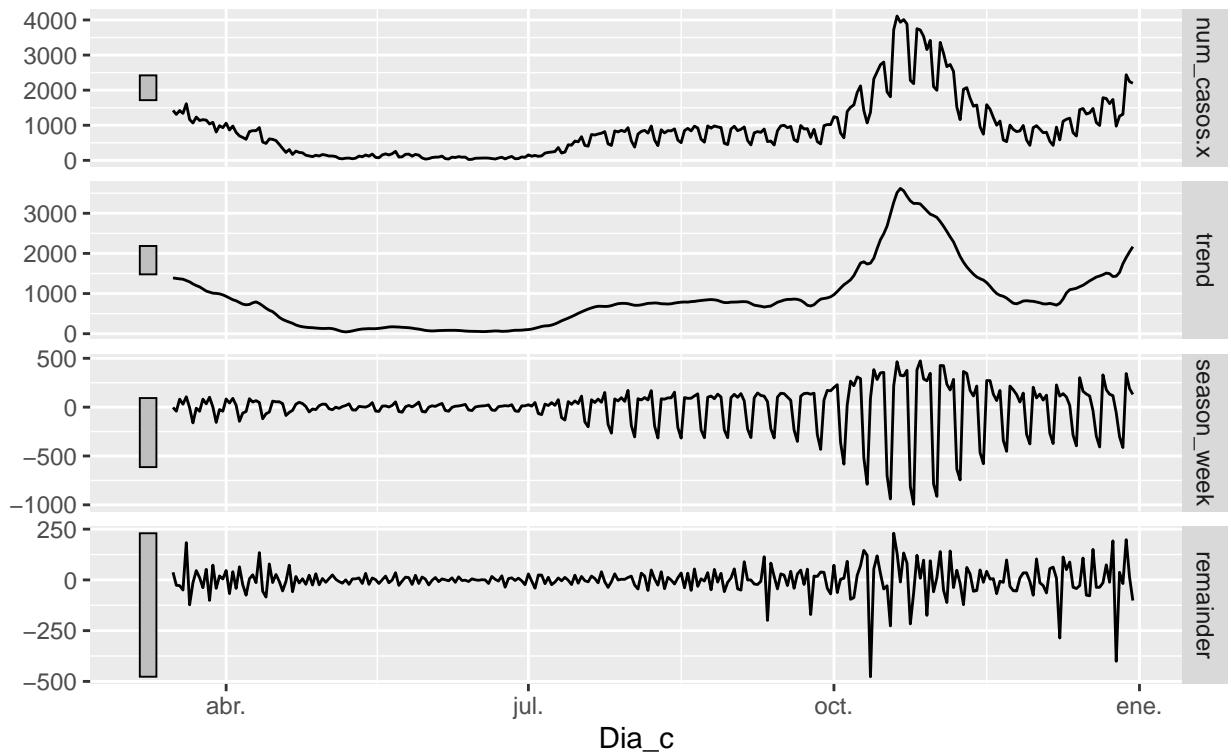
- Unlike SEATS and X-11, STL will handle any type of seasonality, not only monthly and quarterly data.
- The seasonal component is allowed to change over time, and the rate of change can be controlled by the user.
- The smoothness of the trend-cycle can also be controlled by the user.
- It can be robust to outliers (i.e., the user can specify a robust decomposition), so that occasional unusual observations will not affect the estimates of the trend-cycle and seasonal components. They will, however, affect the remainder component"...

Note: Due to some issues detected with the original data (extreme variance in num_casos.x variable) we have to check the different features observed, using the STL “seasonal adjustment” / “trend” as value for the serie. Data has 365 frequency but STL detects weekly seasonality.

```
# Check Seasonal and trend plus transformed one
Total_ts %>%
  #filter_index("2020-09-1" ~ "2020-12-31") %>%
  filter(sub_region_2 == "Barcelona") %>%
  model(STL(num_casos.x ~ season(window = 7) +
            trend(window = 7))) %>%
  components() %>%
  autoplot() + labs(title="Barcelona")
```

Barcelona

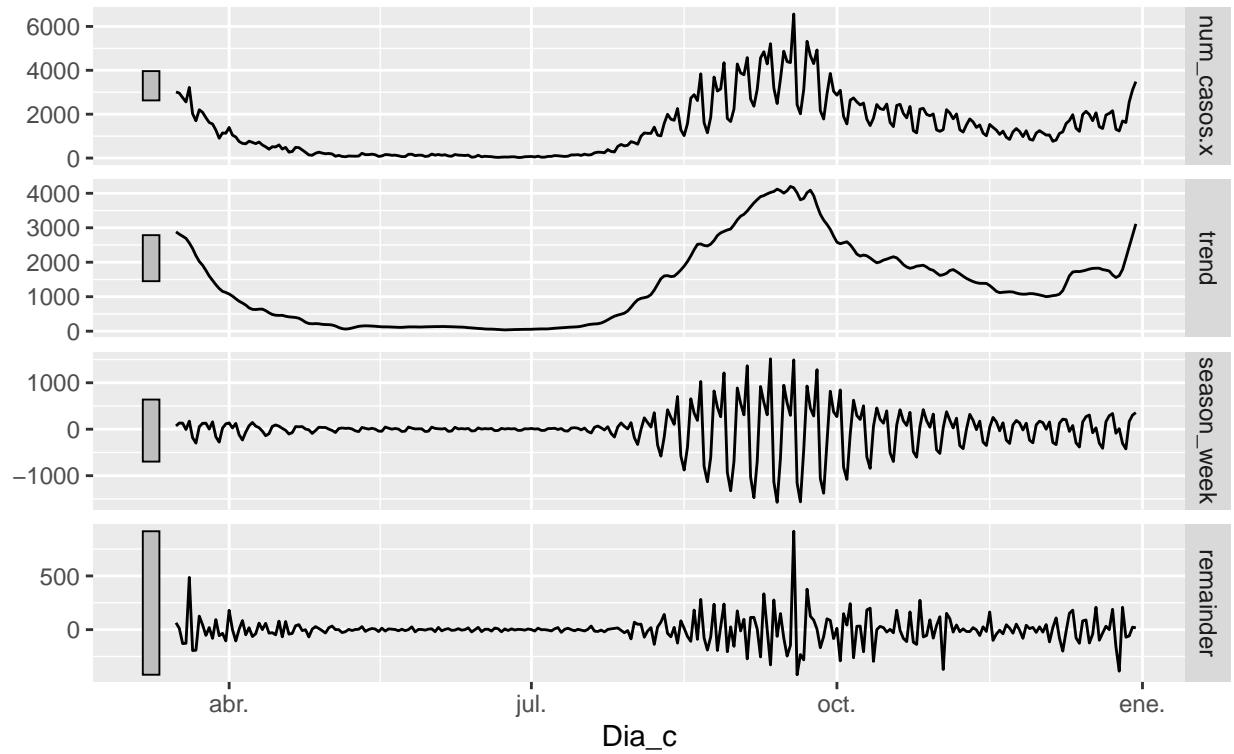
num_casos.x = trend + season_week + remainder



```
Total_ts %>%
  filter(sub_region_2 == "Madrid") %>%
  model(STL(num_casos.x ~ season(window = 7) +
            trend(window = 7))) %>%
  components() %>%
  autoplot() + labs(title="Madrid")
```

Madrid

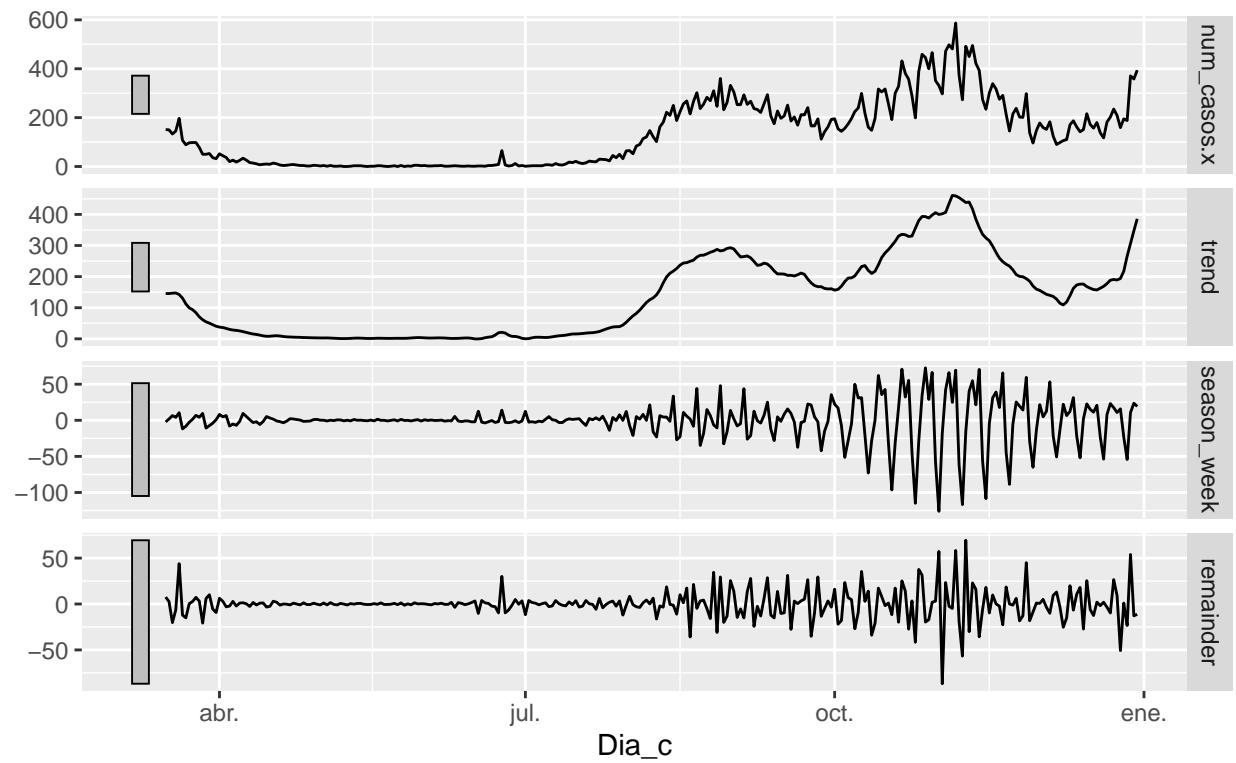
num_casos.x = trend + season_week + remainder



```
Total_ts %>%
  filter(sub_region_2 == "Málaga") %>%
  model(STL(num_casos.x ~ season(window = 7) +
            trend(window = 7))) %>%
  components() %>%
  autoplot() + labs(title="Málaga")
```

Málaga

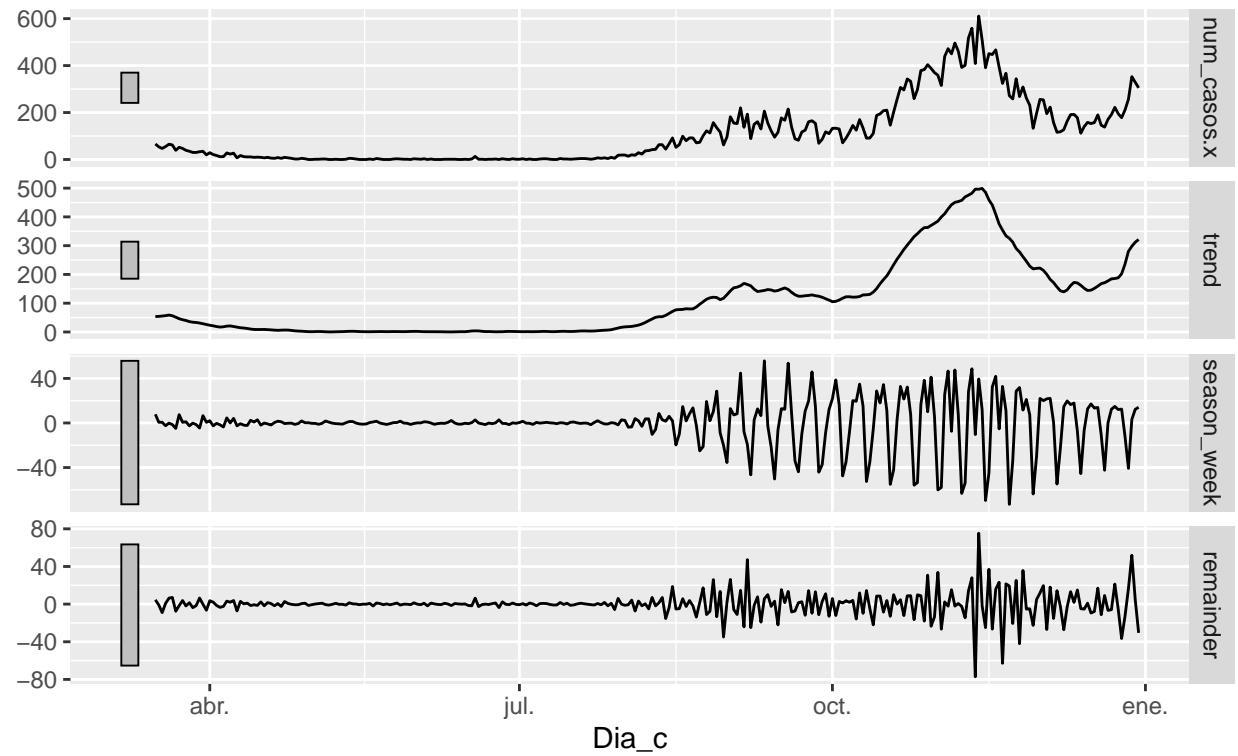
num_casos.x = trend + season_week + remainder



```
Total_ts %>%
  filter(sub_region_2 == "Cádiz") %>%
  model(STL(num_casos.x ~ season(window = 7) +
            trend(window = 7))) %>%
  components() %>%
  autoplot() + labs(title="Cádiz")
```

Cádiz

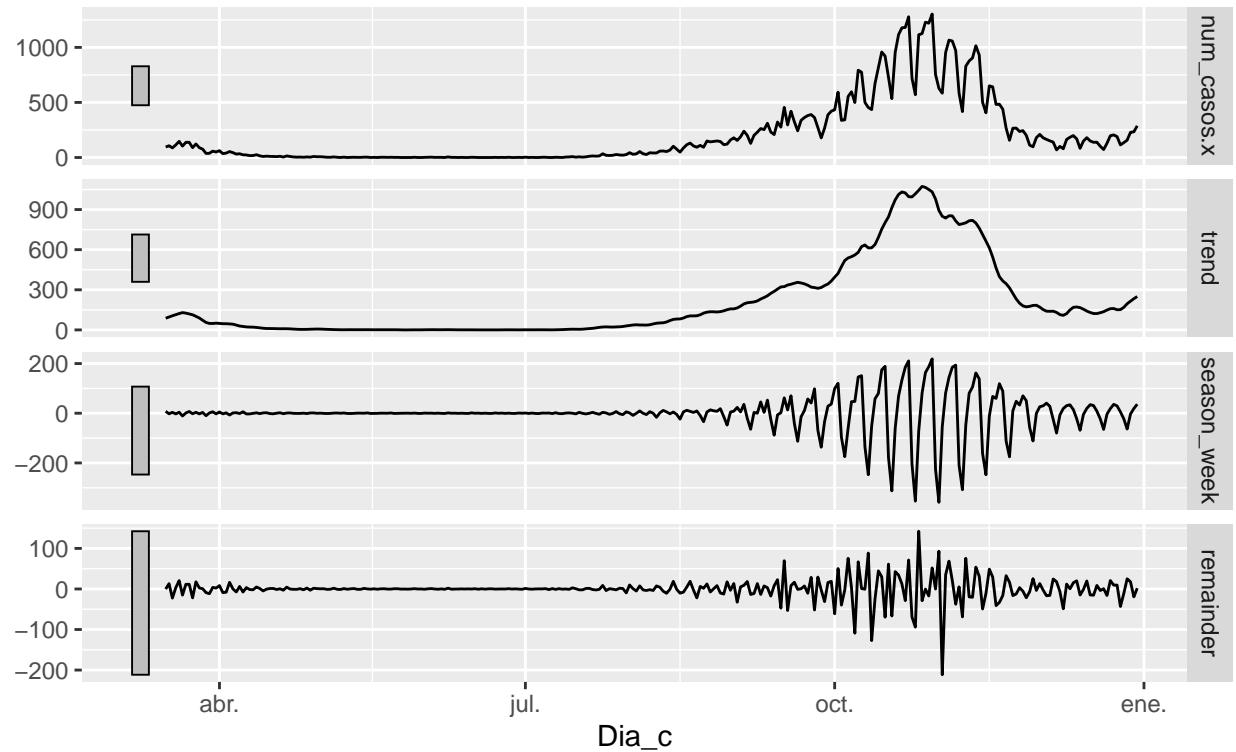
num_casos.x = trend + season_week + remainder



```
Total_ts %>%
  #filter_index("2020-09-1" ~ "2020-12-31")  %>%
  filter(sub_region_2 == "Sevilla") %>%
  model(STL(num_casos.x ~ season(window = 7) +
            trend(window = 7))) %>%
  components() %>%
  autoplot() + labs(title="Sevilla")
```

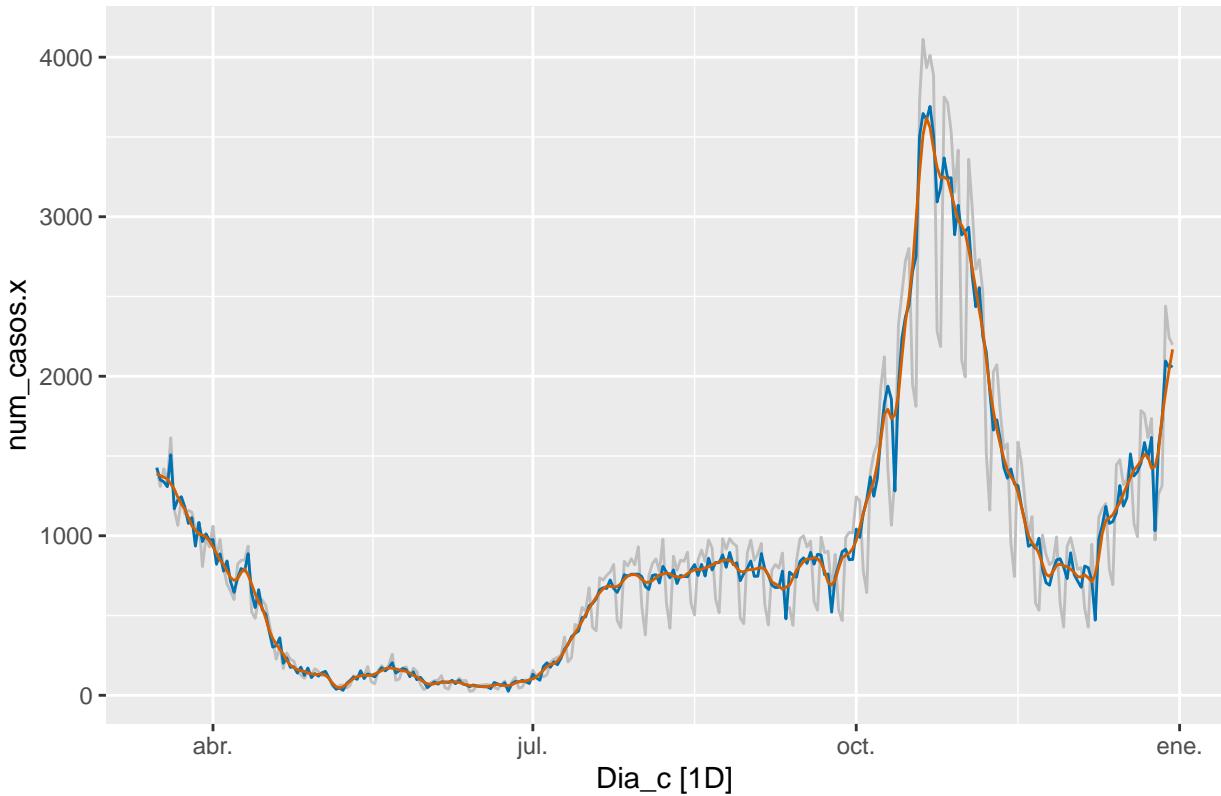
Sevilla

```
num_casos.x = trend + season_week + remainder
```



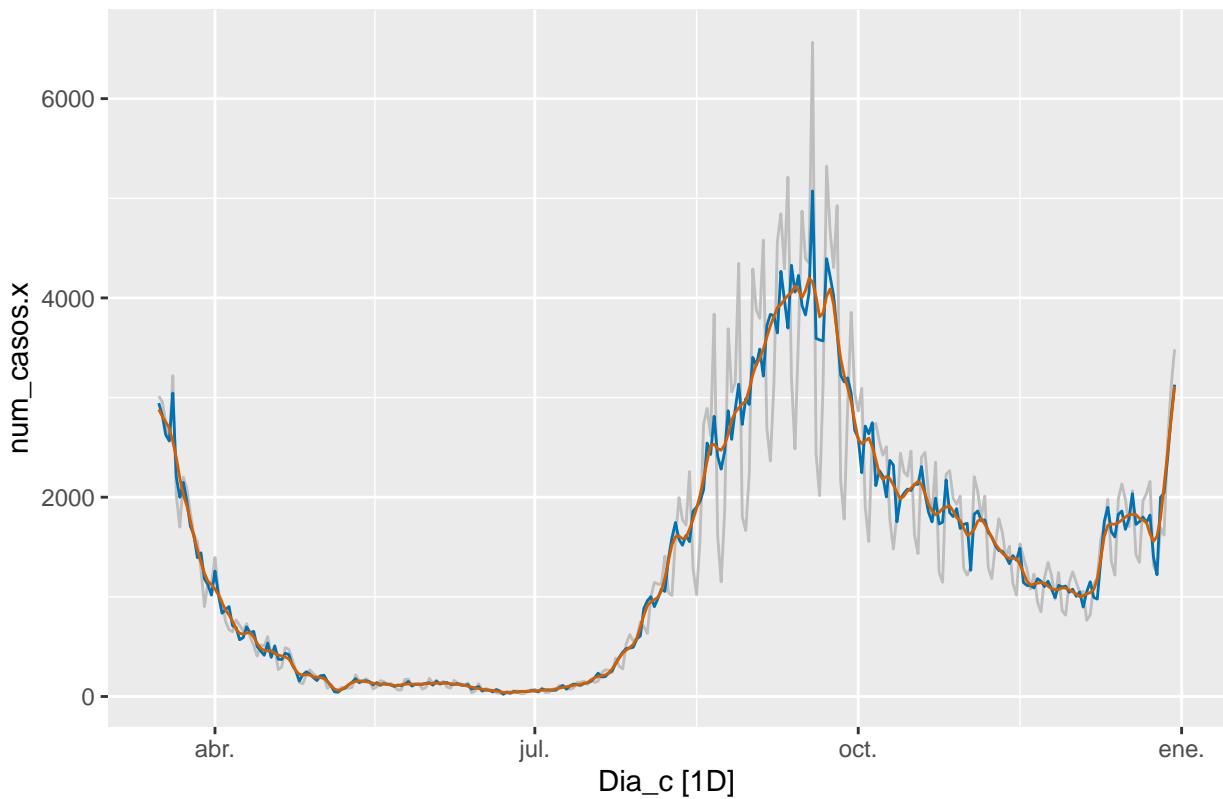
```
# Plot season adjusted and trend
#####
Total_ts %>%
  filter(sub_region_2 == "Barcelona") %>%
  model(STL(num_casos.x ~ season(window = 7) +
            trend(window = 7))) %>%
  components() %>% as_tsibble() %>%
  autoplot(num_casos.x, color = "gray") +
  geom_line(aes(y=season_adjust), color = "#0072B2") +
  geom_line(aes(y=trend), color = "#D55E00") +
  labs(title="Barcelona")
```

Barcelona



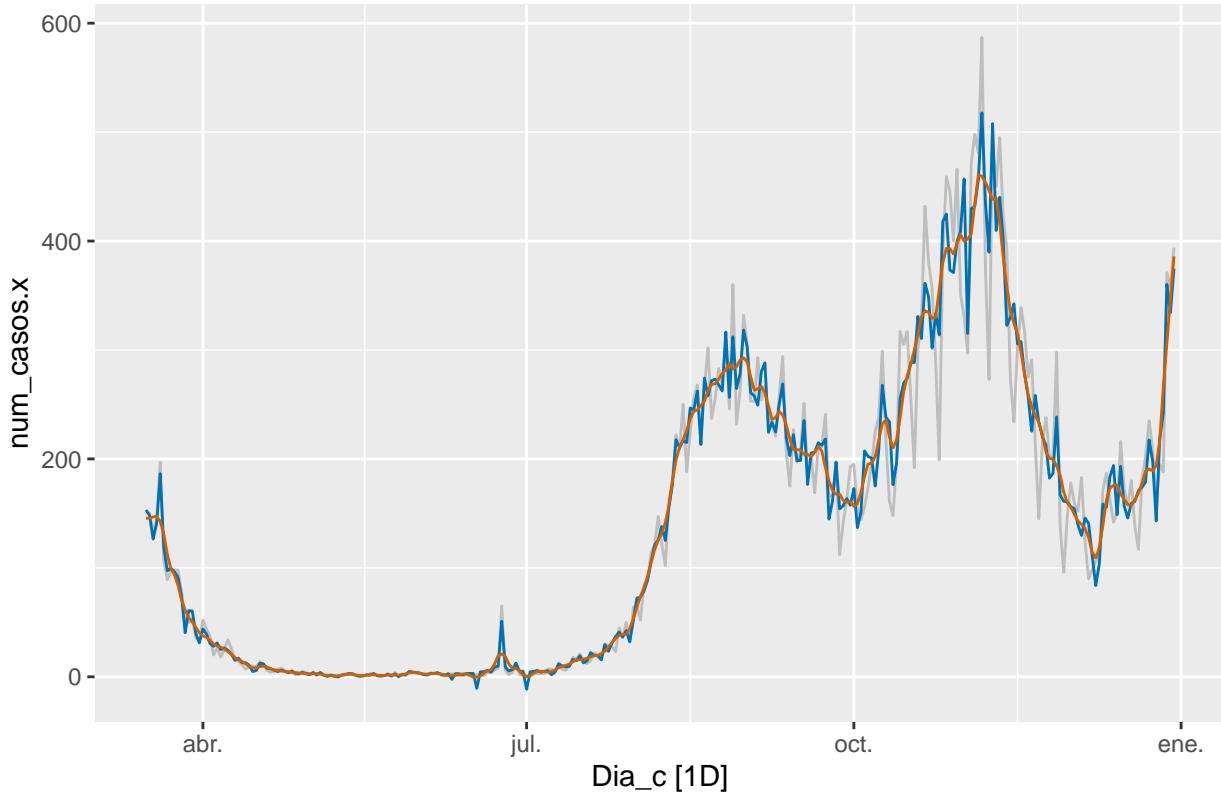
```
Total_ts %>%
  #filter_index("2020-09-1" ~ "2020-12-31") %>%
  filter(sub_region_2 == "Madrid") %>%
  model(STL(num_casos.x ~ season(window = 7) +
            trend(window = 7))) %>%
  components() %>% as_tsibble() %>%
  autoplot(num_casos.x, color = "gray") +
  geom_line(aes(y=season_adjust), color = "#0072B2") +
  geom_line(aes(y=trend), color = "#D55E00") +
  labs(title="Madrid")
```

Madrid



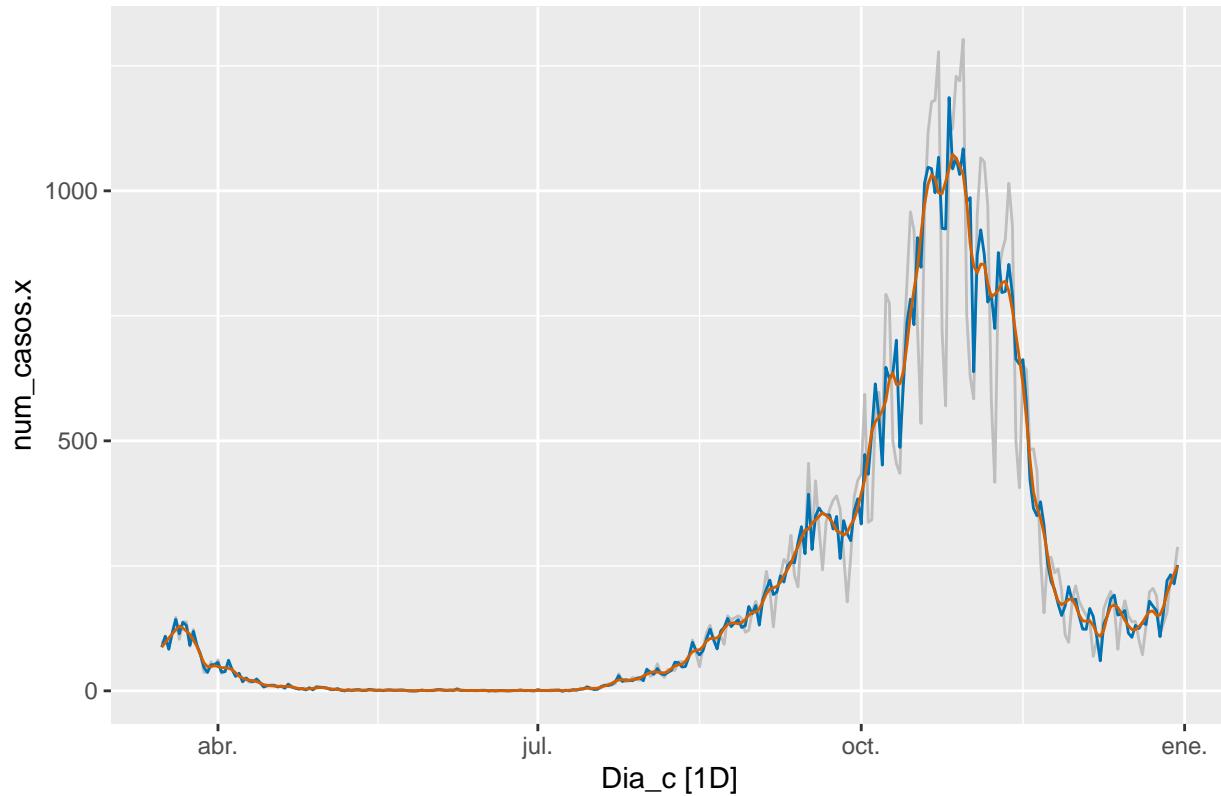
```
Total_ts %>%
  filter(sub_region_2 == "Málaga") %>%
  model(STL(num_casos.x ~ season(window = 7) +
            trend(window = 7))) %>%
  components() %>% as_tsibble() %>%
  autoplot(num_casos.x, color = "gray") +
  geom_line(aes(y=season_adjust), color = "#0072B2") +
  geom_line(aes(y=trend), color = "#D55E00") +
  labs(title="Málaga")
```

Málaga



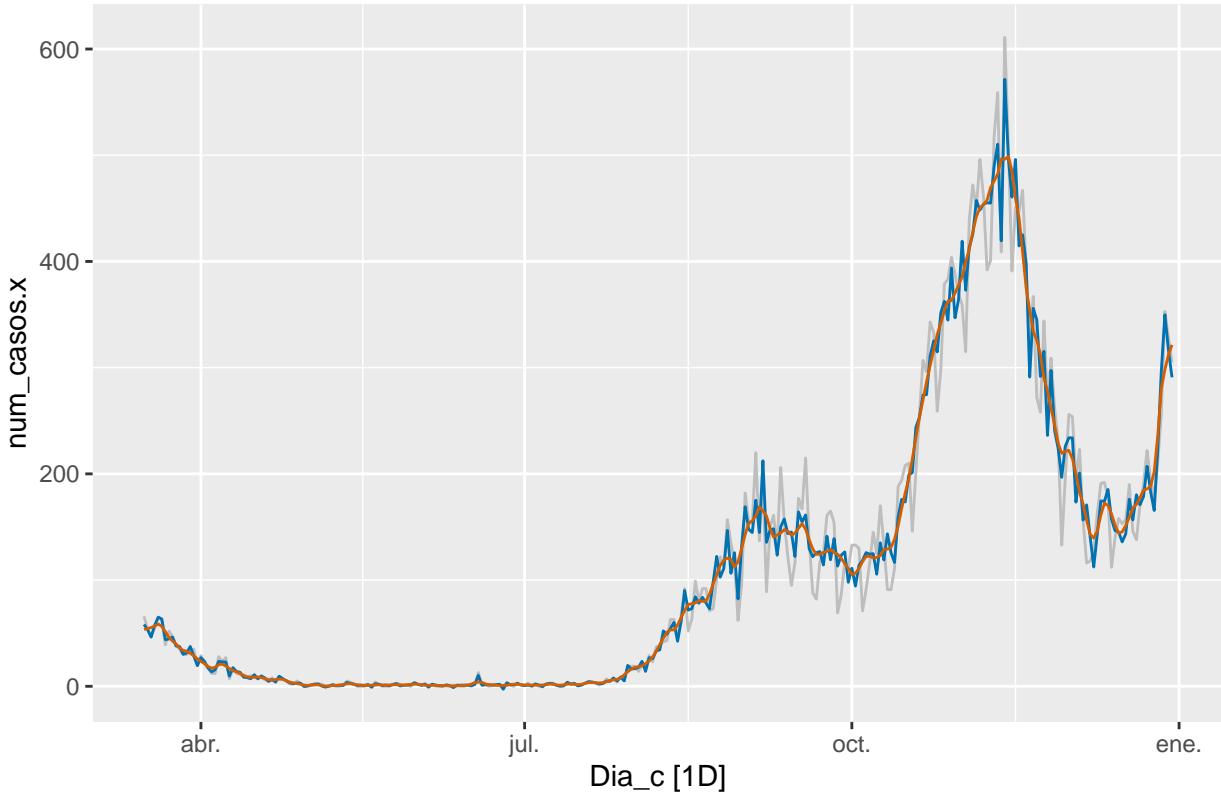
```
Total_ts %>%
  filter(sub_region_2 == "Sevilla") %>%
  model(STL(num_casos.x ~ season(window = 7) +
            trend(window = 7))) %>%
  components() %>% as_tsibble() %>%
  autoplot(num_casos.x, color = "gray") +
  geom_line(aes(y=season_adjust), color = "#0072B2") +
  geom_line(aes(y=trend), color = "#D55E00") +
  labs(title="Sevilla")
```

Sevilla



```
Total_ts %>%
  filter(sub_region_2 == "Cádiz") %>%
  model(STL(num_casos.x ~ season(window = 7) +
            trend(window = 7))) %>%
  components() %>% as_tsibble() %>%
  autoplot(num_casos.x, color = "gray") +
  geom_line(aes(y=season_adjust), color = "#0072B2") +
  geom_line(aes(y=trend), color = "#D55E00") +
  labs(title="Cádiz")
```

Cádiz



We create a separate time series by province our study (Barcelona, Madrid, Málaga, Sevilla and Cádiz).

```
# New time-series for Bar, Mad, Mal, Cor and, Cad
Total_ts %>%
  filter(sub_region_2 == "Barcelona") -> Bar_N_cases
Total_ts %>%
  filter(sub_region_2 == "Madrid") -> Mad_N_cases
Total_ts %>%
  filter(sub_region_2 == "Málaga") -> Mal_N_cases
Total_ts %>%
  filter(sub_region_2 == "Cádiz") -> Cad_N_cases
Total_ts %>%
  filter(sub_region_2 == "Sevilla") -> Sev_N_cases

#filter_index("2020-03-15" ~ "2020-12-31") %>%
```

Note: The fable package will automatically back-transform the forecasts whenever a transformation has been used in the model definition (Hyndman and Athanasopoulos 2021).

Note: Box-Cox transformation (lambda is obtained using “guerrero” feature -fpp3-) and double difference is plotted (lag=7) ((Hyndman and Athanasopoulos 2021) Sec 3 - Transformations).

We proceed with the analysis of the stationary.

```
# Lambda values
lambda_bar <- Bar_N_cases %>%
  features(num_casos.x, features = guerrero) %>%
  pull(lambda_guerrero)
```

```

lambda_mad <- Mad_N_cases %>%
  features(num_casos.x, features = guerrero) %>%
  pull(lambda_guerrero)

lambda_mal <- Mal_N_cases %>%
  features(num_casos.x, features = guerrero) %>%
  pull(lambda_guerrero)

lambda_cad <- Cad_N_cases %>%
  features(num_casos.x, features = guerrero) %>%
  pull(lambda_guerrero)

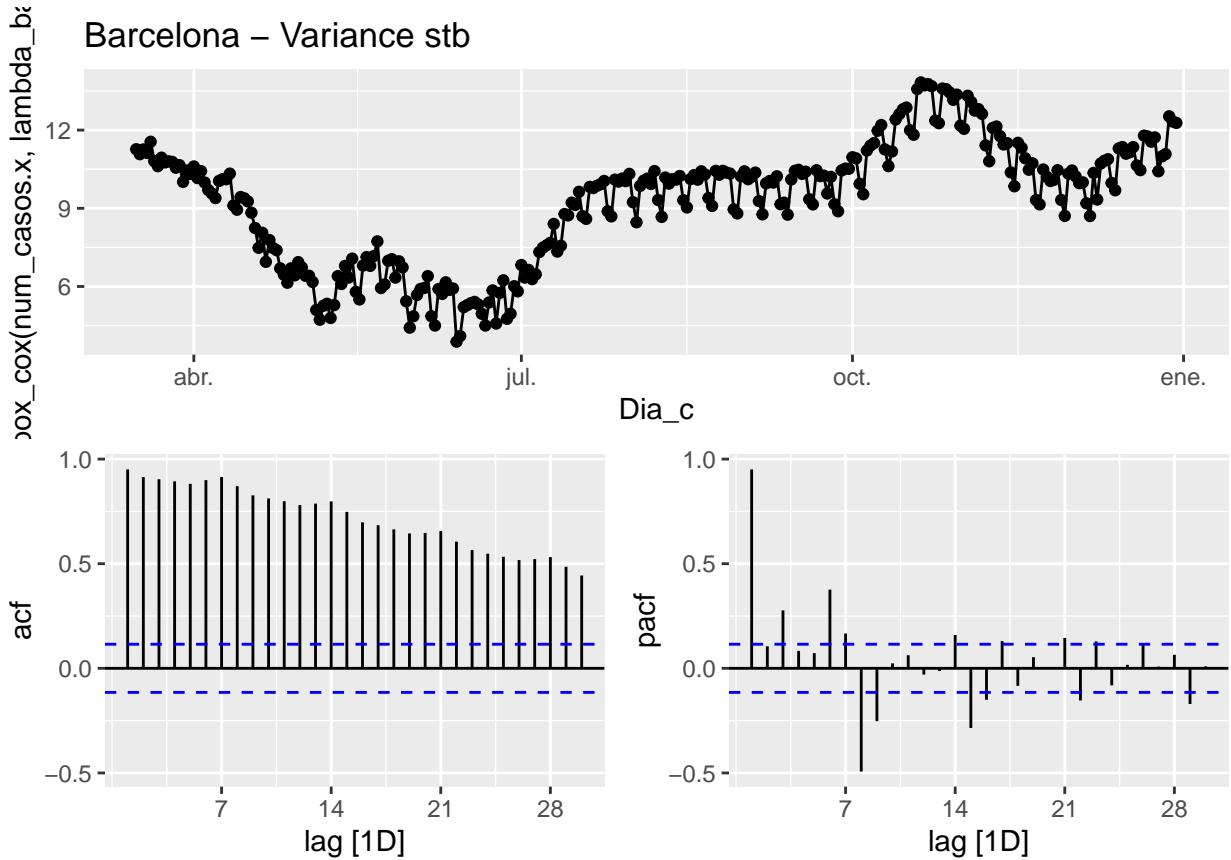
lambda_sev <- Sev_N_cases %>%
  features(num_casos.x, features = guerrero) %>%
  pull(lambda_guerrero)

# Variance stb checks (variance + difference)
#####
Bar_N_cases %>%
  gg_tsdisplay(((box_cox(num_casos.x,lambda_bar))),  

               plot_type='partial', lag=30)+  

  labs(title="Barcelona – Variance stb")

```



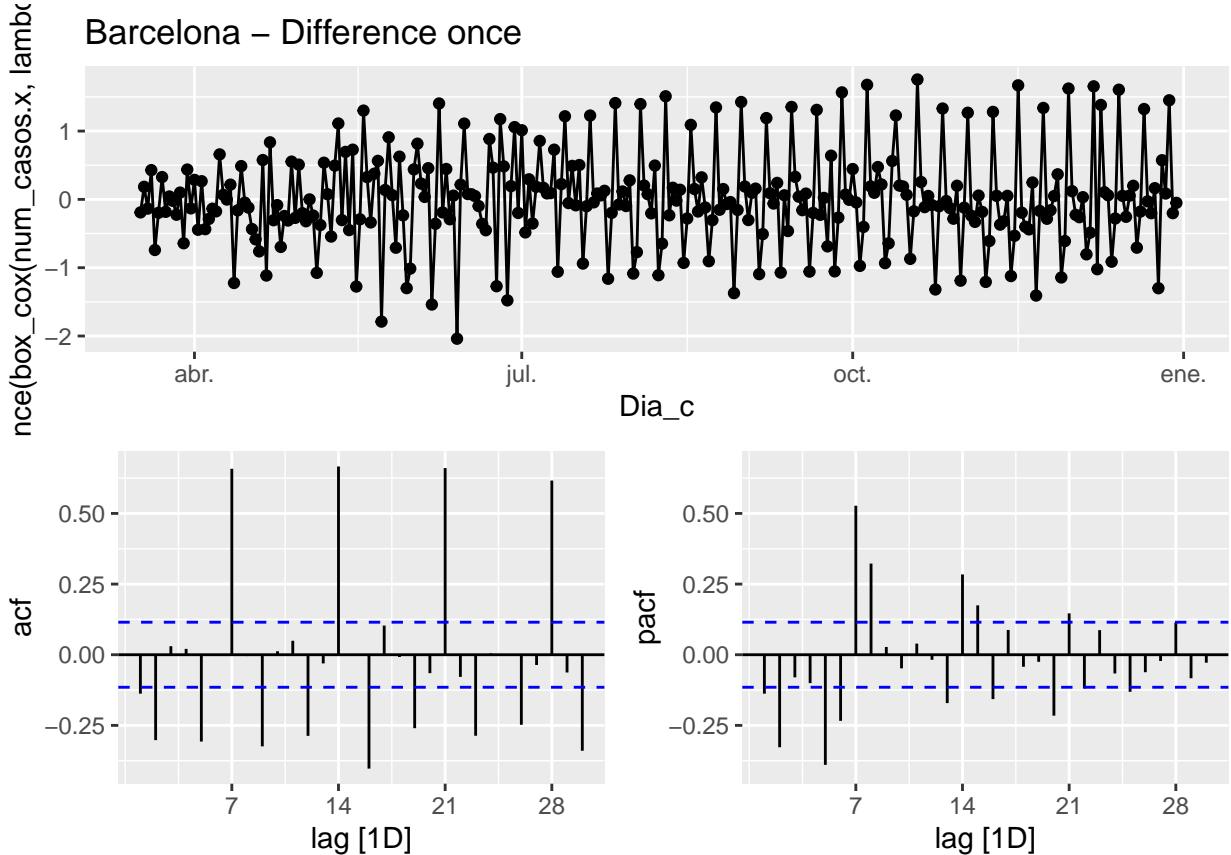
```

Bar_N_cases %>%
  gg_tsdisplay((difference(box_cox(num_casos.x,lambda_bar))),  

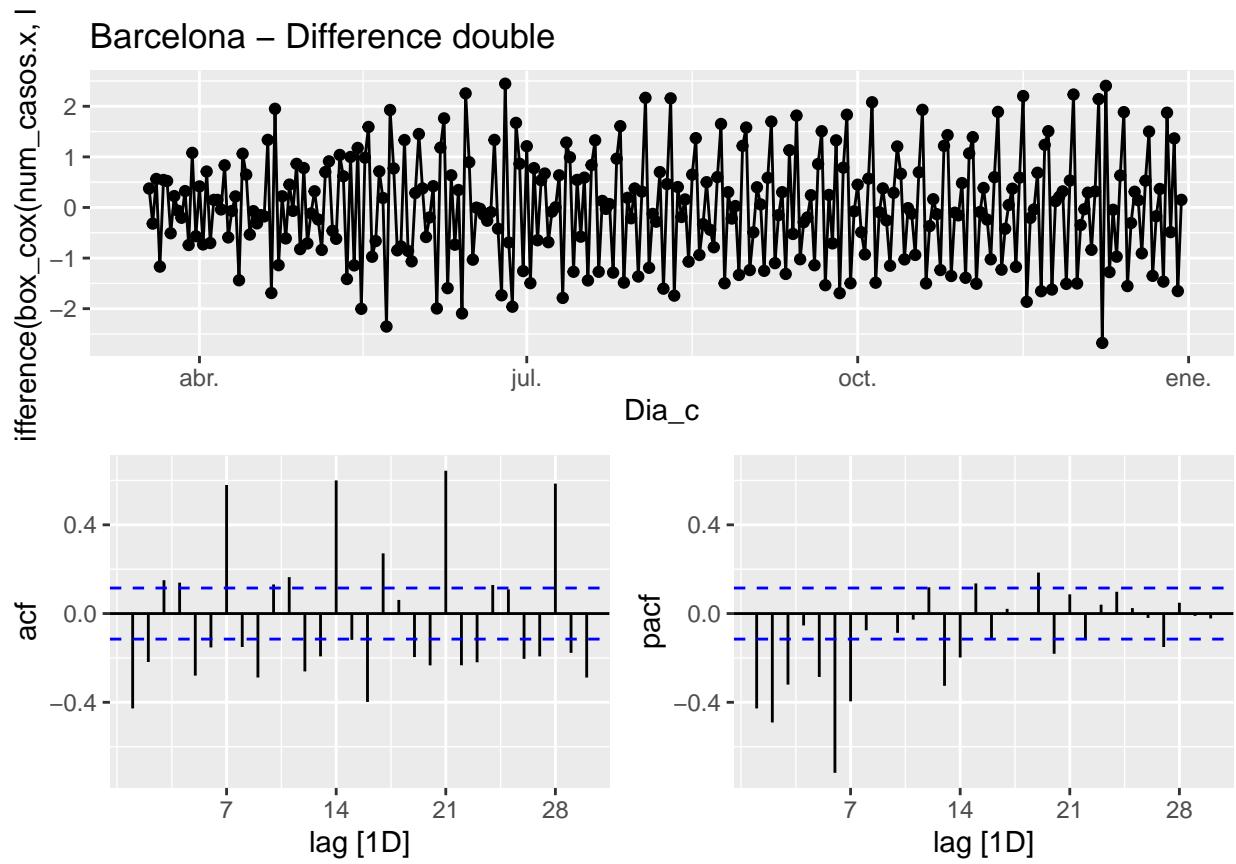
               plot_type='partial', lag=30)+

```

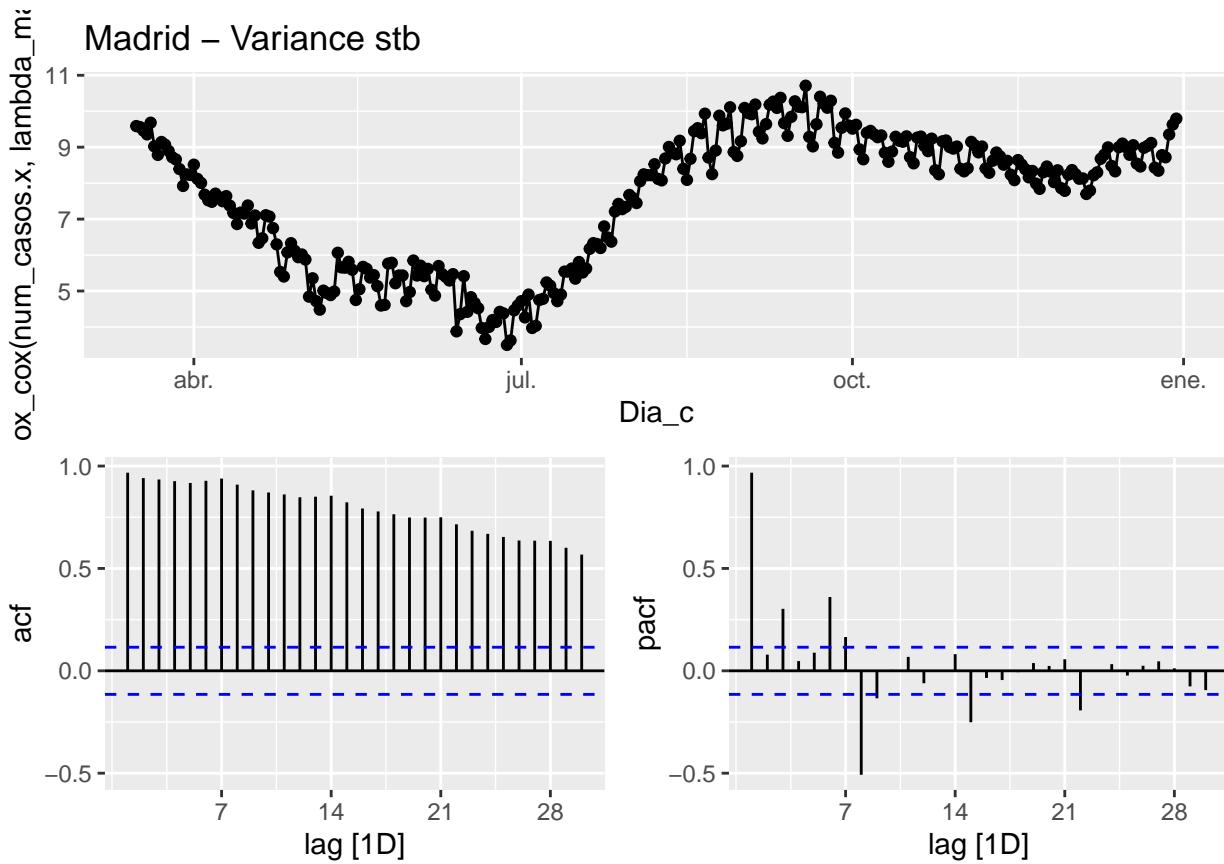
```
labs(title="Barcelona - Difference once")
```



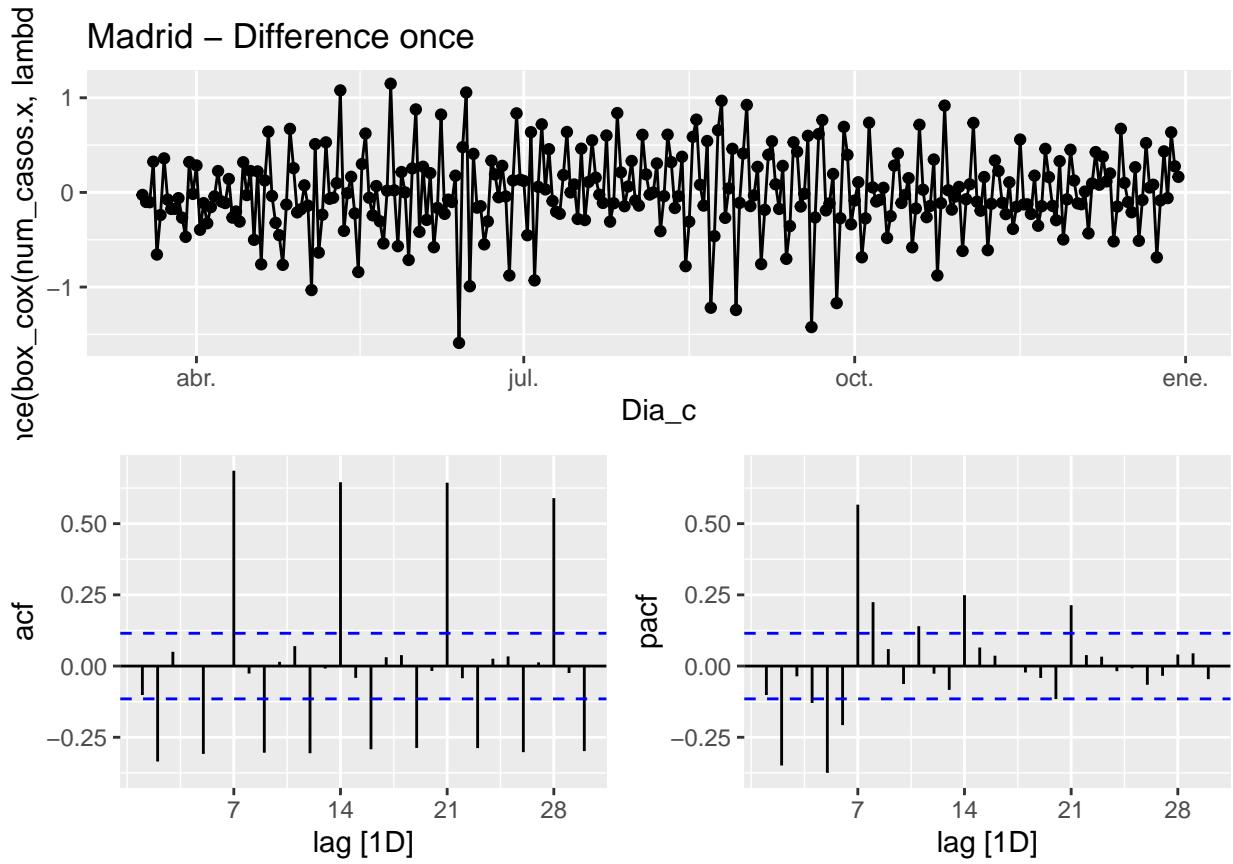
```
Bar_N_cases %>%
  gg_tsdisplay(difference(difference(box_cox(num_casos.x, lambda_bar))),
               plot_type='partial', lag=30) +
  labs(title="Barcelona - Difference double")
```



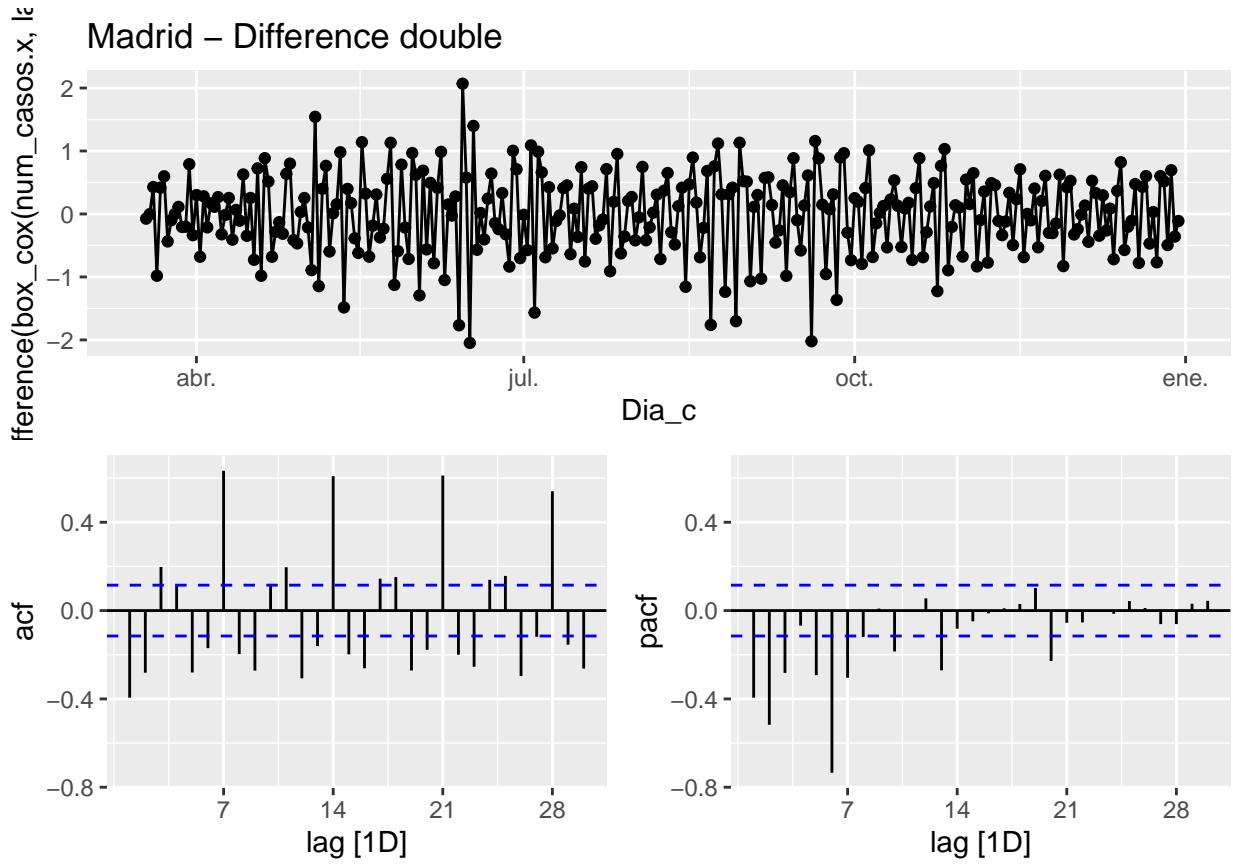
```
#####
#Mad_N_cases %>%
gg_tsdisplay((box_cox(num_casos.x,lambda_mad))),
  plot_type='partial', lag=30) +
  labs(title="Madrid - Variance stb")
```



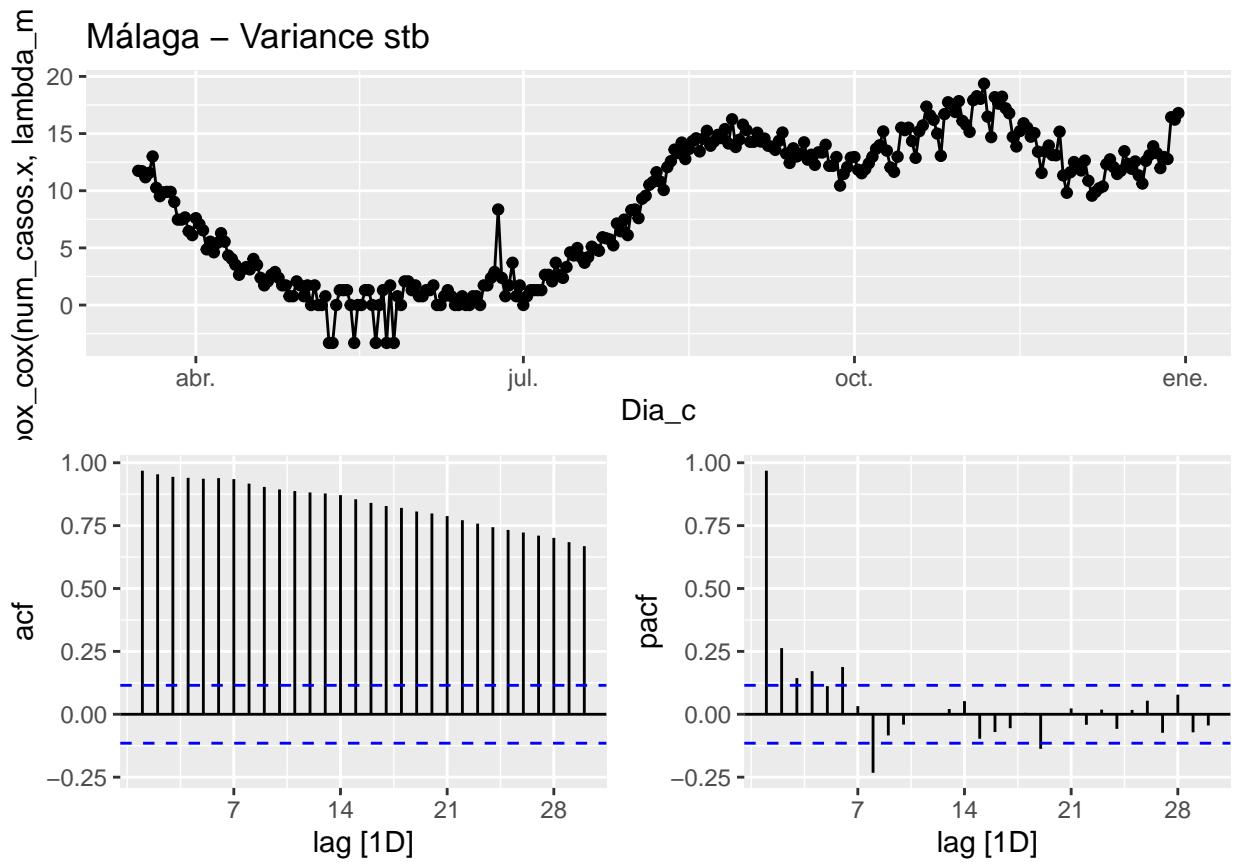
```
Mad_N_cases %>%
  gg_tsdisplay((difference(box_cox(num_casos.x, lambda_mad))),
               plot_type='partial', lag=30) +
  labs(title="Madrid – Difference once")
```



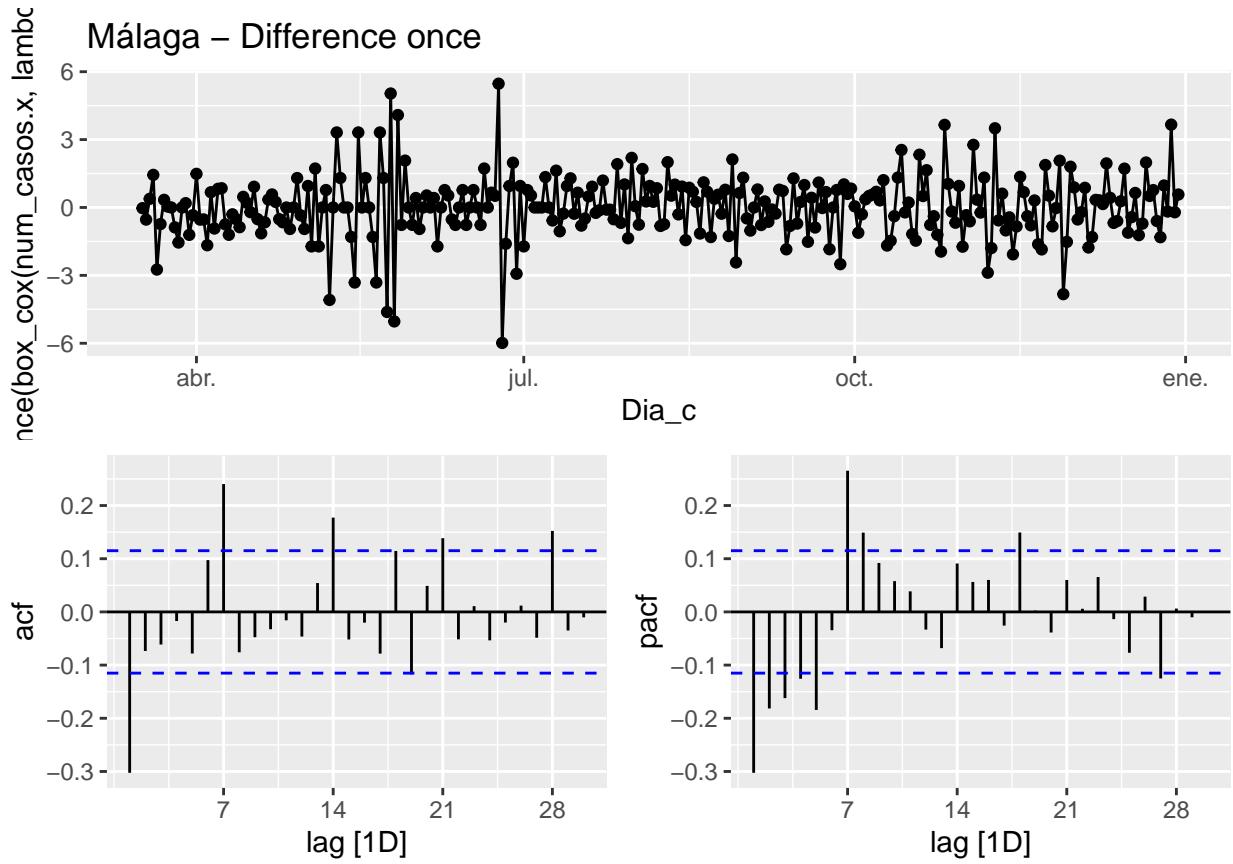
```
Mad_N_cases %>%
  gg_tsdisplay(difference(difference(box_cox(num_casos.x, lambda_mad))),
               plot_type='partial', lag=30) +
  labs(title="Madrid - Difference double")
```



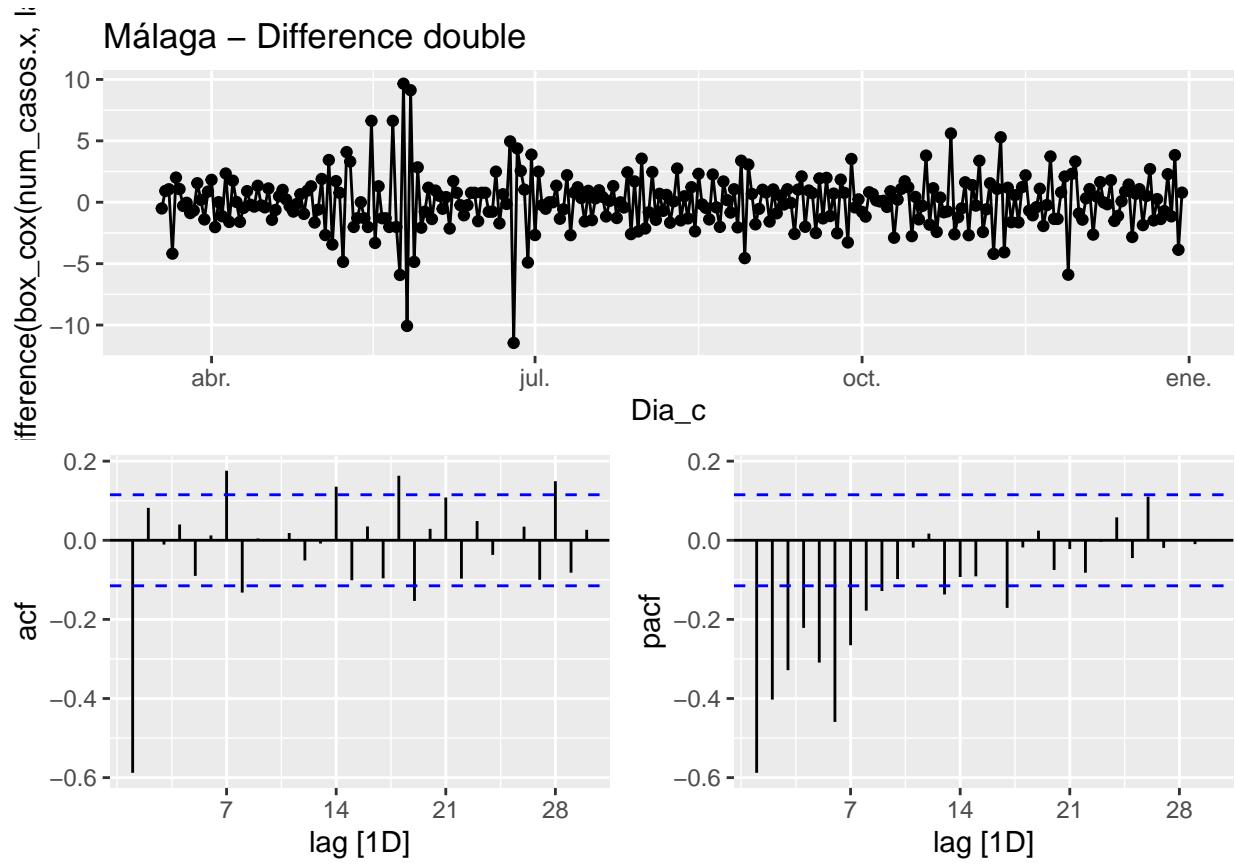
```
#####
Mal_N_cases %>%
  gg_tsdisplay((box_cox(num_casos.x,lambda_mal))),
  plot_type='partial', lag=30) +
  labs(title="Málaga - Variance stb")
```



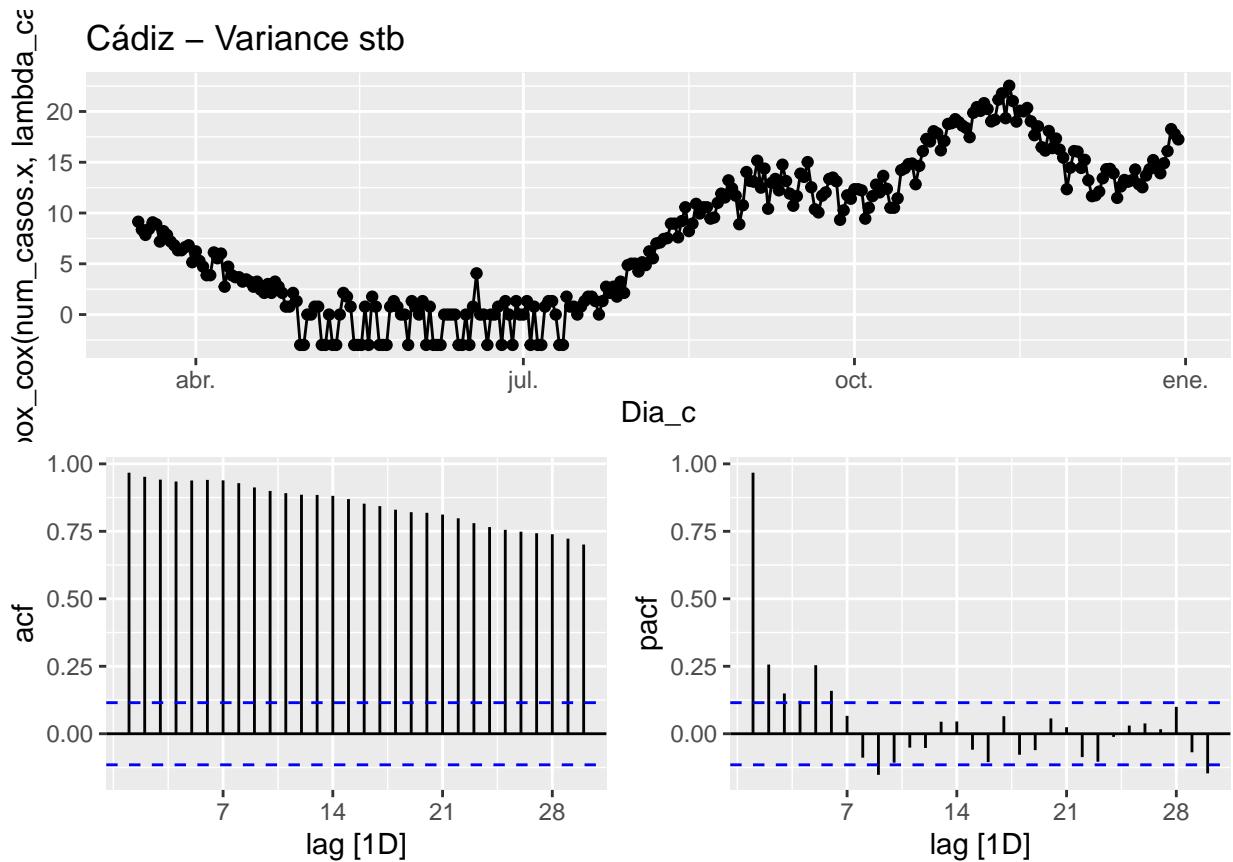
```
Mal_N_cases %>%
  gg_tsdisplay((difference(box_cox(num_casos.x, lambda_mal))),
               plot_type='partial', lag=30)+
  labs(title="Málaga - Difference once")
```



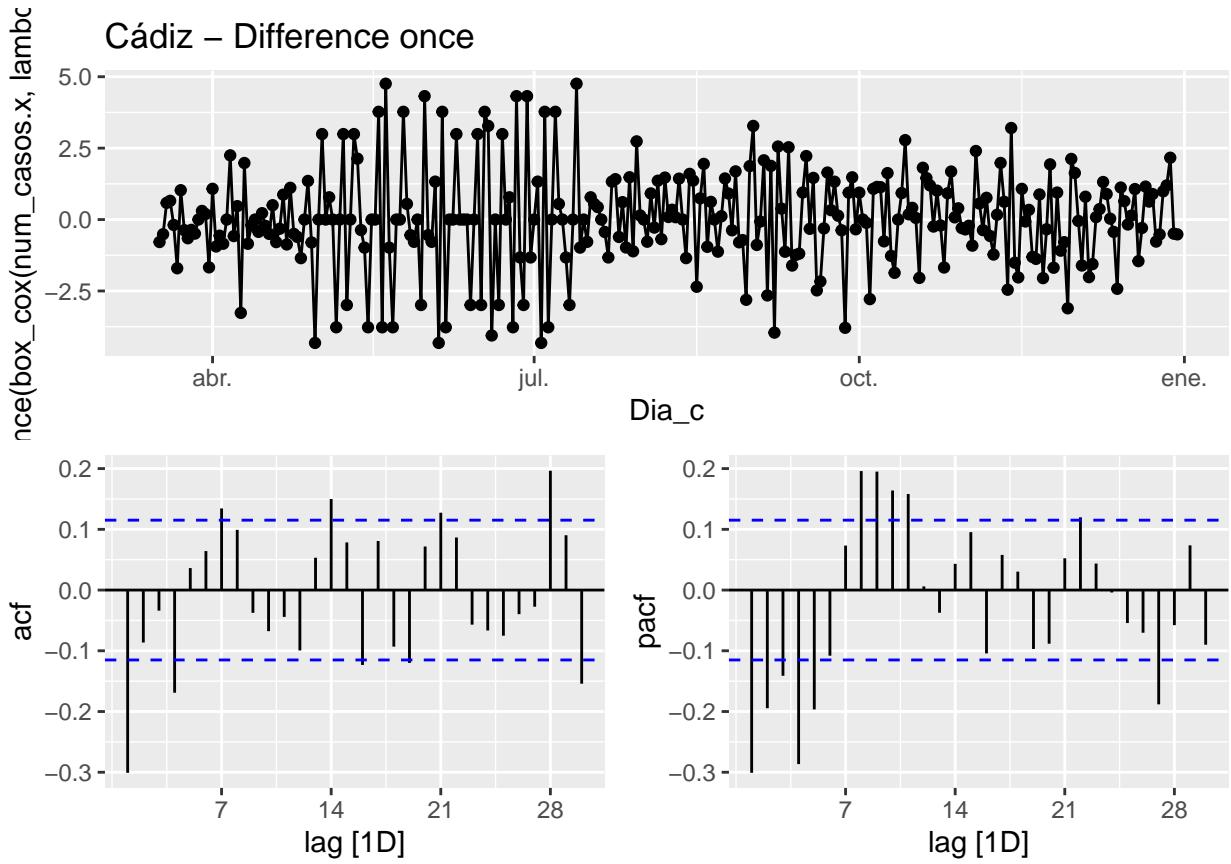
```
Mal_N_cases %>%
  gg_tsdisplay(difference(difference(box_cox(num_casos.x, lambda_mal))),
               plot_type='partial', lag=30) +
  labs(title="Málaga - Difference double")
```



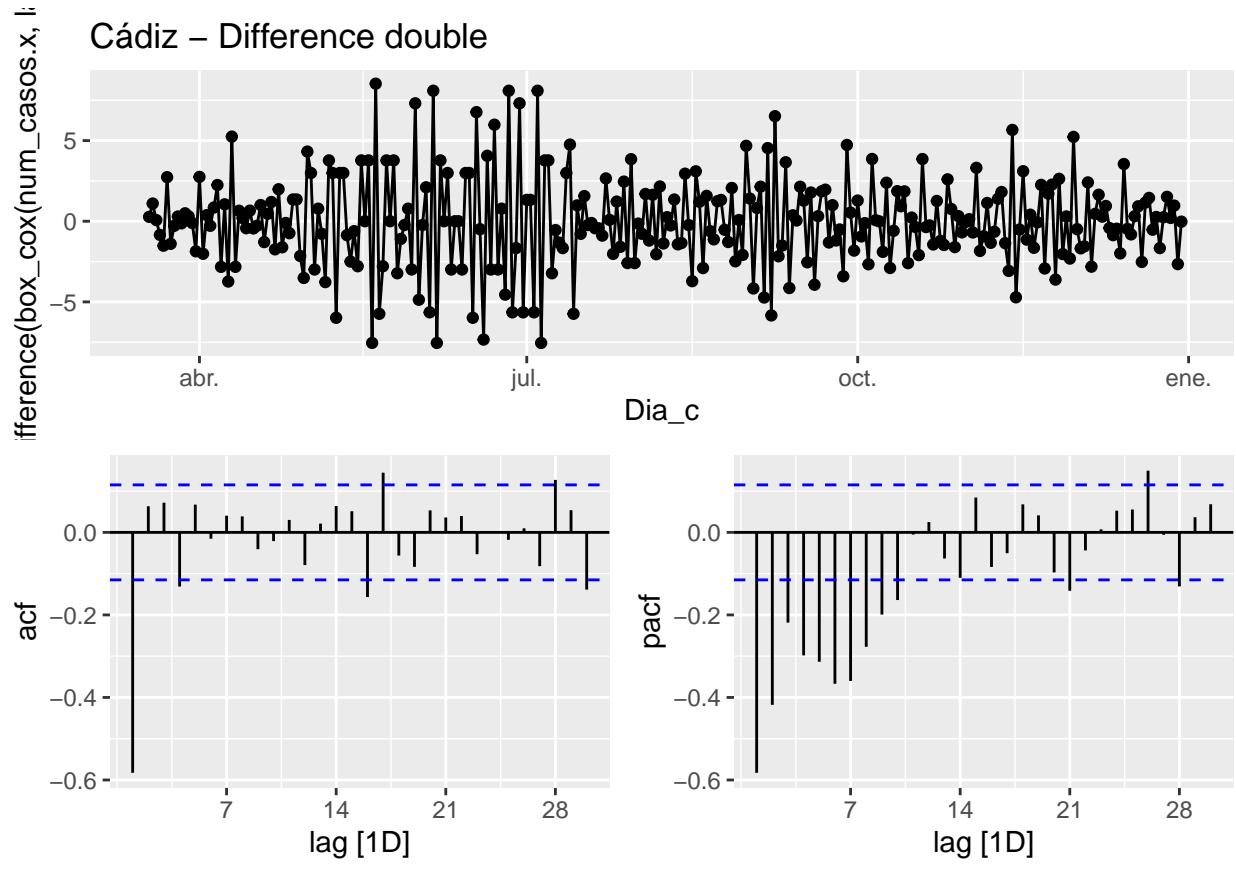
```
#####
# Cad_N_cases %>%
gg_tsdisplay((box_cox(num_casos.x,lambda_cad))),
  plot_type='partial',lag=30) +
  labs(title="Cádiz - Variance stb")
```



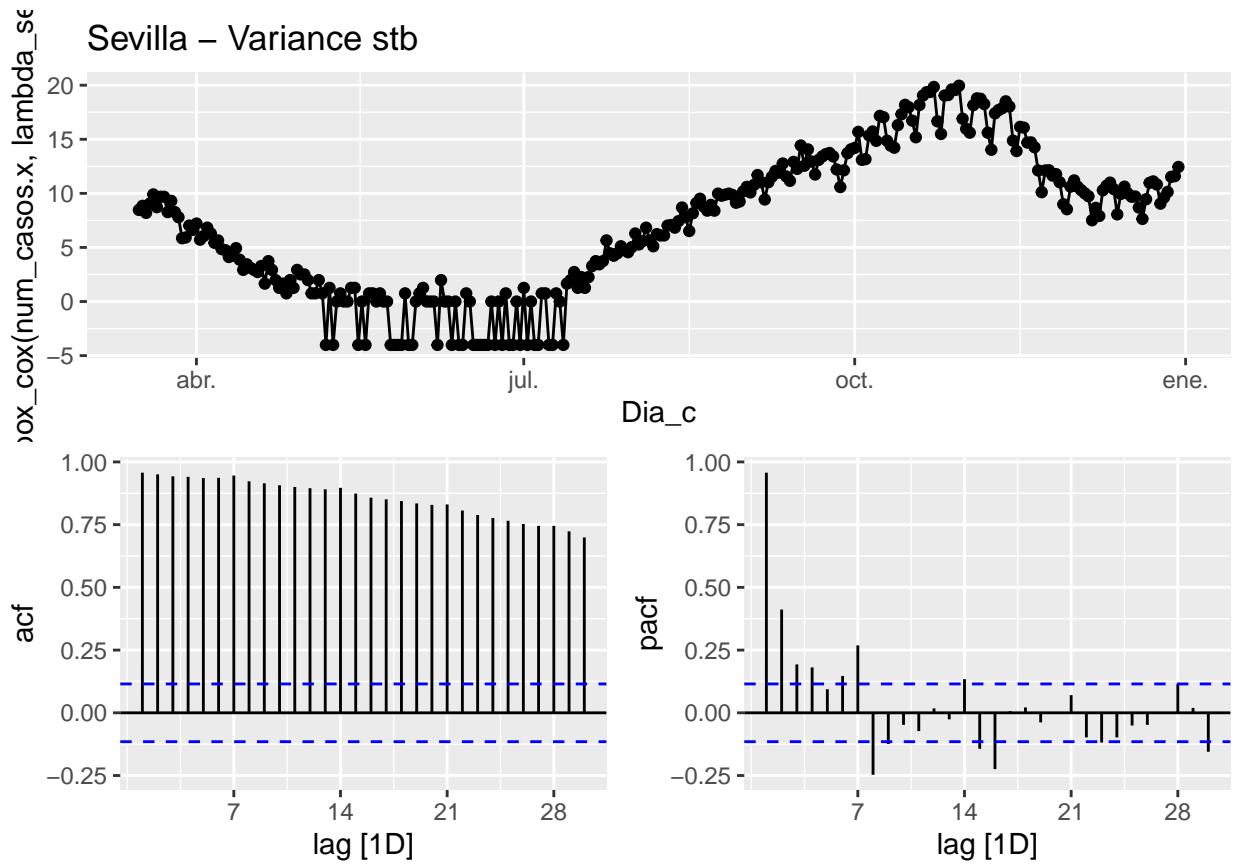
```
Cad_N_cases %>%
  gg_tsdisplay((difference(box_cox(num_casos.x, lambda_cad))),
               plot_type='partial', lag=30) +
  labs(title="Cádiz - Difference once")
```



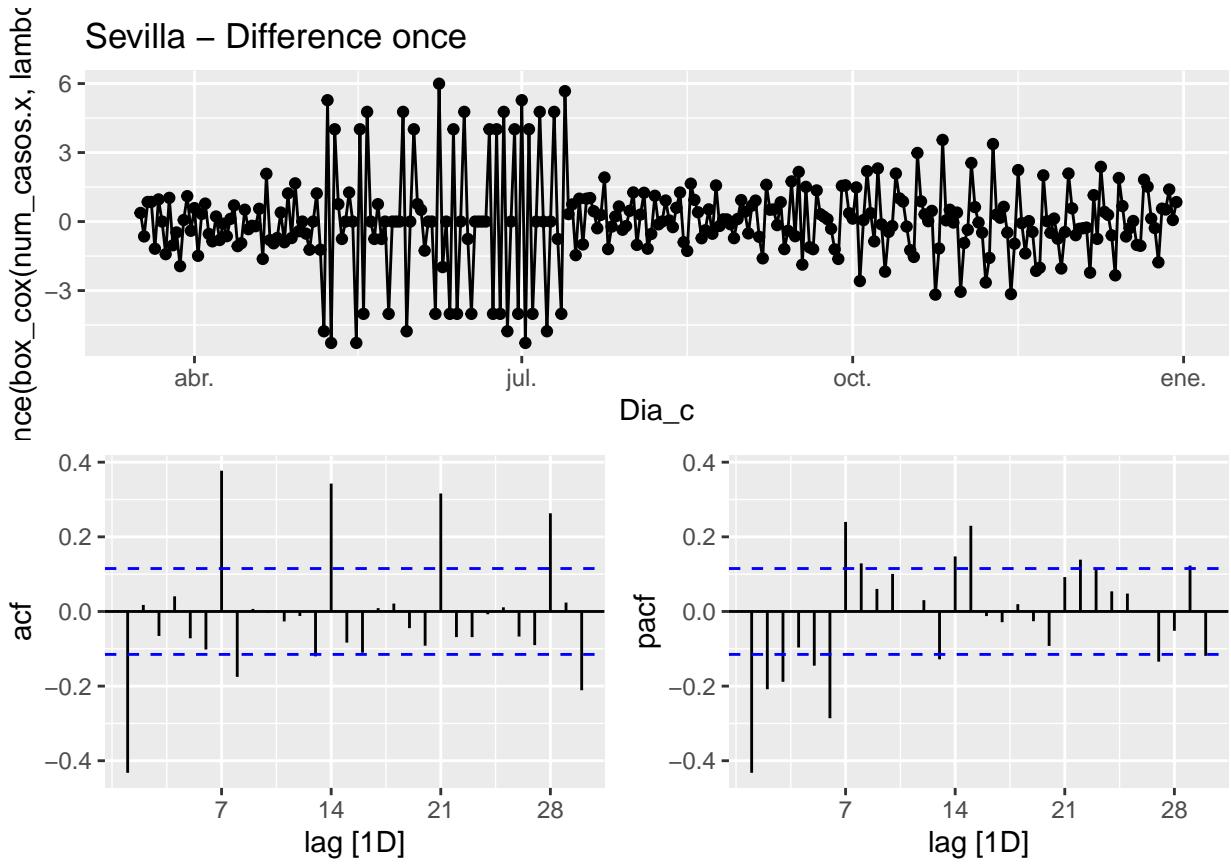
```
Cad_N_cases %>%
  gg_tsdisplay(difference(difference(box_cox(num_casos.x, lambda_cad))),
               plot_type='partial', lag=30) +
  labs(title="Cádiz - Difference double")
```



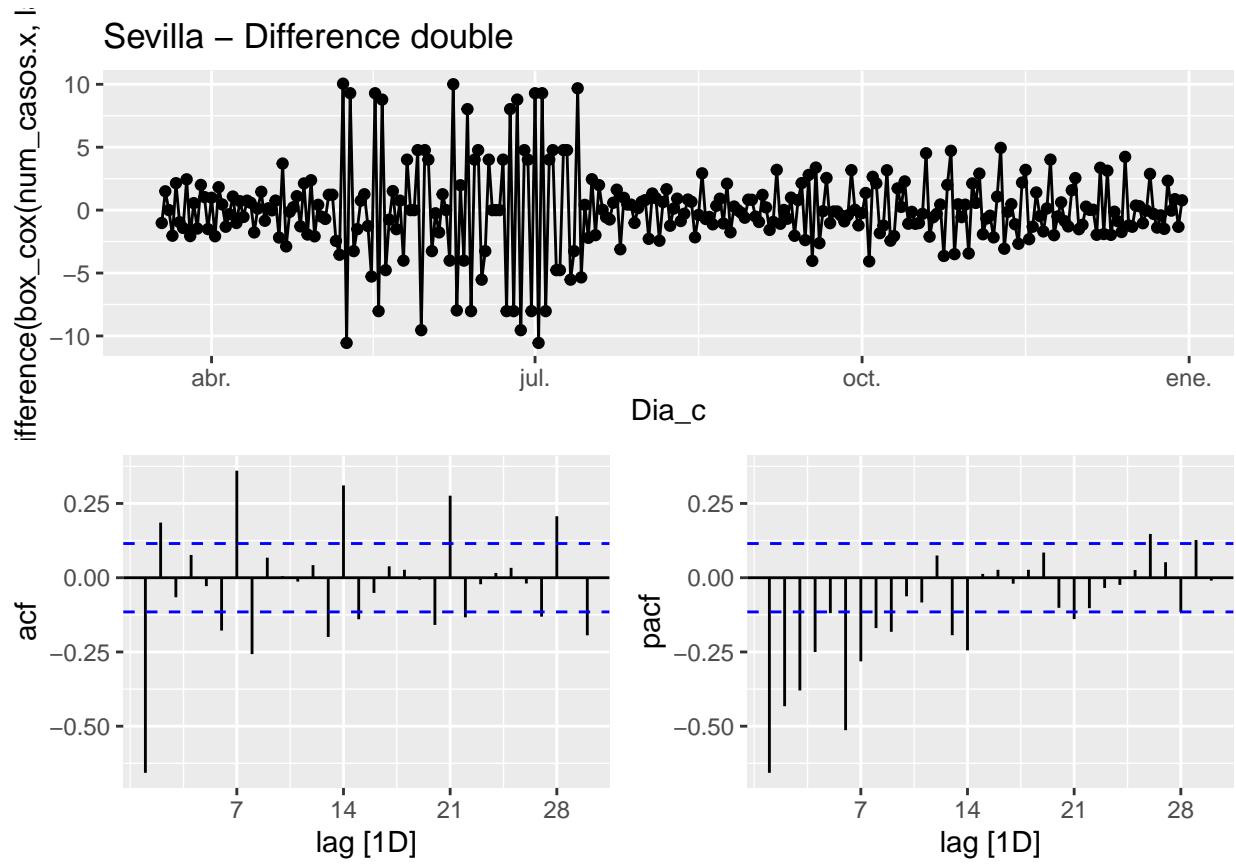
```
#####
Sev_N_cases %>%
  gg_tsdisplay((box_cox(num_casos.x,lambda_sev))),
  plot_type='partial', lag=30) +
  labs(title="Sevilla - Variance stb")
```



```
Sev_N_cases %>%
  gg_tsdisplay((difference(box_cox(num_casos.x, lambda_sev))),
               plot_type='partial', lag=30) +
  labs(title="Sevilla - Difference once")
```



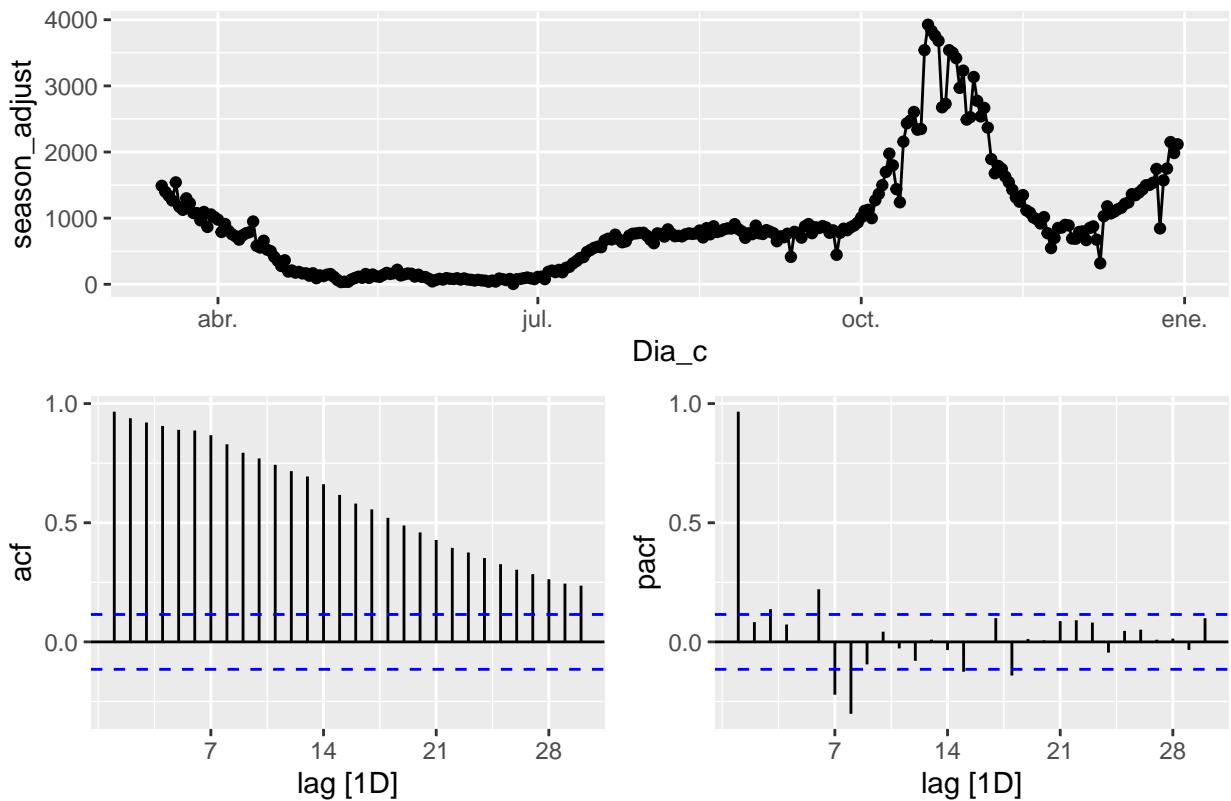
```
Sev_N_cases %>%
  gg_tsdisplay(difference(difference(box_cox(num_casos.x, lambda_sev))),
               plot_type='partial', lag=30) +
  labs(title="Sevilla - Difference double")
```



We have also used the STL season_adjust or trend to help / check to have a time series with less variance.

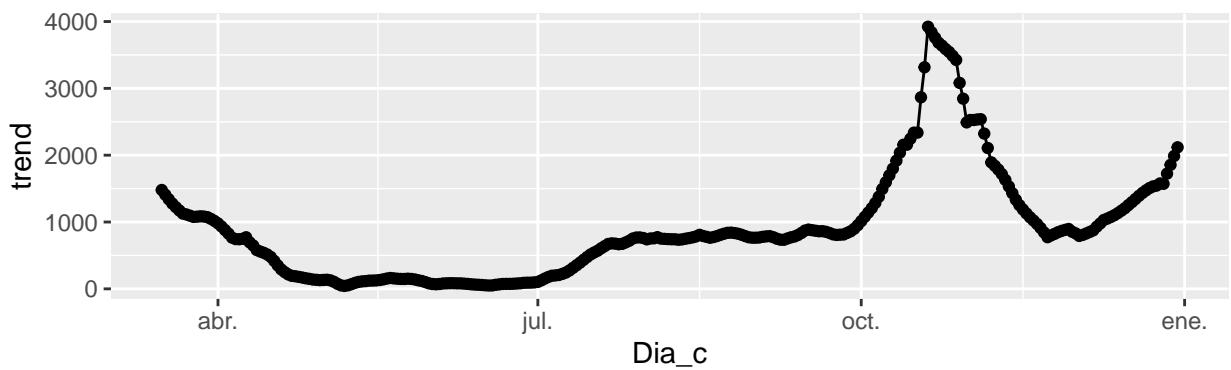
```
## season_adjust
#####
# Variance stb checks (variance + difference)
#####
Bar_N_cases %>% model(STL(num_casos.x ~
  season(window = 7) +
  trend(window = 7), robust = TRUE)) %>%
components() %>%
select(-.model) %>%
gg_tsdisplay(season_adjust,
  plot_type='partial', lag=30) +
  labs(title="Barcelona - Orginal STL season_adjust")
```

Barcelona – Orginal STL season_adjust

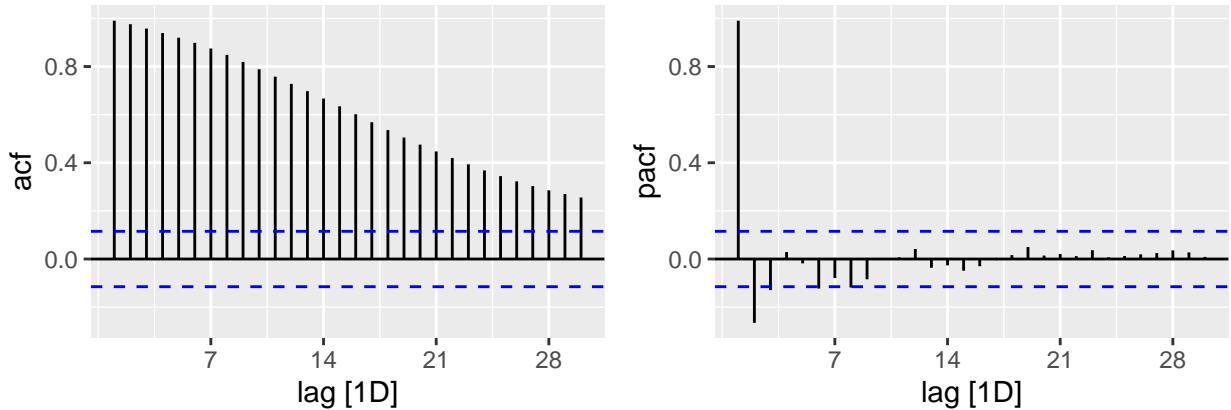


```
Bar_N_cases %>% model(STL(num_casos.x ~
  season(window = 7) +
  trend(window = 7), robust = TRUE)) %>%
components() %>%
select(-.model) %>%
gg_tsdisplay(trend,
  plot_type='partial', lag=30) +
  labs(title="Barcelona - Orginal STL trend")
```

Barcelona – Orginal STL trend

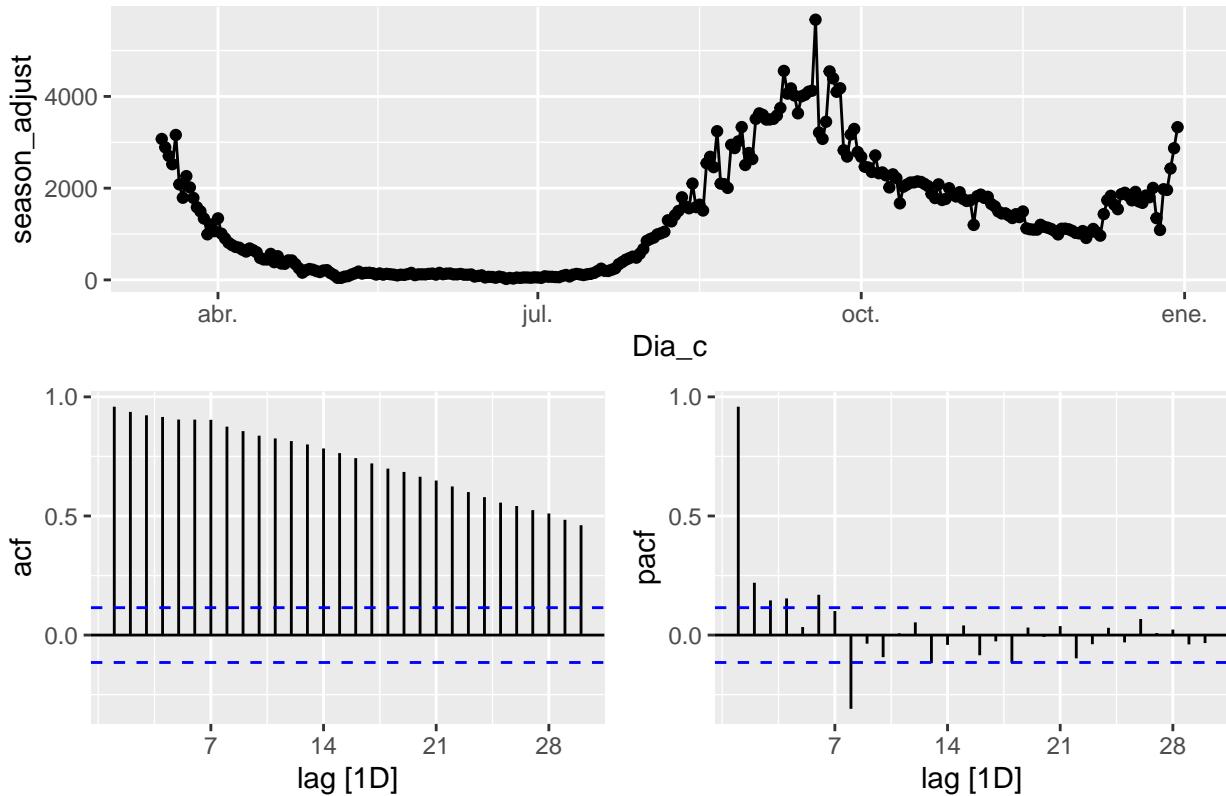


Dia_c



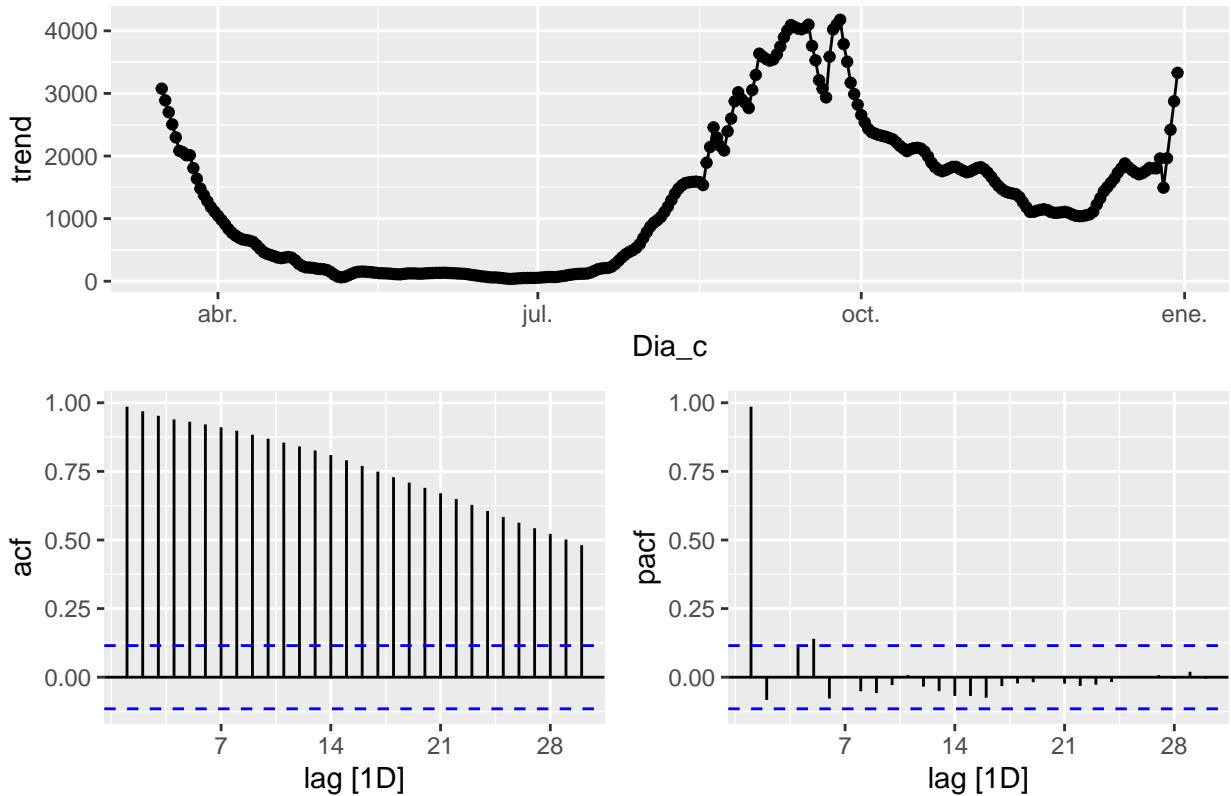
```
#####
Mad_N_cases %>%
  model(STL(num_casos.x ~
    season(window = 7) +
    trend(window = 7), robust = TRUE)) %>%
  components() %>%
  select(-.model) %>%
  gg_tsdisplay(season_adjust,
    plot_type='partial', lag=30) +
  labs(title="Madrid - Original STL season_adjust")
```

Madrid – Original STL season_adjust



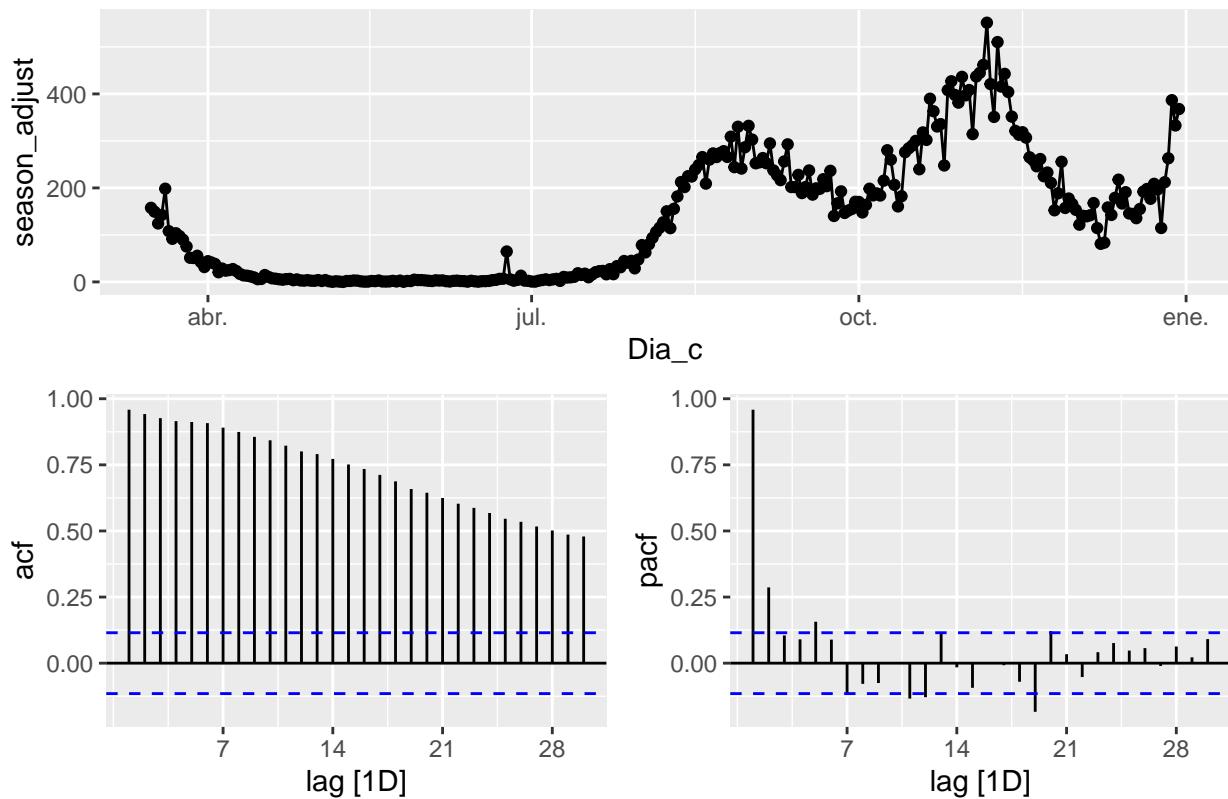
```
Mad_N_cases %>% model(STL(num_casos.x ~
  season(window = 7) +
  trend(window = 7), robust = TRUE)) %>%
components() %>%
select(-.model) %>%
gg_tsdisplay(trend,
  plot_type='partial', lag=30) +
  labs(title="Madrid - Original STL trend")
```

Madrid – Original STL trend

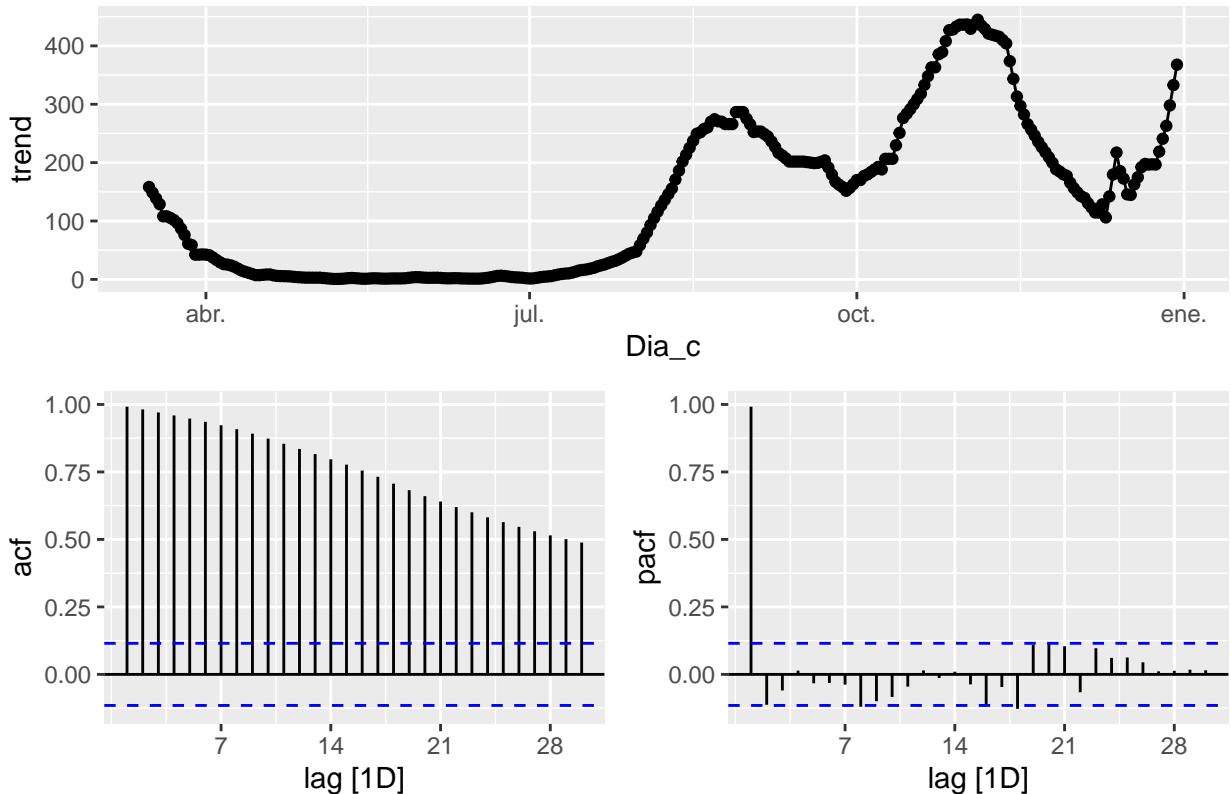


```
#####
Mal_N_cases %>%
  model(STL(num_casos.x ~
    season(window = 7) +
    trend(window = 7), robust = TRUE)) %>%
  components() %>%
  select(-.model) %>%
  gg_tsdisplay(season_adjust,
    plot_type='partial', lag=30) +
  labs(title="Málaga - STL Original season_adjust")
```

Málaga – STL Original season_adjust

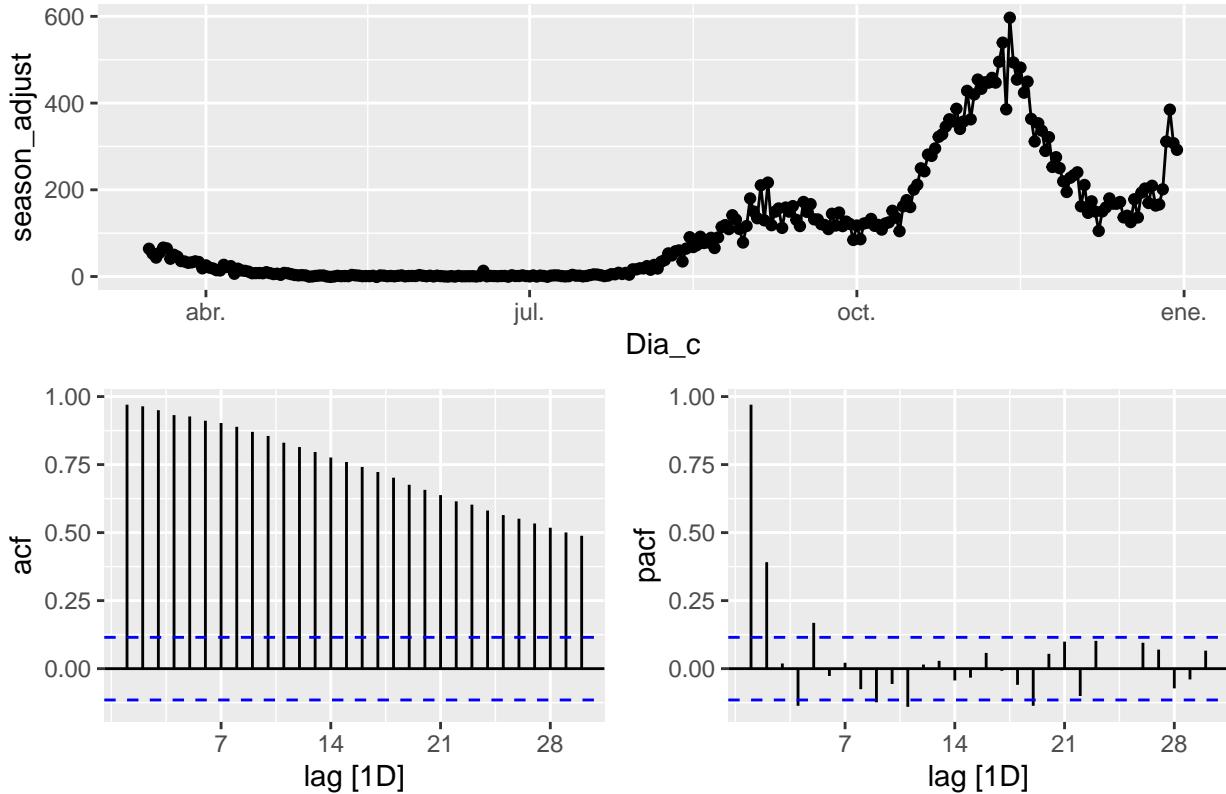


Málaga – STL Original trend



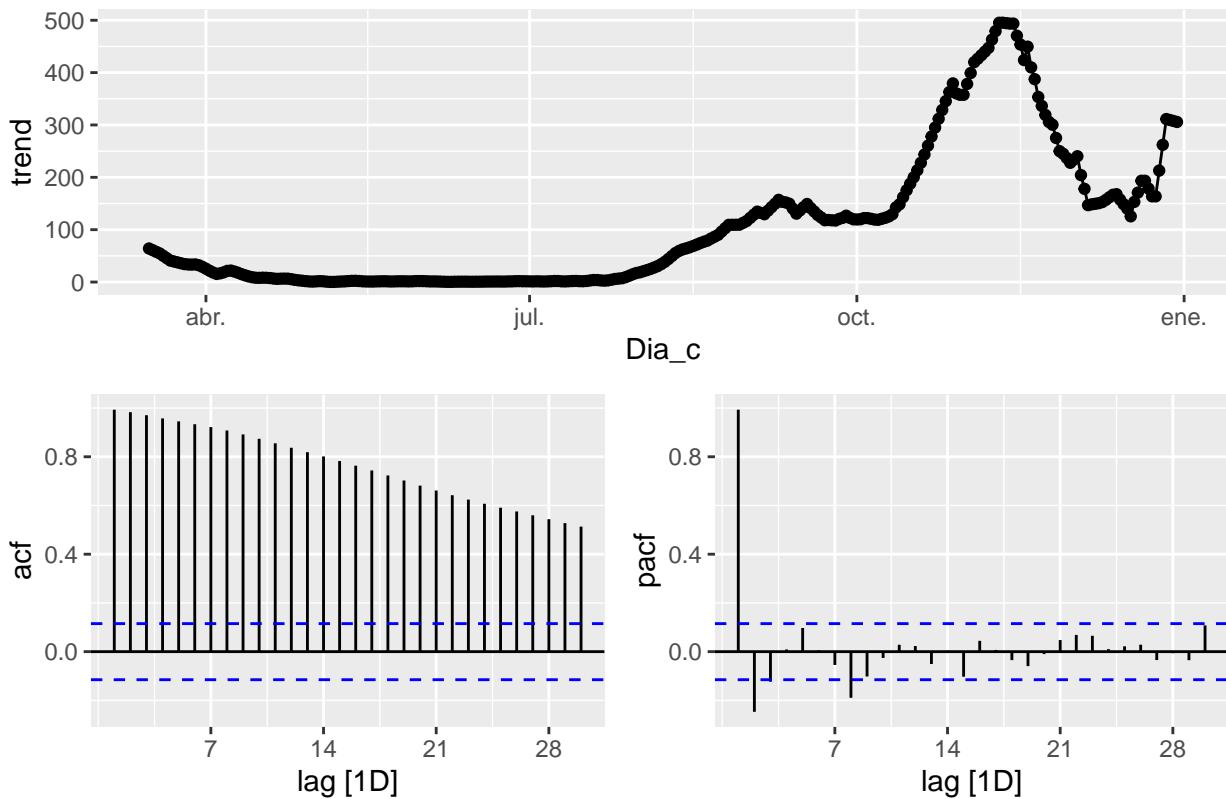
```
#####
Cad_N_cases %>%
  model(STL(num_casos.x ~
    season(window = 7) +
    trend(window = 7), robust = TRUE)) %>%
  components() %>%
  select(-.model) %>%
  gg_tsdisplay(season_adjust,
    plot_type='partial',lag=30) +
  labs(title="Cádiz - STL Original season_adjust")
```

Cádiz – STL Original season_adjust



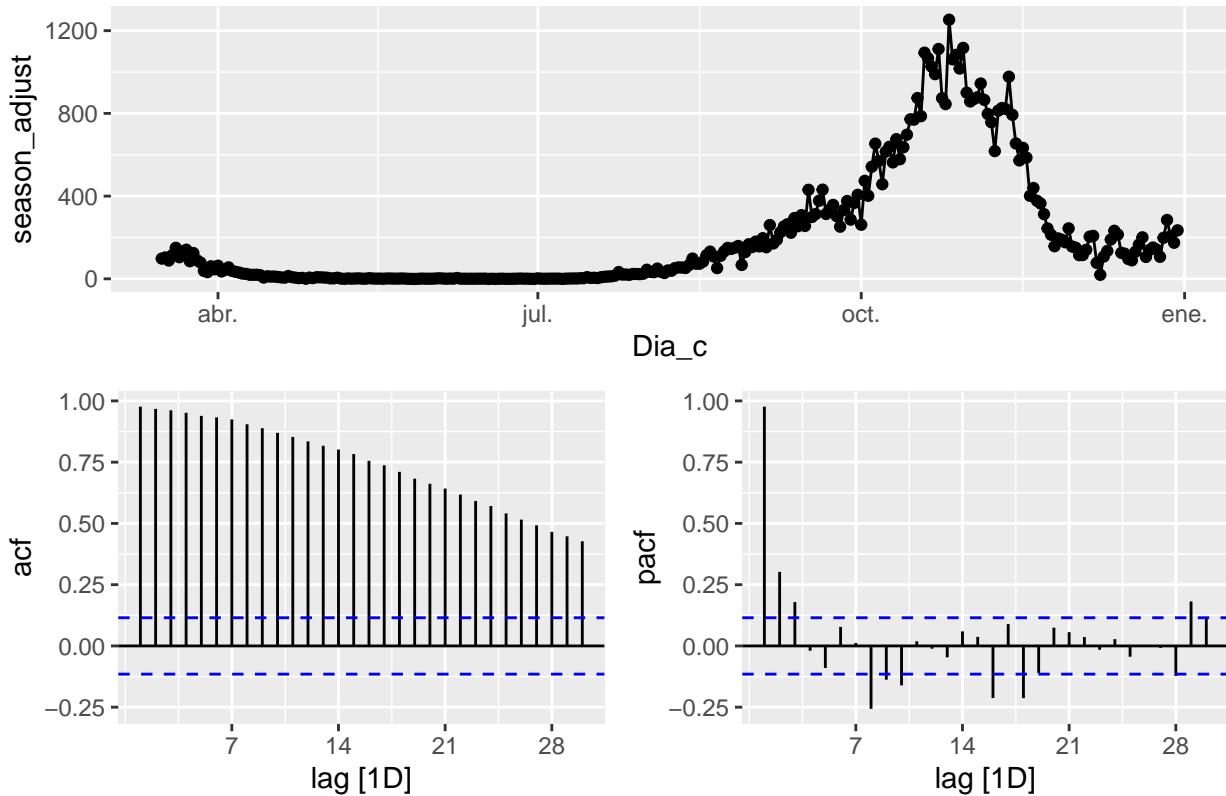
```
Cad_N_cases %>% model(STL(num_casos.x ~
  season(window = 7) +
  trend(window = 7), robust = TRUE)) %>%
components() %>%
select(-.model) %>%
gg_tsdisplay(trend,
  plot_type='partial',lag=30)+
  labs(title="Cádiz - STL Original trend")
```

Cádiz – STL Original trend



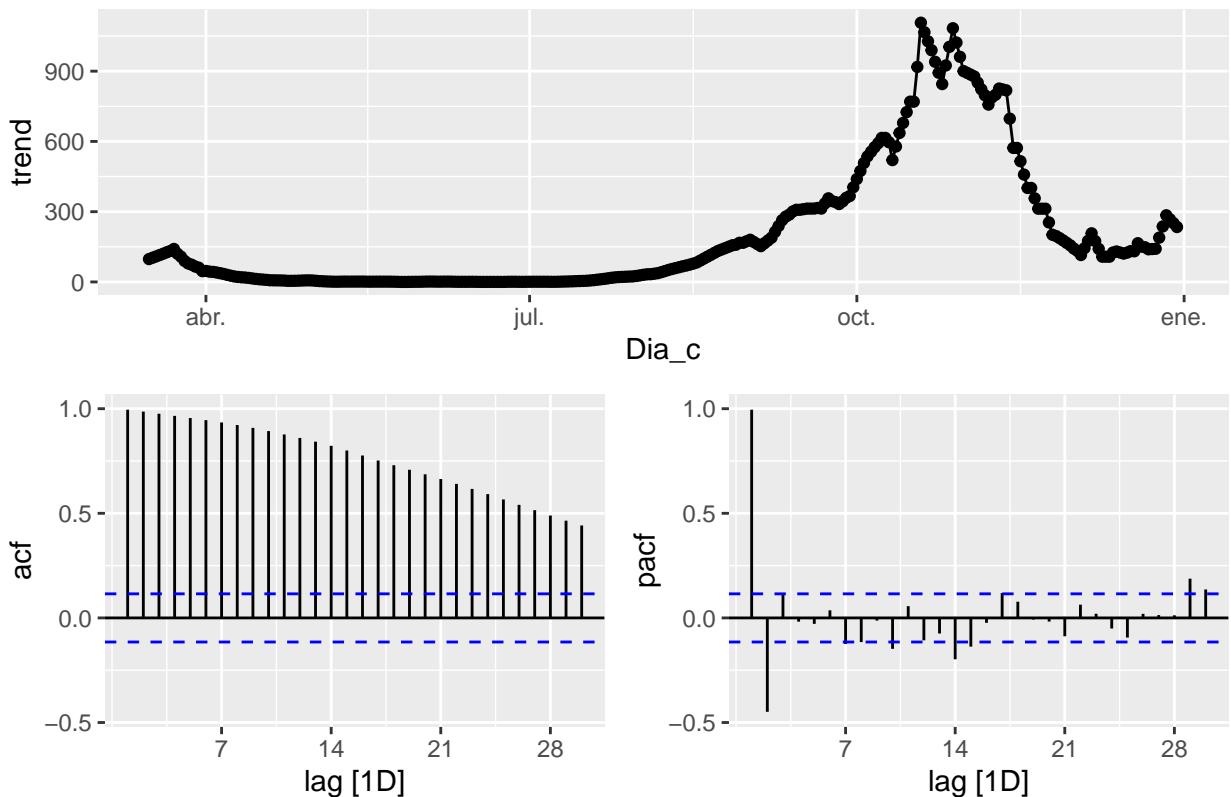
```
#####
Sev_N_cases %>%
  model(STL(num_casos.x ~
    season(window = 7) +
    trend(window = 7), robust = TRUE)) %>%
  components() %>%
  select(-.model) %>%
  gg_tsdisplay(season_adjust,
    plot_type='partial', lag=30) +
  labs(title="Sevilla - STL Original season_adjust")
```

Sevilla – STL Original season_adjust



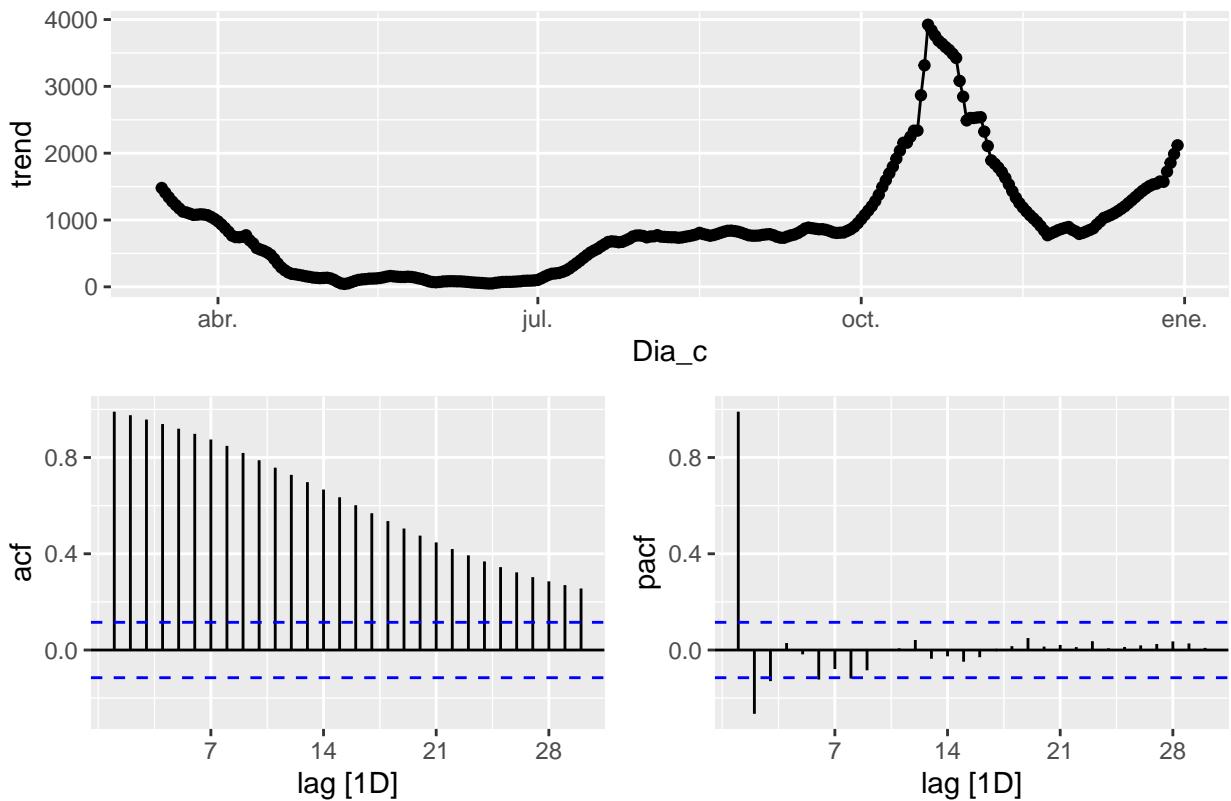
```
Sev_N_cases %>% model(STL(num_casos.x ~
  season(window = 7) +
  trend(window = 7), robust = TRUE)) %>%
components() %>%
select(-.model) %>%
gg_tsdisplay(trend,
  plot_type='partial', lag=30) +
  labs(title="Sevilla - STL Original trend")
```

Sevilla – STL Original trend



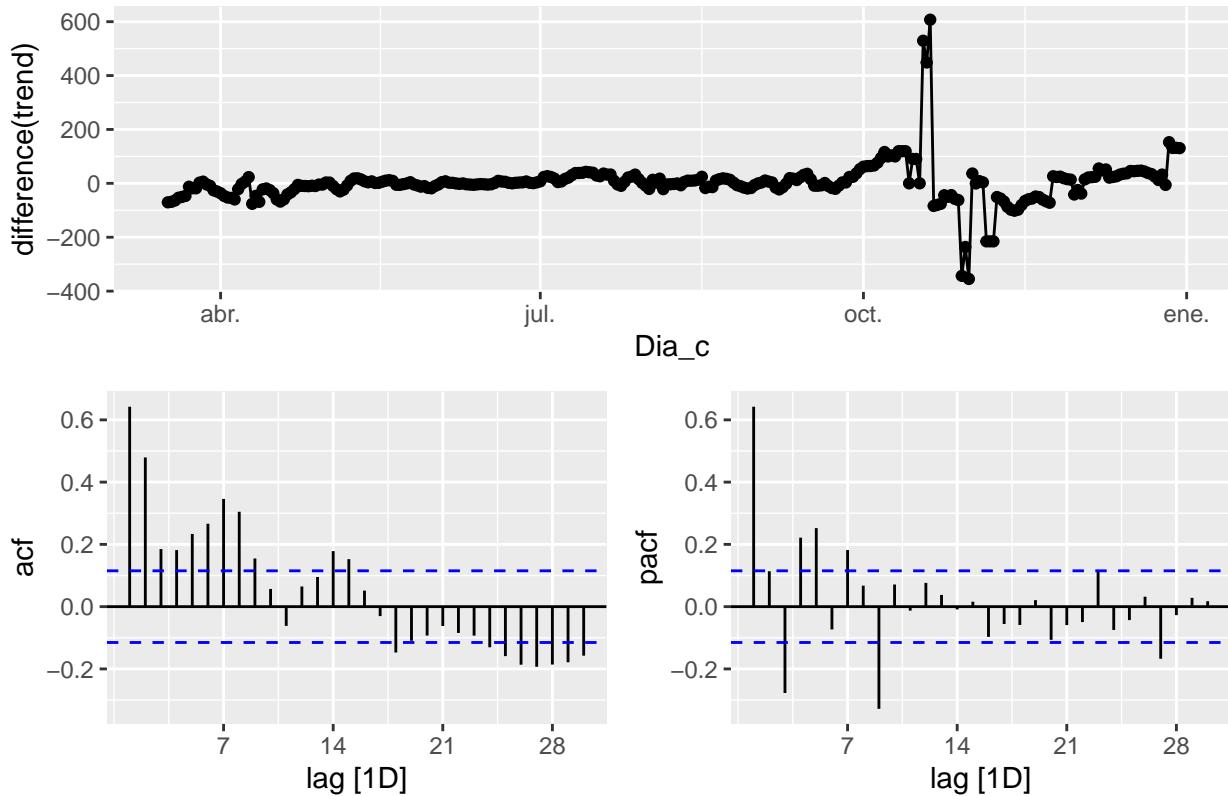
```
# trend
#####
# Variance stb checks (variance + difference)
#####
Bar_N_cases %>% model(STL(num_casos.x ~
  season(window = 7) +
  trend(window = 7), robust = TRUE)) %>%
components() %>%
select(trend) %>%
gg_tsdisplay(trend,
  plot_type='partial', lag=30) +
  labs(title="Barcelona - Orginal STL")
```

Barcelona – Orginal STL

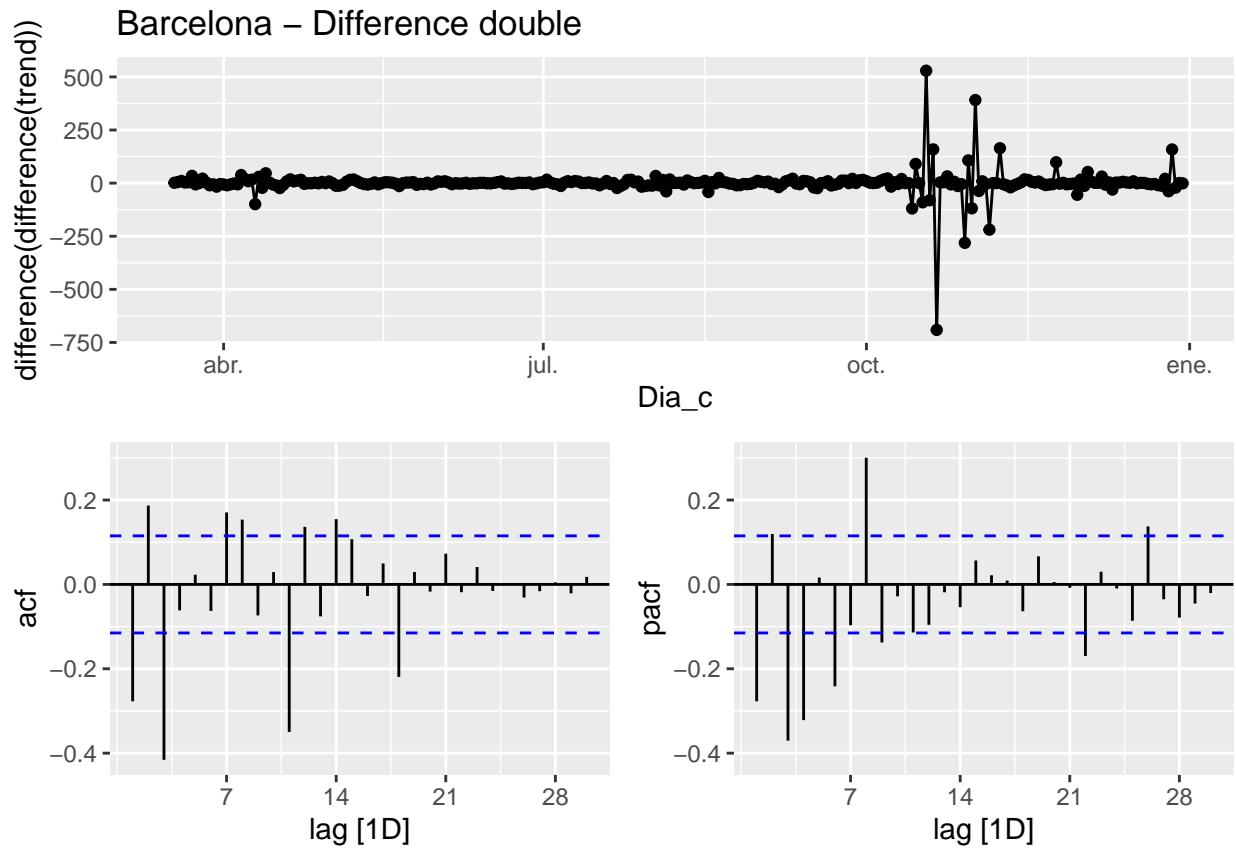


```
Bar_N_cases %>% model(STL(num_casos.x ~
  season(window = 7) +
  trend(window = 7), robust = TRUE)) %>%
components() %>%
select(-.model) %>%
gg_tsdisplay(difference(trend),
  plot_type='partial', lag=30) +
  labs(title="Barcelona - Difference once")
```

Barcelona – Difference once

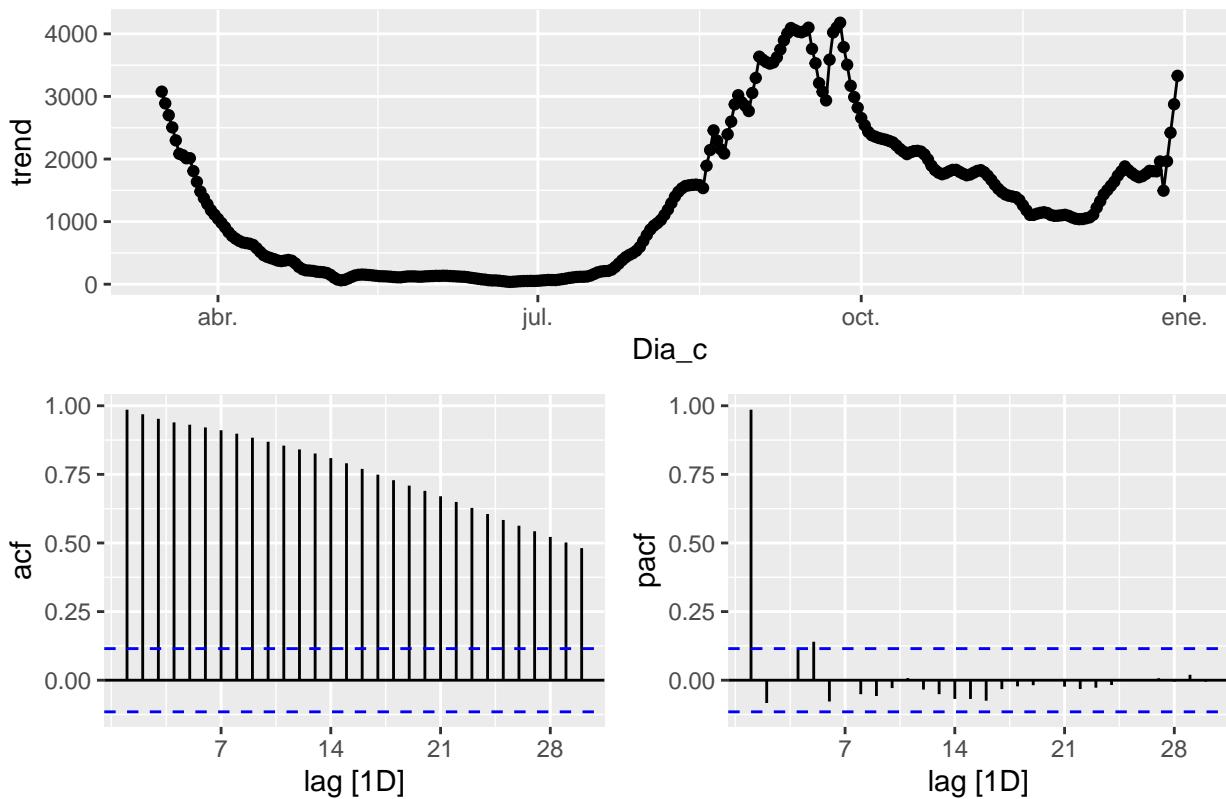


```
Bar_N_cases %>% model(STL(num_casos.x ~
  season(window = 7) +
  trend(window = 7), robust = TRUE)) %>%
components() %>%
select(-.model) %>%
gg_tsdisplay(difference(difference(trend)),
  plot_type='partial', lag=30) +
  labs(title="Barcelona - Difference double")
```



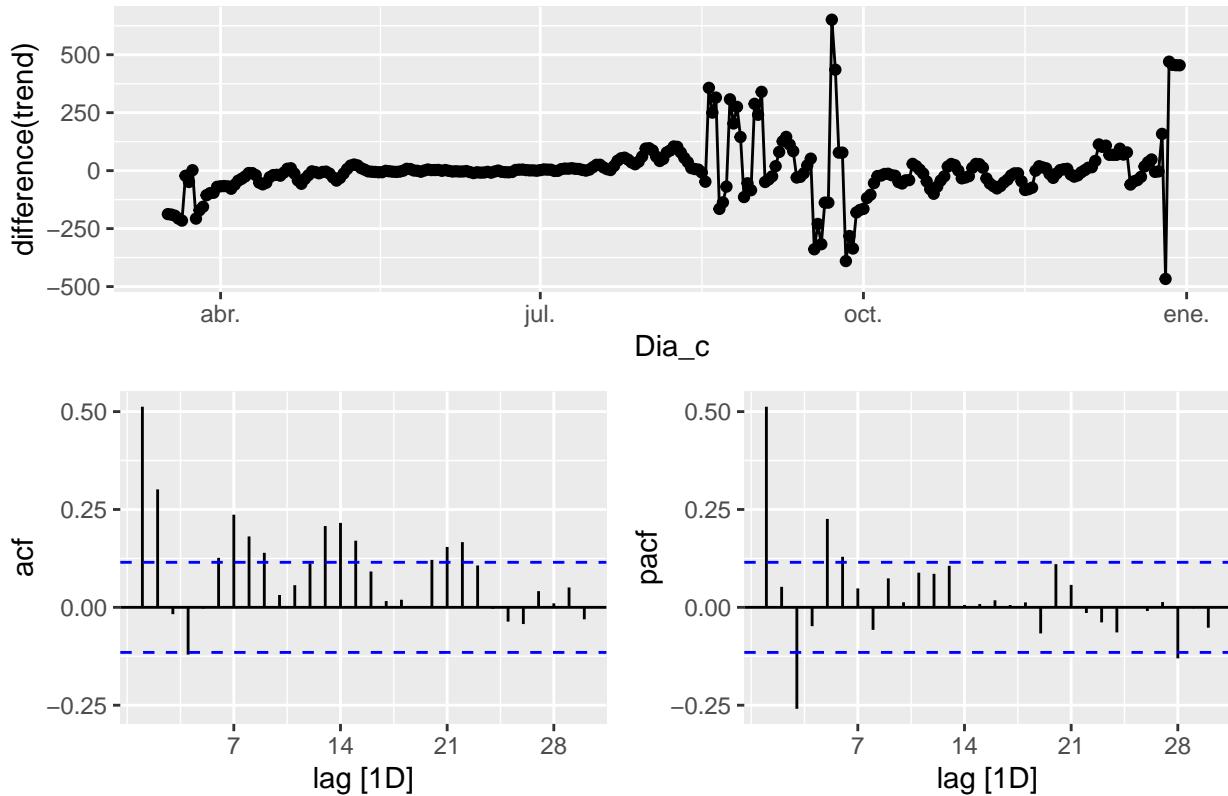
```
#####
Mad_N_cases %>% model(STL(num_casos.x ~
  season(window = 7) +
  trend(window = 7), robust = TRUE)) %>%
components() %>%
select(-.model) %>%
gg_tsdisplay(trend,
  plot_type='partial', lag=30) +
  labs(title="Madrid - Original STL")
```

Madrid – Original STL

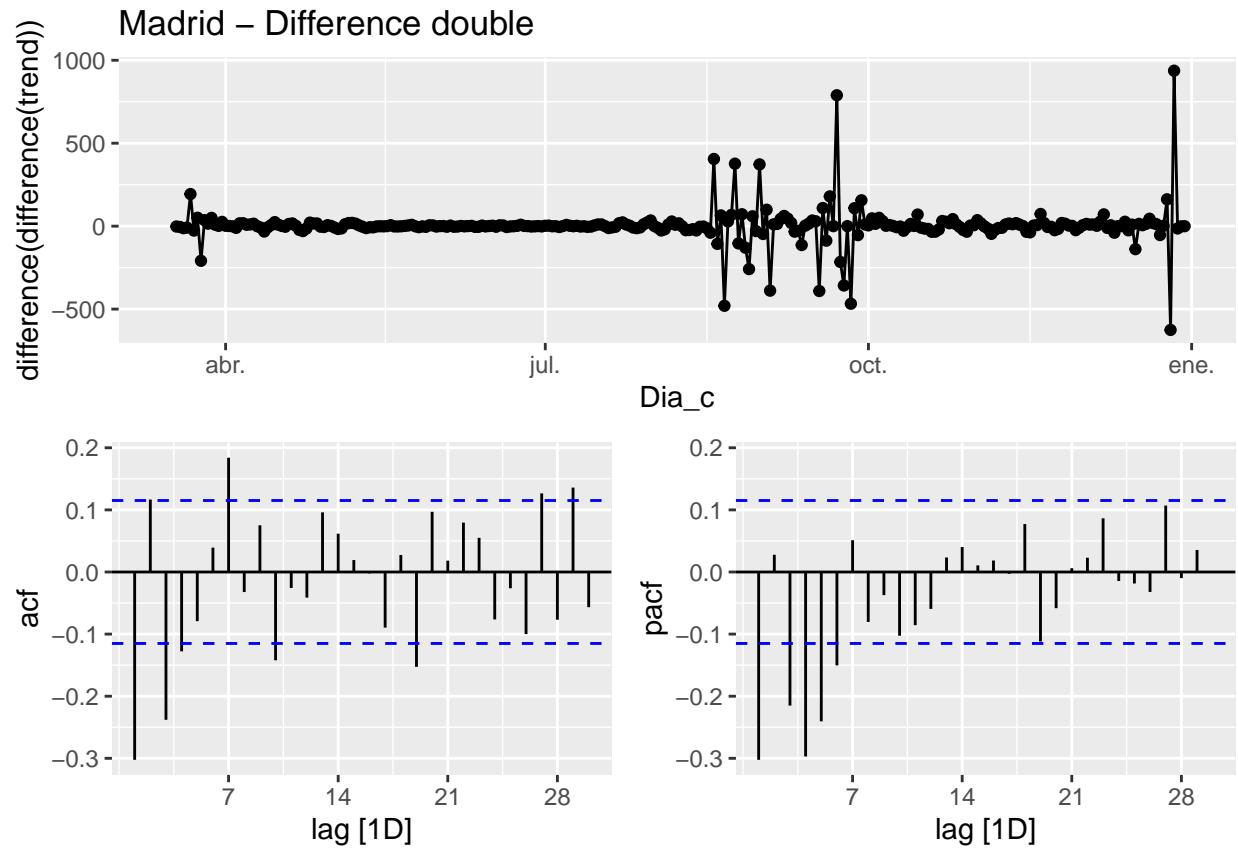


```
Mad_N_cases %>% model(STL(num_casos.x ~
  season(window = 7) +
  trend(window = 7), robust = TRUE)) %>%
components() %>%
select(-.model) %>%
gg_tsdisplay(difference(trend),
  plot_type='partial', lag=30) +
  labs(title="Madrid - Difference once")
```

Madrid – Difference once

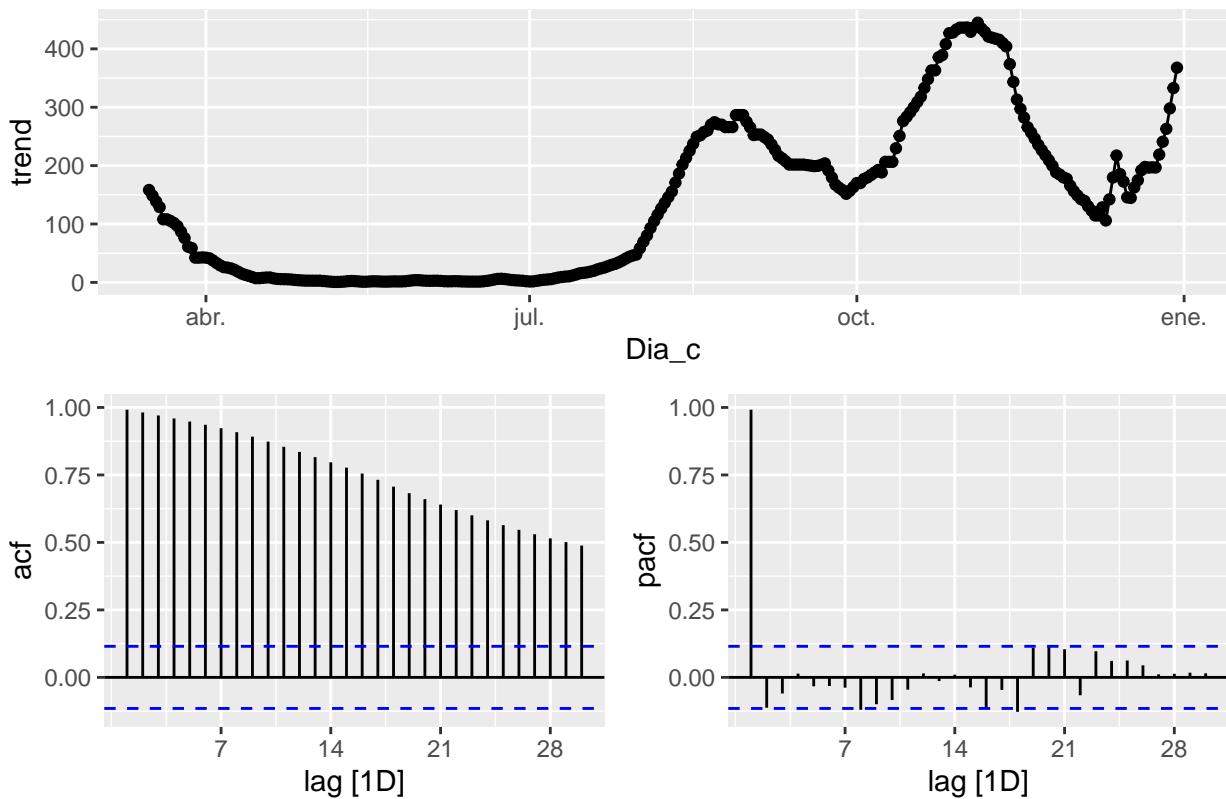


```
Mad_N_cases %>% model(STL(num_casos.x ~
  season(window = 7) +
  trend(window = 7), robust = TRUE)) %>%
components() %>%
select(-.model) %>%
gg_tsdisplay(difference(difference(trend)),
  plot_type='partial', lag=30) +
  labs(title="Madrid - Difference double")
```



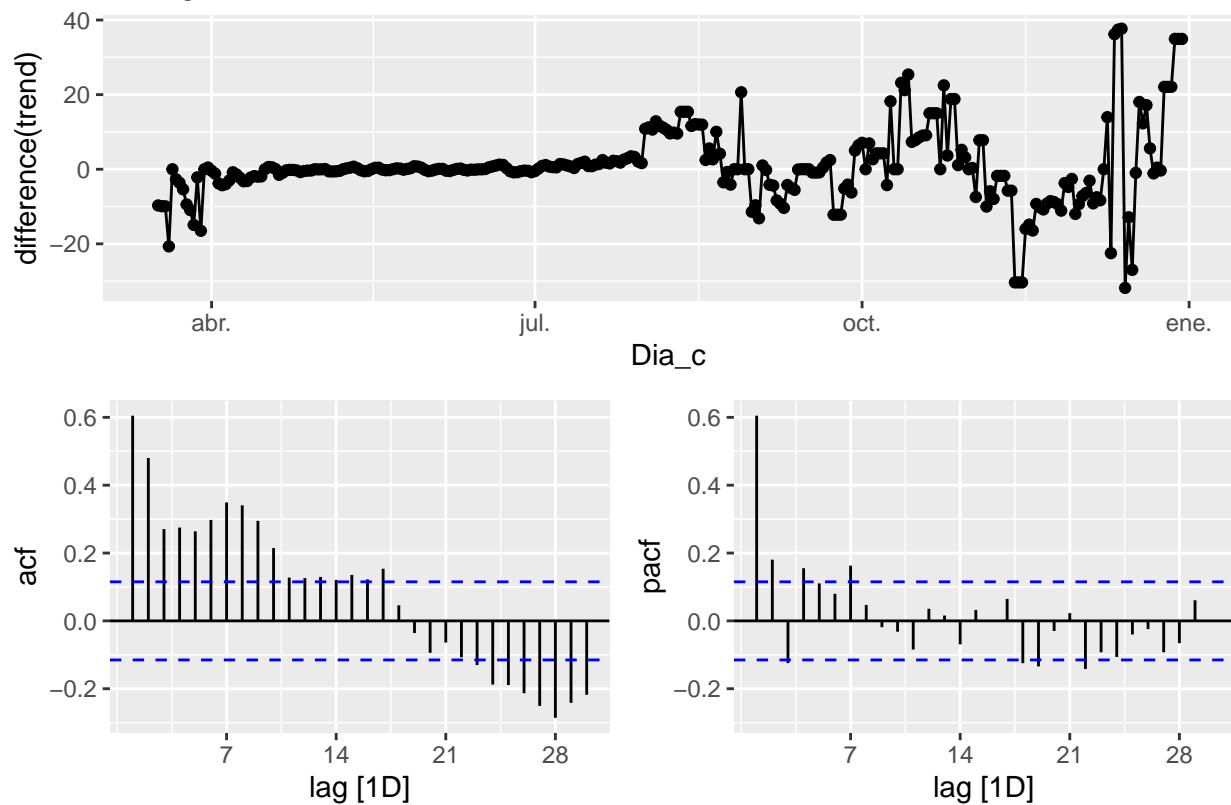
```
#####
Mal_N_cases %>%
  model(STL(num_casos.x ~
    season(window = 7) +
    trend(window = 7), robust = TRUE)) %>%
  components() %>%
  select(-.model) %>%
  gg_tsdisplay(trend,
    plot_type='partial', lag=30) +
  labs(title="Málaga - STL Original")
```

Málaga – STL Original

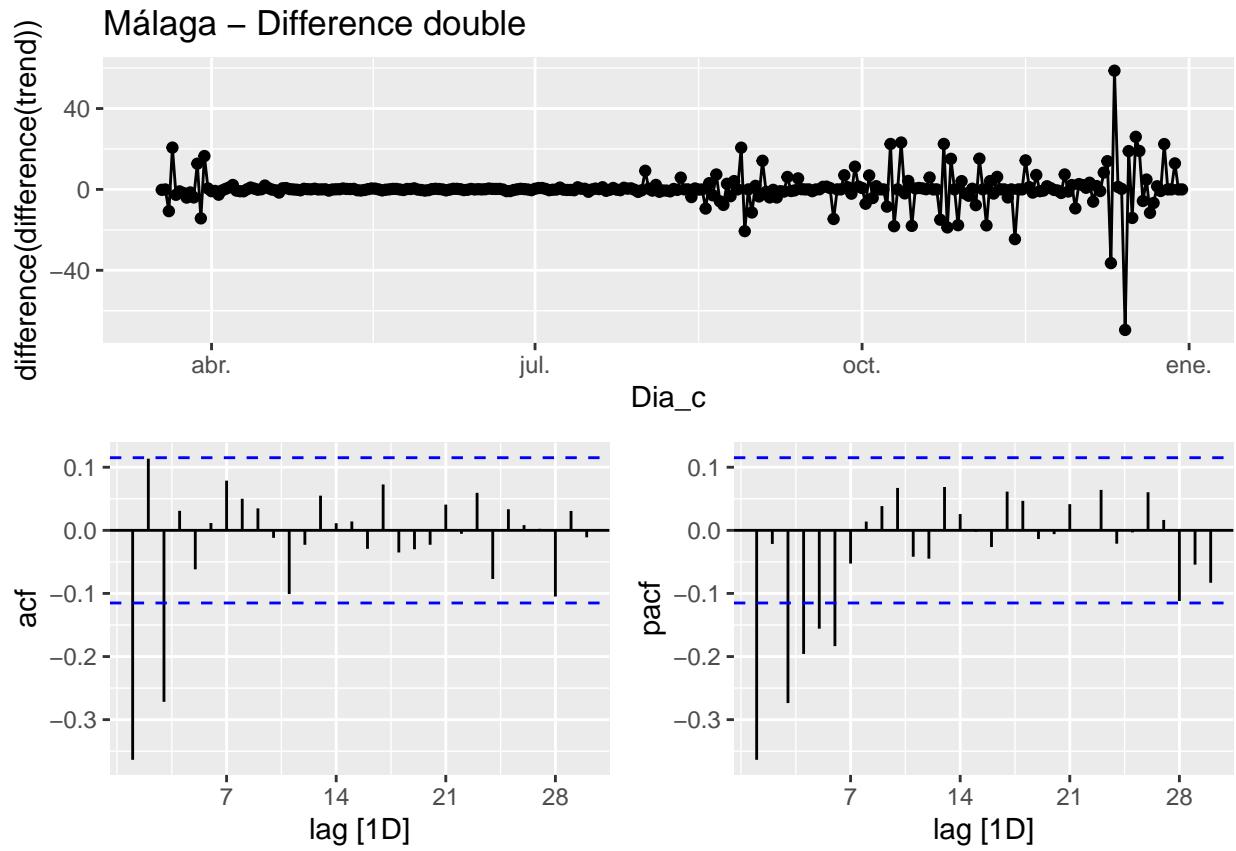


```
Mal_N_cases %>% model(STL(num_casos.x ~
  season(window = 7) +
  trend(window = 7), robust = TRUE)) %>%
components() %>%
select(-.model) %>%
gg_tsdisplay(difference(trend),
  plot_type='partial', lag=30) +
  labs(title="Málaga - Difference once")
```

Málaga – Difference once

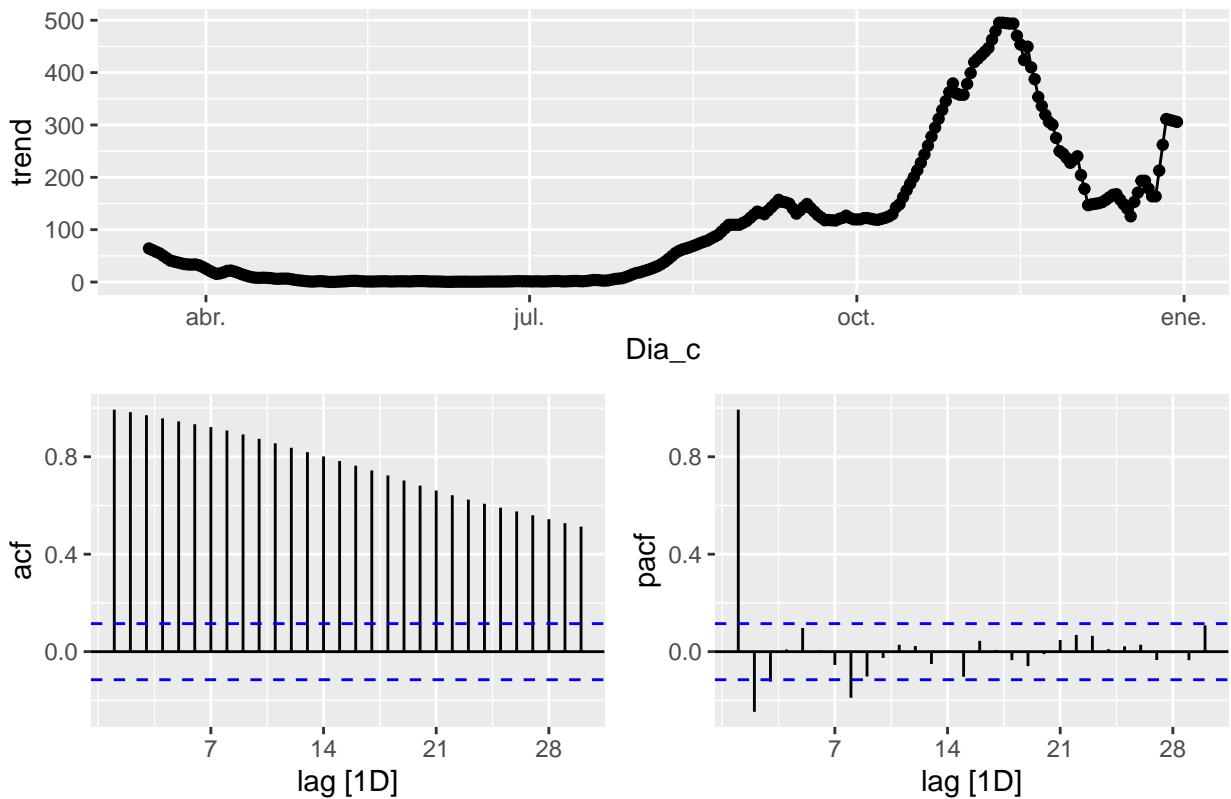


```
Mal_N_cases %>%
  model(STL(num_casos.x ~
    season(window = 7) +
    trend(window = 7), robust = TRUE)) %>%
  components() %>%
  select(-.model) %>%
  gg_tsdisplay(difference(difference(trend)),
    plot_type='partial', lag=30) +
  labs(title="Málaga - Difference double")
```



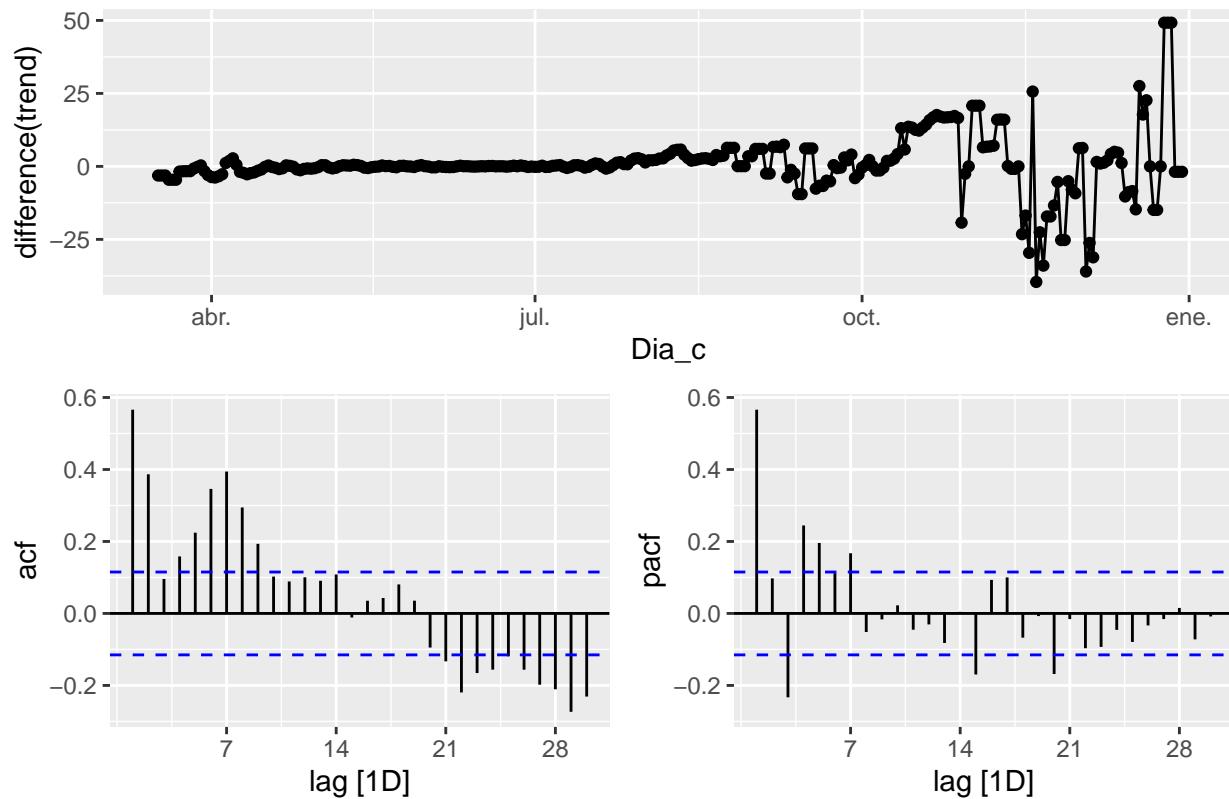
```
#####
Cad_N_cases %>%
  model(STL(num_casos.x ~
    season(window = 7) +
    trend(window = 7), robust = TRUE)) %>%
  components() %>%
  select(-.model) %>%
  gg_tsdisplay(trend,
    plot_type='partial', lag=30) +
  labs(title="Cádiz - STL Original")
```

Cádiz – STL Original

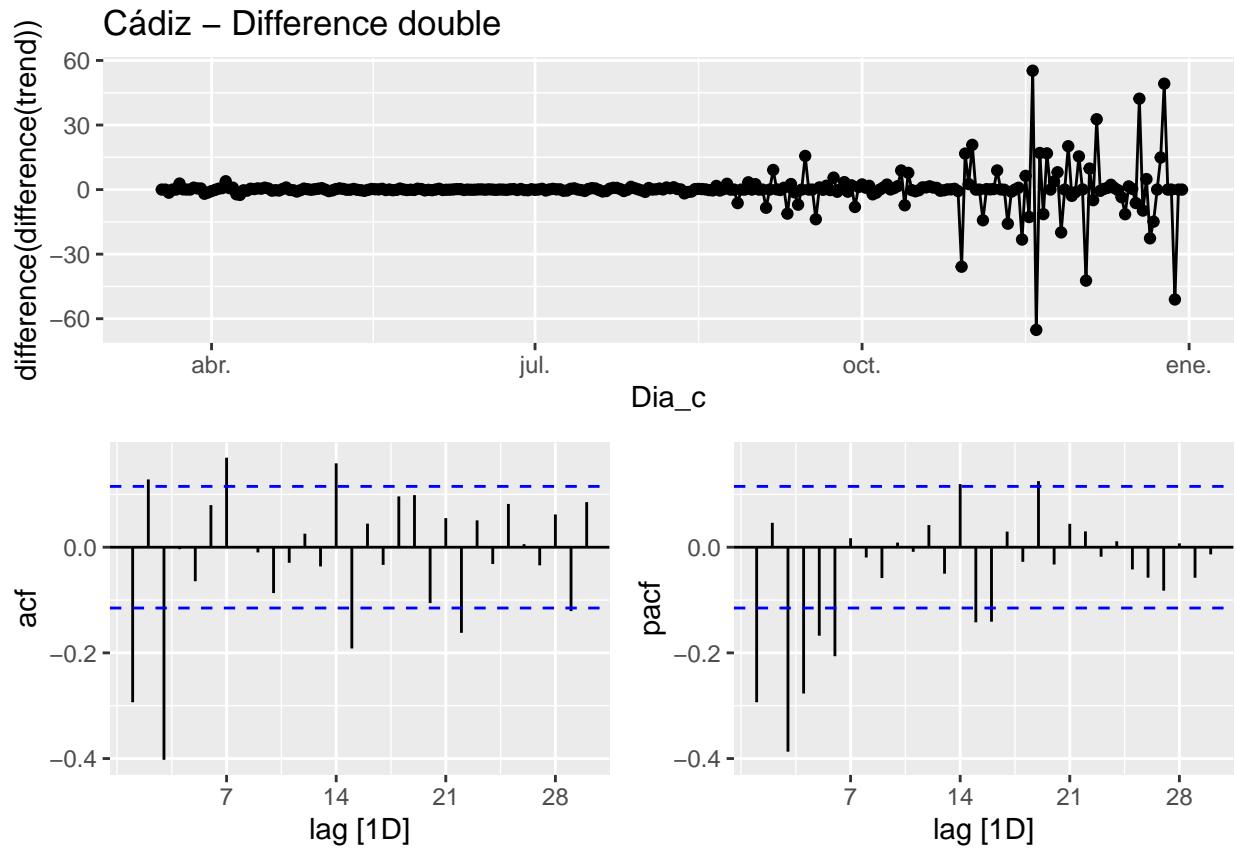


```
Cad_N_cases %>% model(STL(num_casos.x ~
  season(window = 7) +
  trend(window = 7), robust = TRUE)) %>%
components() %>%
select(-.model) %>%
gg_tsdisplay(difference(trend),
  plot_type='partial',lag=30) +
  labs(title="Cádiz - Difference once")
```

Cádiz – Difference once

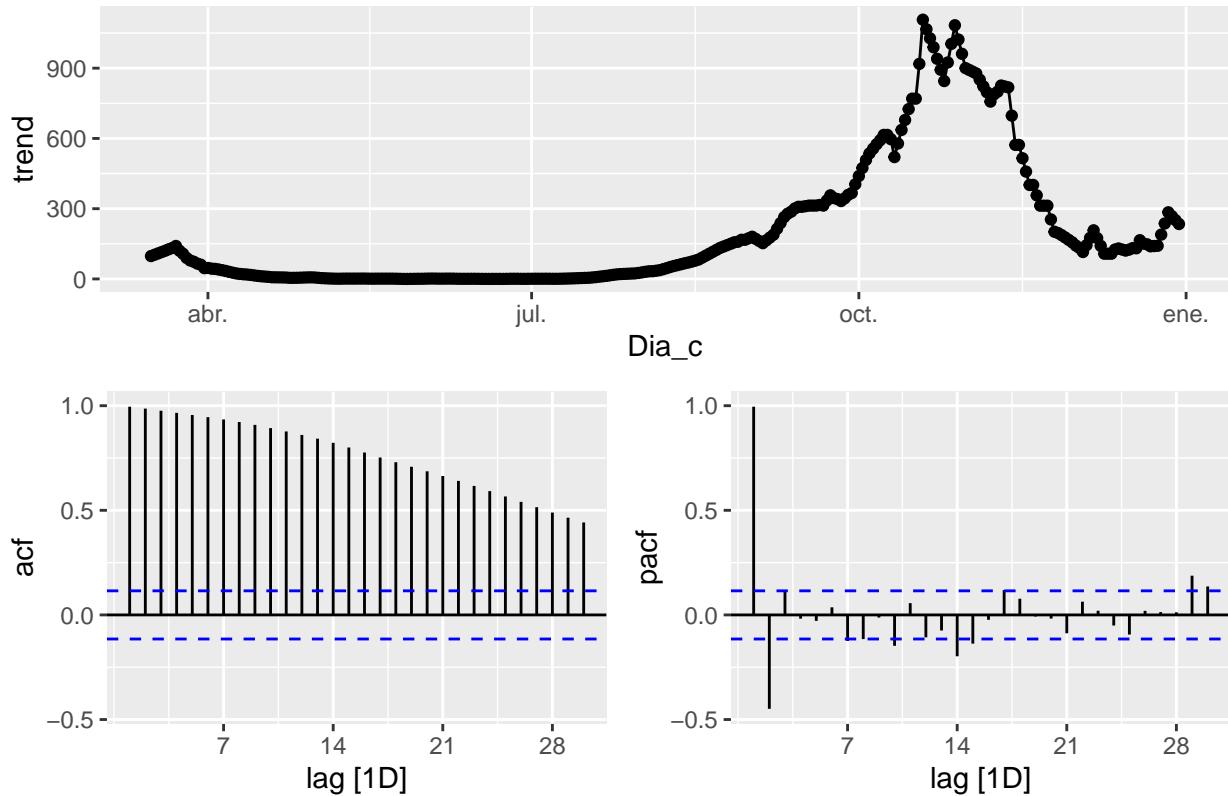


```
Cad_N_cases %>% model(STL(num_casos.x ~
  season(window = 7) +
  trend(window = 7), robust = TRUE)) %>%
components() %>%
select(-.model) %>%
gg_tsdisplay(difference(difference(trend)),
  plot_type='partial',lag=30) +
  labs(title="Cádiz - Difference double")
```



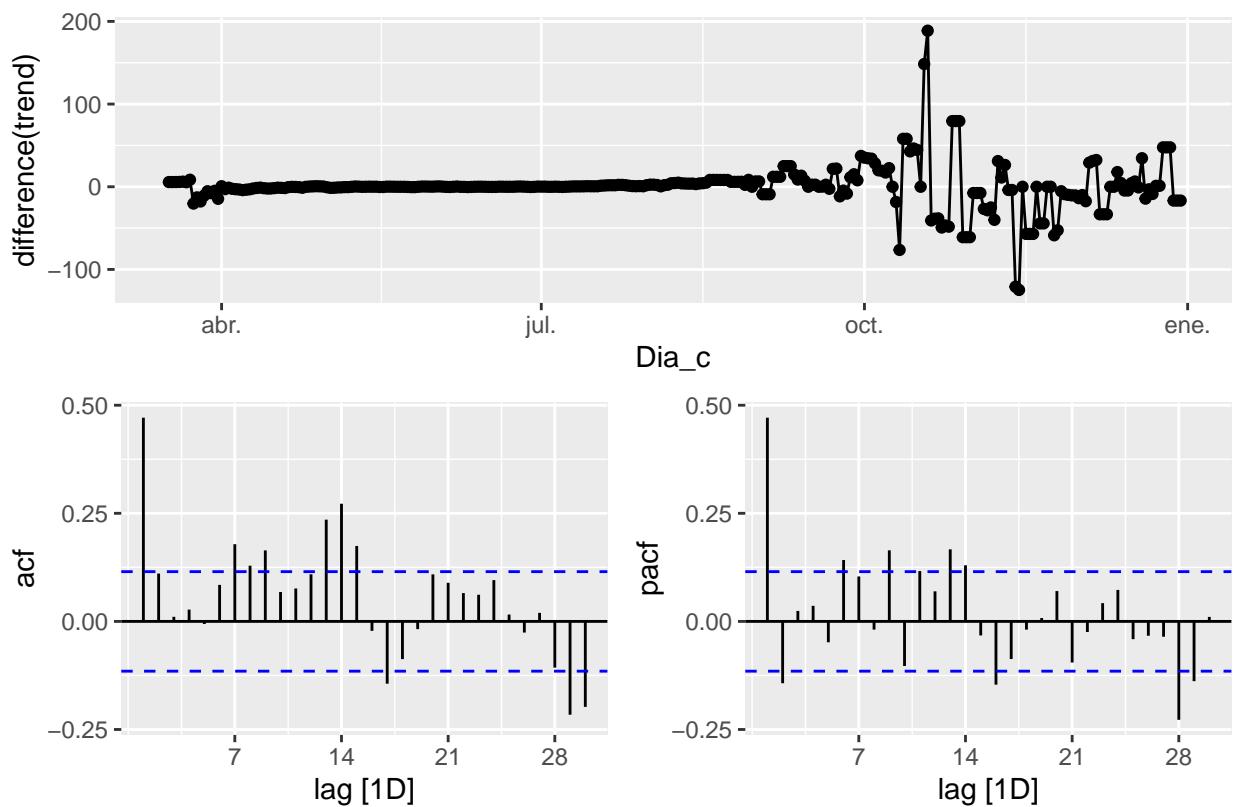
```
#####
Sev_N_cases %>% model(STL(num_casos.x ~
  season(window = 7) +
  trend(window = 7), robust = TRUE)) %>%
components() %>%
select(-.model) %>%
gg_tsdisplay(trend,
  plot_type='partial', lag=30) +
  labs(title="Sevilla - STL Original")
```

Sevilla – STL Original

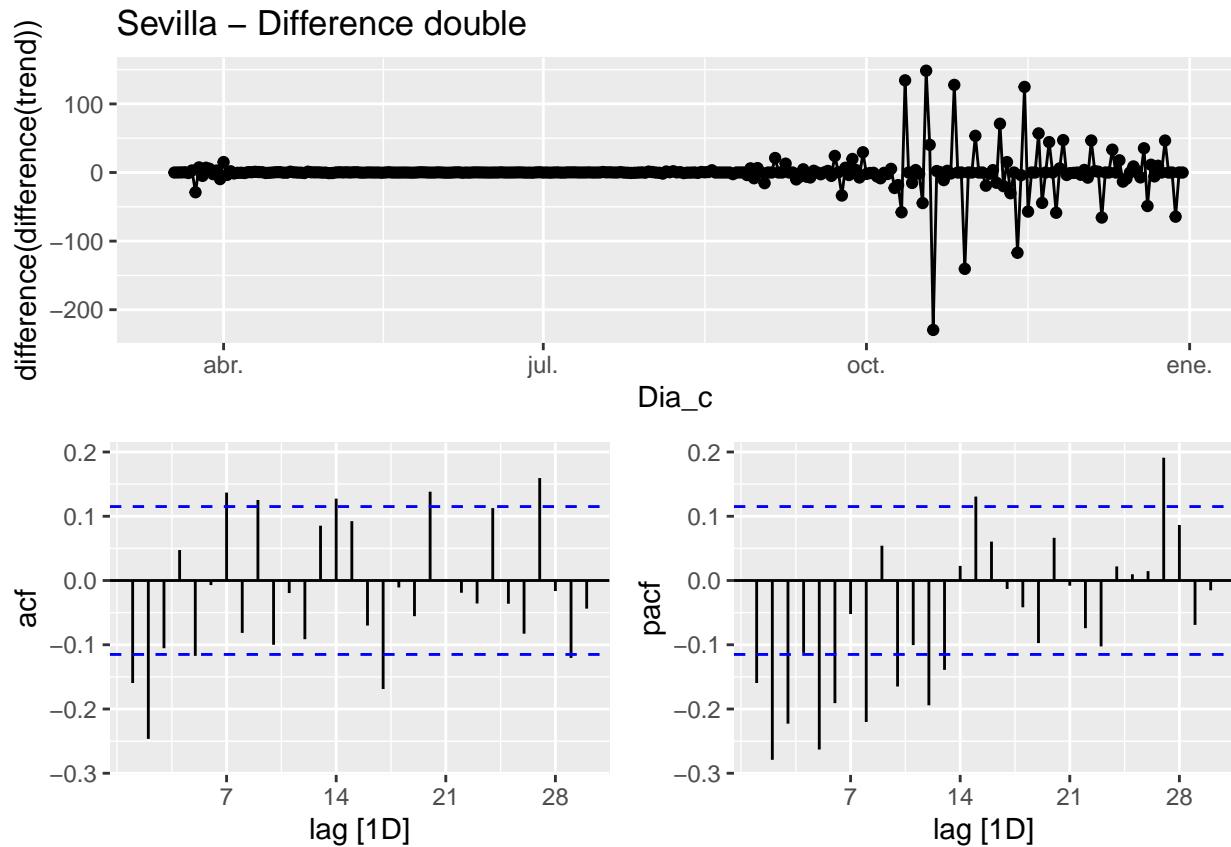


```
Sev_N_cases %>% model(STL(num_casos.x ~
  season(window = 7) +
  trend(window = 7), robust = TRUE)) %>%
components() %>%
select(-.model) %>%
gg_tsdisplay(difference(trend),
  plot_type='partial', lag=30) +
  labs(title="Sevilla - Difference once")
```

Sevilla – Difference once



```
Sev_N_cases %>% model(STL(num_casos.x ~
  season(window = 7) +
  trend(window = 7), robust = TRUE)) %>%
components() %>%
select(-.model) %>%
gg_tsdisplay(difference(difference(trend)),
  plot_type='partial', lag=30) +
  labs(title="Sevilla - Difference double")
```



3.3 Model and Forecast (Barcelona, Madrid, Málaga, Sevilla and Cádiz)

3.3.1 Univariate (7, 14, 17 days) Barcelona, Madrid, Málaga, Sevilla and Cádiz

As stated by (Hyndman and Athanasopoulos 2021)... “The ARIMA() function uses unitroot_nsdiffs() to determine D (the number of seasonal differences to use), and unitroot_ndiffs() to determine d (the number of ordinary differences to use), when these are not specified.”

Due to we have lots of or close to zero infections between april and august, we are going to star the analysis the **second week** of august to avoid problems (negative values when transform) during the ARIMA process (We have used, for each analysis and filter, the 1st day of the week as starting point due to the “weekly seasonality” detected).

```
# Train and Test data set creation
# Train
Bar_N_cases_tr <- Bar_N_cases %>%
  filter_index("2020-08-10" ~ "2020-12-13")
# Test
Bar_N_cases_tt <- Bar_N_cases %>%
  filter_index("2020-12-14" ~ "2020-12-31")

# Modeling over train
Bar_N_cases_tr %>%
  model(arima_at1=ARIMA(box_cox(num_casos.x,lambda_bar)),
        arima_at2=ARIMA(box_cox(num_casos.x,lambda_bar),
                        stepwise = FALSE, approx = FALSE),
```

```

Snaive=SNAIVE(box_cox(num_casos.x,lambda_bar))) -> fit_model

# Show and report model
fit_model

```

3.3.1.1 Barcelona

```

## # A mable: 1 x 4
## # Key:   sub_region_2 [1]
##   sub_region_2           arima_at1           arima_at2   Snaive
##   <chr>                 <model>             <model>   <model>
## 1 Barcelona    <ARIMA(1,0,1)(2,1,0)[7]> <ARIMA(1,0,1)(2,1,0)[7]> <SNAIVE>
report(fit_model)

## # A tibble: 3 x 9
##   sub_region_2 .model   sigma2 log_lik   AIC   AICc   BIC ar_roots ma_roots
##   <chr>        <chr>     <dbl>  <dbl> <dbl> <dbl> <dbl> <list>   <list>
## 1 Barcelona    arima_at1  0.132   -48.3 107. 107. 120. <cpl [15]> <cpl [1]>
## 2 Barcelona    arima_at2  0.132   -48.3 107. 107. 120. <cpl [15]> <cpl [1]>
## 3 Barcelona    Snaive     0.505     NA     NA     NA  <NULL>  <NULL>

# Good model >> Less Sigma / AICc
fit_model %>% pivot_longer(!sub_region_2,
                           names_to = "Model name",
                           values_to = "Orders")

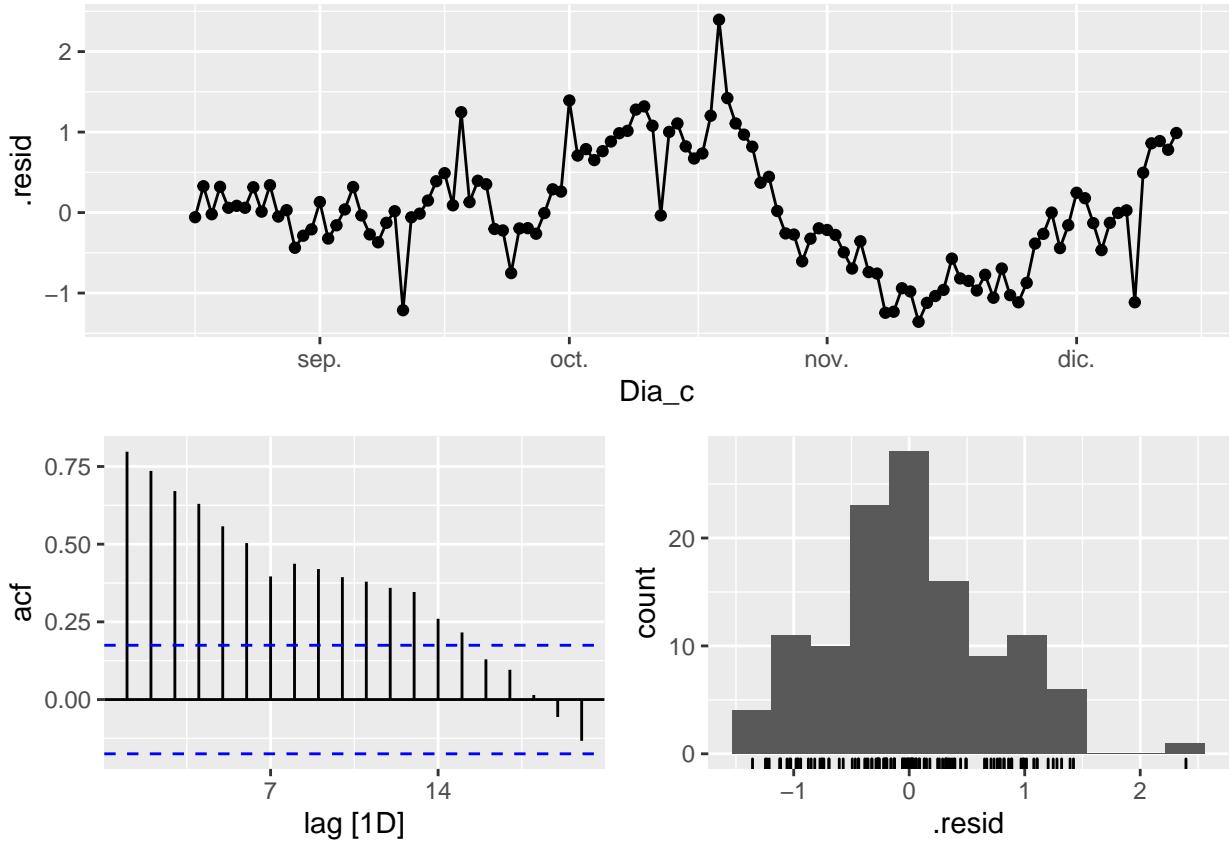
## # A mable: 3 x 3
## # Key:   sub_region_2, Model name [3]
##   sub_region_2 `Model name`          Orders
##   <chr>        <chr>                <model>
## 1 Barcelona    arima_at1    <ARIMA(1,0,1)(2,1,0)[7]>
## 2 Barcelona    arima_at2    <ARIMA(1,0,1)(2,1,0)[7]>
## 3 Barcelona    Snaive      <SNAIVE>

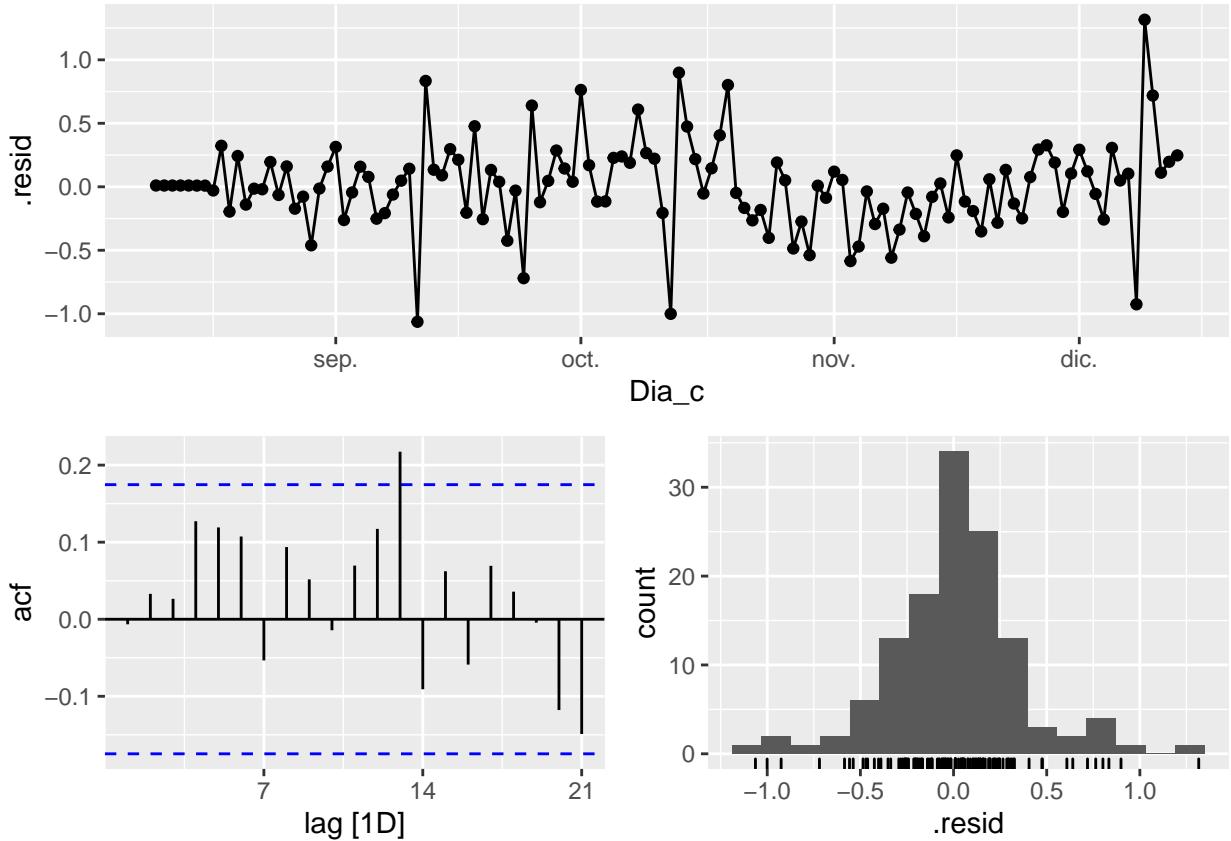
glance(fit_model) %>% arrange(AICc) %>% select(.model:BIC)

## # A tibble: 3 x 6
##   .model   sigma2 log_lik   AIC   AICc   BIC
##   <chr>     <dbl>  <dbl> <dbl> <dbl> <dbl>
## 1 arima_at1  0.132   -48.3 107. 107. 120.
## 2 arima_at2  0.132   -48.3 107. 107. 120.
## 3 Snaive     0.505     NA     NA     NA     NA

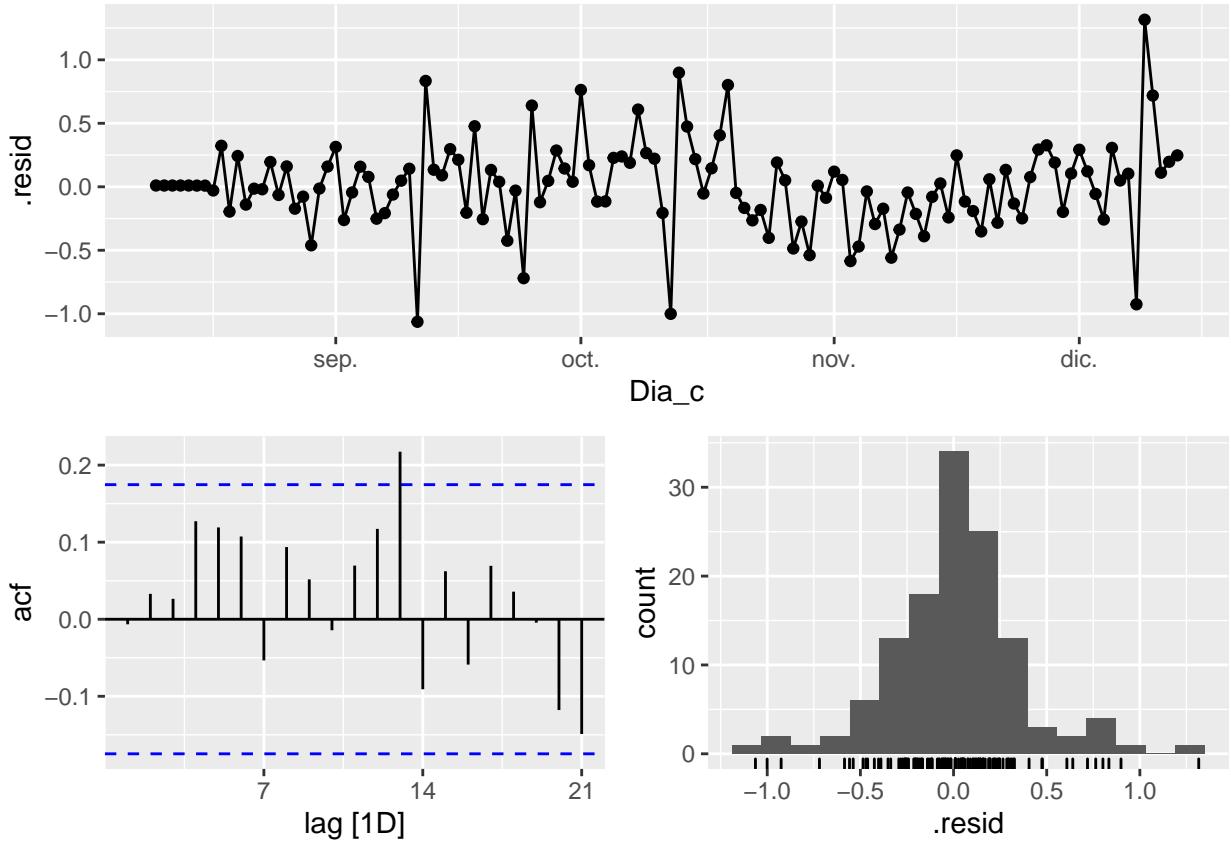
# We use a Ljung-Box test >> large p-value, confirms residuals are similar / considered to white noise
fit_model %>% select(Snaive) %>% gg_tsresiduals()

```





```
fit_model %>% select(arima_at2) %>% gg_tsresiduals()
```



```

augment(fit_model) %>%
  features(.innov, ljung_box, lag=7)

## # A tibble: 3 x 4
##   sub_region_2 .model    lb_stat lb_pvalue
##   <chr>        <chr>      <dbl>    <dbl>
## 1 Barcelona    arima_at1  6.20    0.516
## 2 Barcelona    arima_at2  6.20    0.516
## 3 Barcelona    Snaive     342.     0

augment(fit_model) %>%
  features(.innov, ljung_box, lag=14)

## # A tibble: 3 x 4
##   sub_region_2 .model    lb_stat lb_pvalue
##   <chr>        <chr>      <dbl>    <dbl>
## 1 Barcelona    arima_at1  18.4    0.191
## 2 Barcelona    arima_at2  18.4    0.191
## 3 Barcelona    Snaive     472.     0

augment(fit_model) %>%
  features(.innov, ljung_box, lag=21)

## # A tibble: 3 x 4
##   sub_region_2 .model    lb_stat lb_pvalue
##   <chr>        <chr>      <dbl>    <dbl>
## 1 Barcelona    arima_at1  25.9    0.212
## 2 Barcelona    arima_at2  25.9    0.212

```

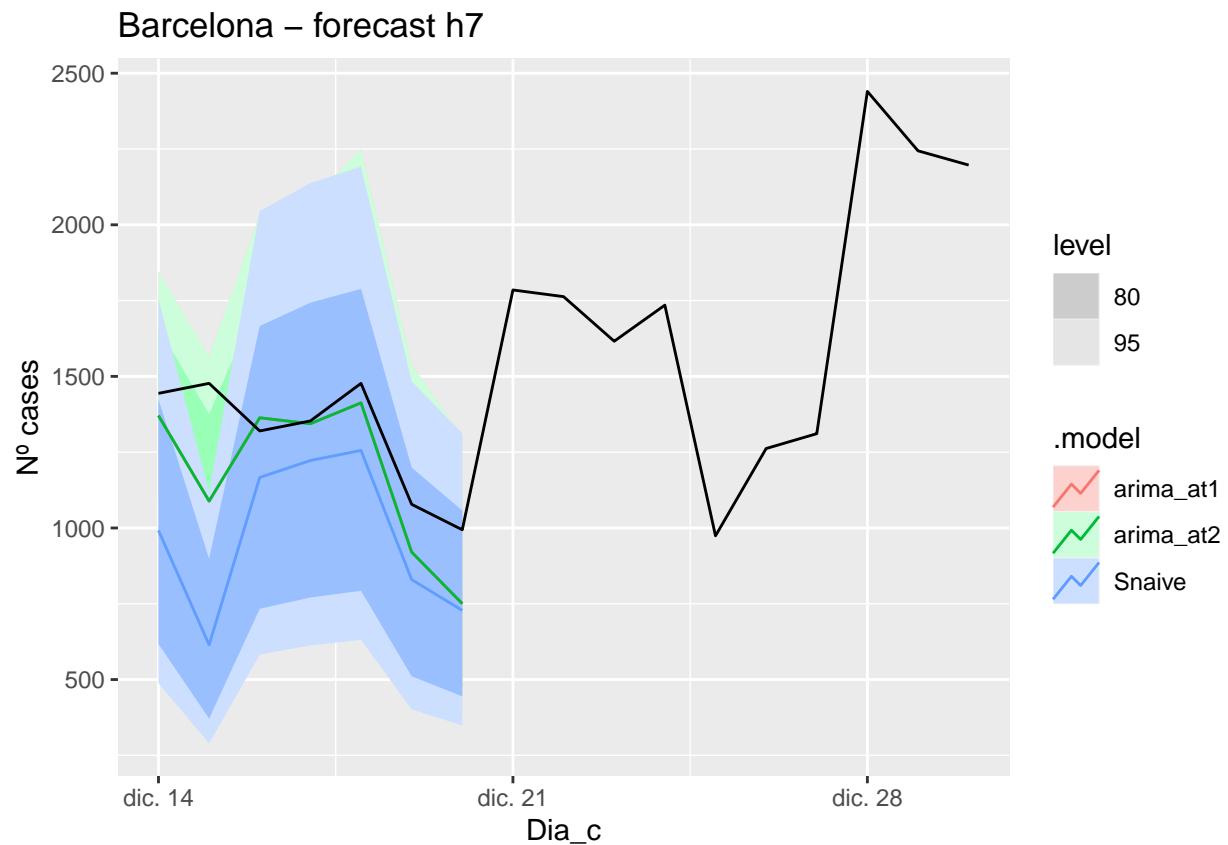
```

## 3 Barcelona      Snaive      488.      0
# Significant spikes out of 30 is still consistent with white noise
# To be sure, use a Ljung-Box test, which has a large p-value, confirming that
# the residuals are similar to white noise.
# Note that the alternative models also passed this test.

# Forecast
fc_h7<-fabletools::forecast(fit_model, h=7)
fc_h14<-fabletools::forecast(fit_model, h=14)
fc_h17<-fabletools::forecast(fit_model, h=17)

# Plots
fc_h7 %>%
  autoplot(Bar_N_cases_tt) +
  labs(y = "Nº cases", title = "Barcelona - forecast h7")

```

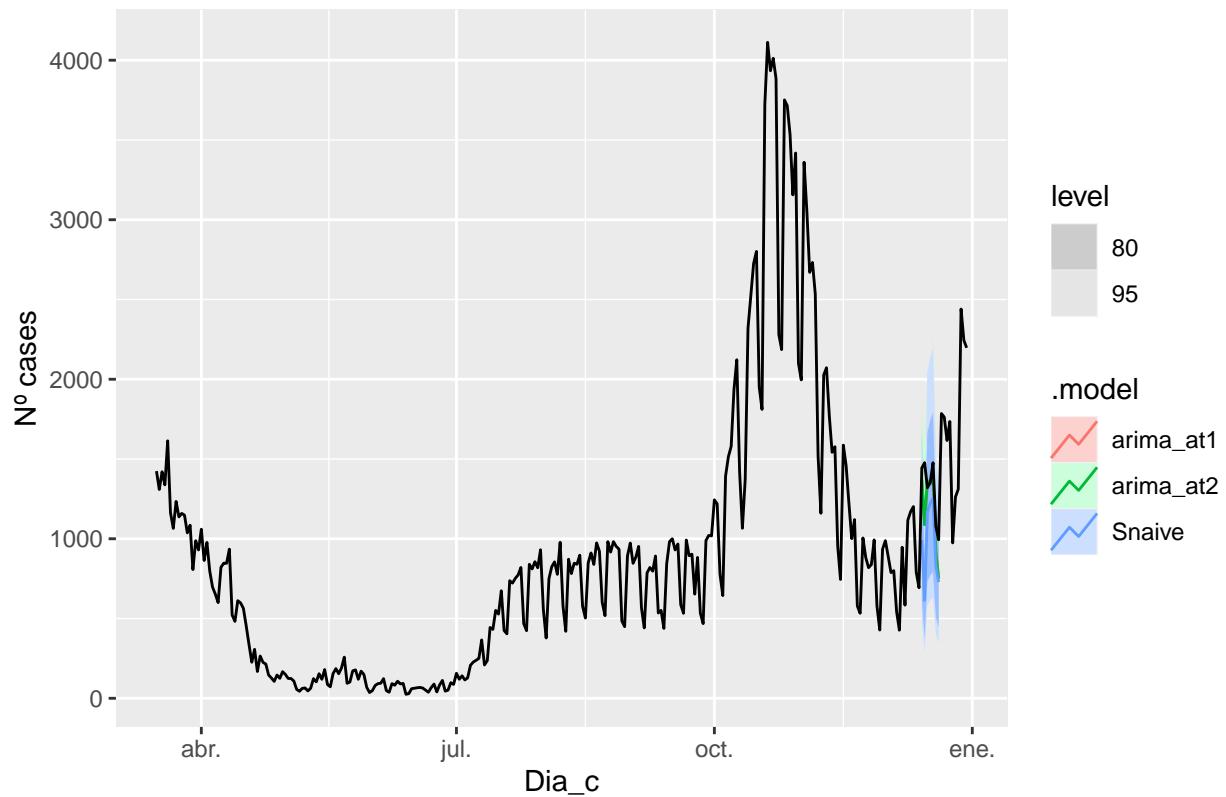


```

fc_h7 %>%
  autoplot(Bar_N_cases) +
  labs(y = "Nº cases", title = "Barcelona - forecast h7")

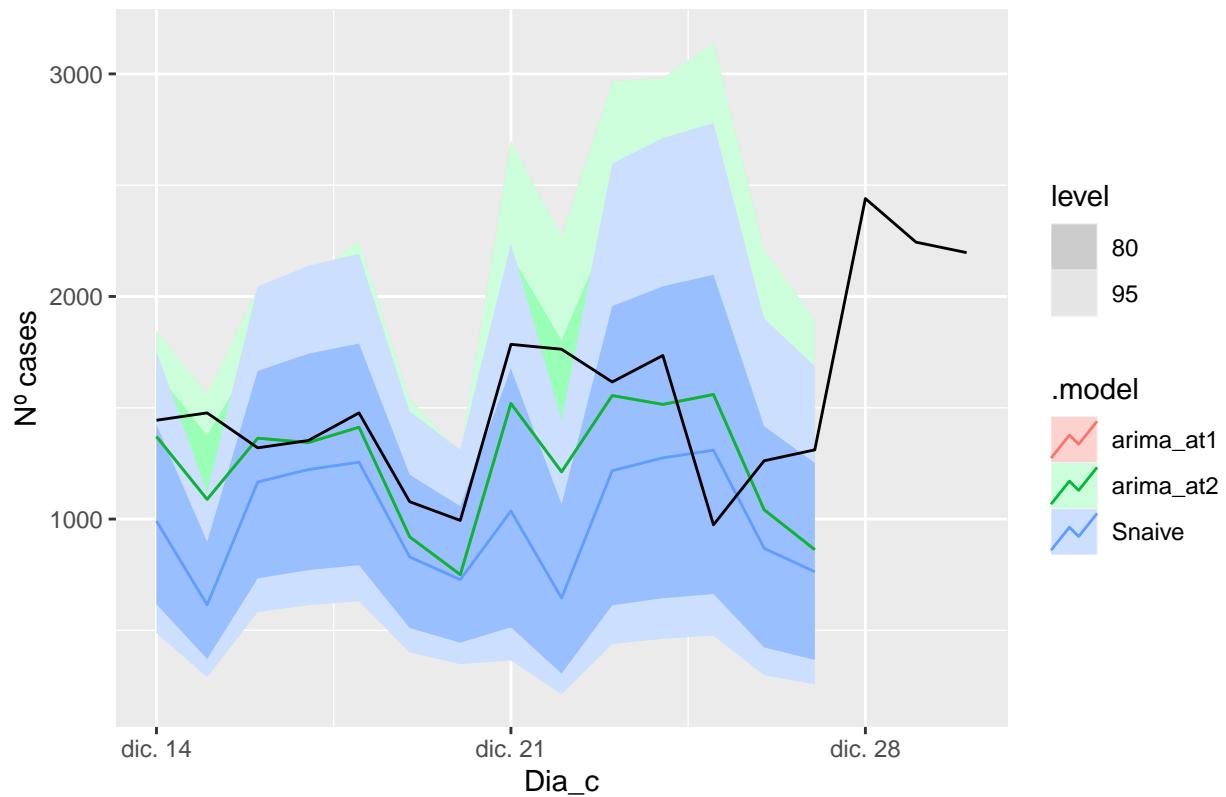
```

Barcelona – forecast h7



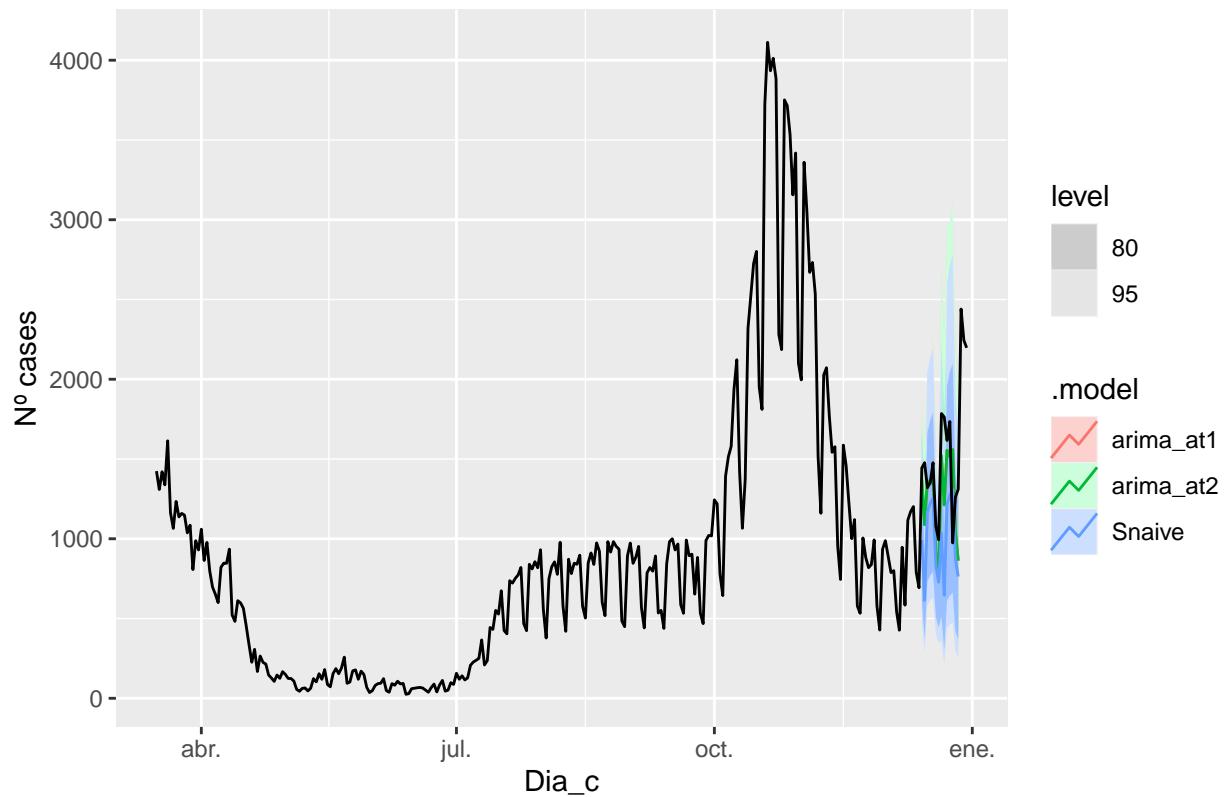
```
fc_h14 %>%
  autoplot(Bar_N_cases_tt) +
  labs(y = "Nº cases", title = "Barcelona - forecast h14")
```

Barcelona – forecast h14



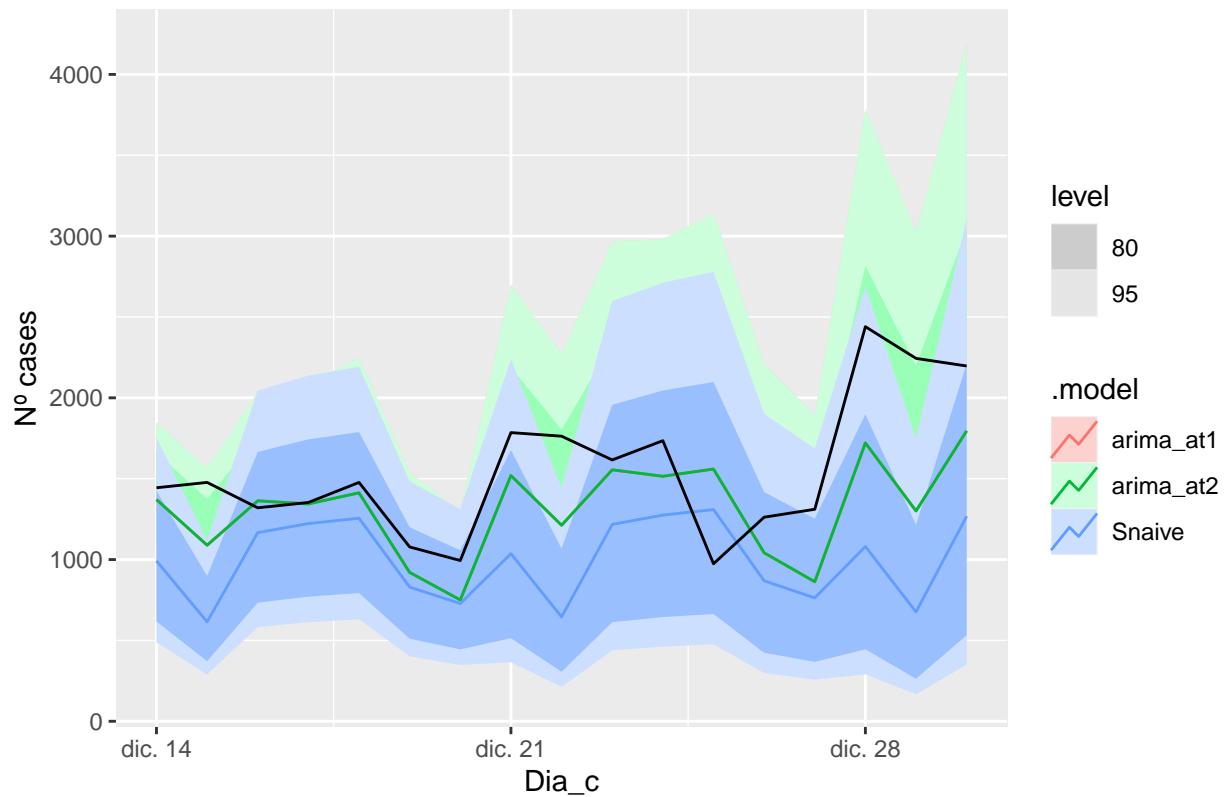
```
fc_h14 %>%
  autoplot(Bar_N_cases) +
  labs(y = "Nº cases", title = "Barcelona - forecast h14")
```

Barcelona – forecast h14



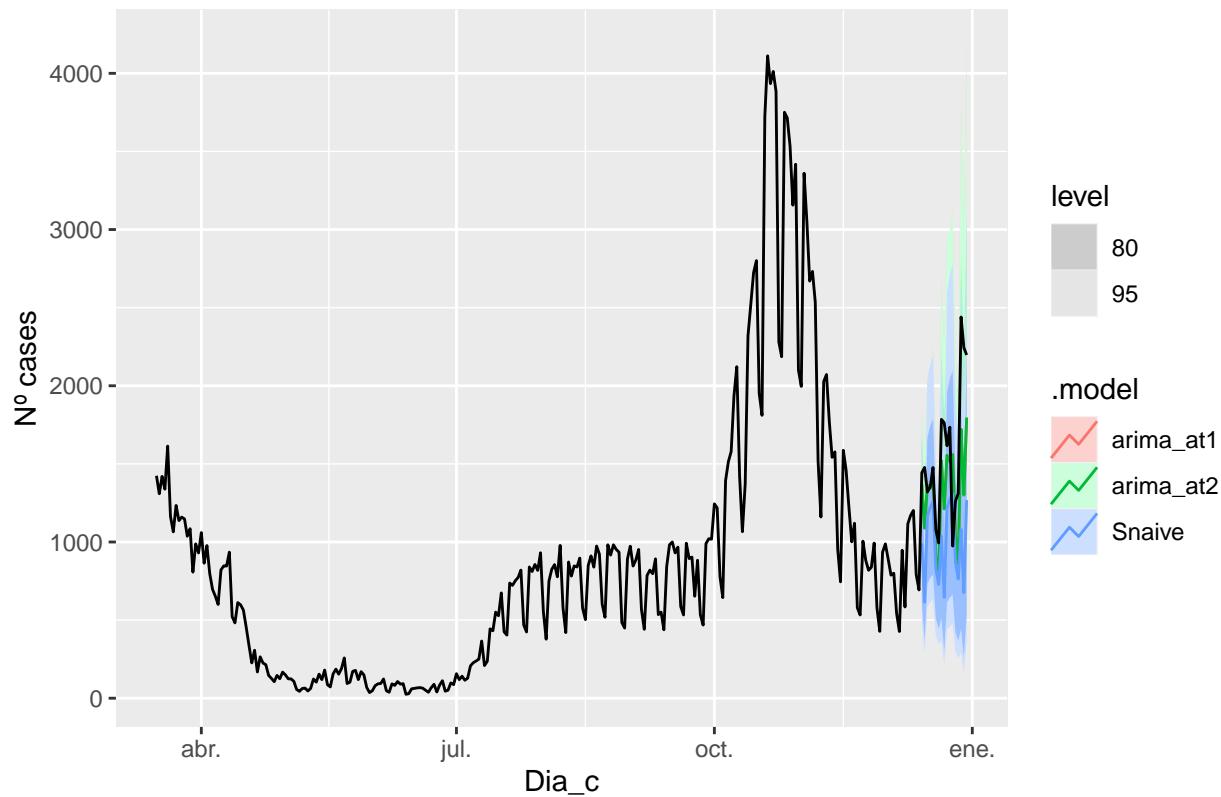
```
fc_h17 %>%
  autoplot(Bar_N_cases_tt) +
  labs(y = "Nº cases", title = "Barcelona - forecast h17")
```

Barcelona – forecast h17



```
fc_h17 %>%
  autoplot(Bar_N_cases) +
  labs(y = "Nº cases", title = "Barcelona - forecast h17")
```

Barcelona – forecast h17



```
# Accuracy
fabletools::accuracy(fc_h7, Bar_N_cases)

## # A tibble: 3 x 11
##   .model    sub_region_2 .type     ME   RMSE    MAE    MPE    MAPE    MASE   RMSSE   ACF1
##   <chr>      <chr>     <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 arima_at1 Barcelona   Test    128.  188.  140.  10.3  11.3  0.617  0.489 -0.222
## 2 arima_at2 Barcelona   Test    128.  188.  140.  10.3  11.3  0.617  0.489 -0.222
## 3 Snaive      Barcelona   Test    333.  409.  333.  25.1  25.1  1.47   1.07   0.108

#fabletools::accuracy(fc_h7, Bar_N_cases_tt)
fabletools::accuracy(fc_h14, Bar_N_cases)

## # A tibble: 3 x 11
##   .model    sub_region_2 .type     ME   RMSE    MAE    MPE    MAPE    MASE   RMSSE   ACF1
##   <chr>      <chr>     <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 arima_at1 Barcelona   Test    148.  301.  238.  9.02  18.1  1.05   0.783 -0.0965
## 2 arima_at2 Barcelona   Test    148.  301.  238.  9.02  18.1  1.05   0.783 -0.0965
## 3 Snaive      Barcelona   Test    405.  529.  453.  26.5  31.4  2.00   1.38   0.145

fabletools::accuracy(fc_h17, Bar_N_cases)

## # A tibble: 3 x 11
##   .model    sub_region_2 .type     ME   RMSE    MAE    MPE    MAPE    MASE   RMSSE   ACF1
##   <chr>      <chr>     <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 arima_at1 Barcelona   Test    243.  408.  317.  12.7  20.2  1.40   1.06   0.322
## 2 arima_at2 Barcelona   Test    243.  408.  317.  12.7  20.2  1.40   1.06   0.322
## 3 Snaive      Barcelona   Test    560.  731.  600.  31.7  35.7  2.64   1.91   0.465
```

```

# Train and test data set creation
# Train
Mad_N_cases_tr <- Mad_N_cases %>%
  filter_index("2020-08-10" ~ "2020-12-13")
# Test
Mad_N_cases_tt <- Mad_N_cases %>%
  filter_index("2020-12-14" ~ "2020-12-31")

# Modeling over train
Mad_N_cases_tr %>%
  model(#arima_nm1=ARIMA(season_adjust ~ pdq(1,1,2) + PDQ(1,0,1)),
        arima_at1=ARIMA(box_cox(num_casos.x,lambda_mad)),
        arima_at2=ARIMA(box_cox(num_casos.x,lambda_mad),
                        stepwise = FALSE, approx = FALSE),
        Snaive=SNAIVE(box_cox(num_casos.x,lambda_mad))) -> fit_model

# Show and report model
fit_model

```

3.3.1.2 Madrid

```

## # A mable: 1 x 4
## # Key:      sub_region_2 [1]
##   sub_region_2           arima_at1           arima_at2     Snaive
##   <chr>                 <model>             <model>    <model>
## 1 Madrid    <ARIMA(0,1,2)(2,1,2)[7]> <ARIMA(0,1,2)(2,1,2)[7]> <SNAIVE>
report(fit_model)

## # A tibble: 3 x 9
##   sub_region_2 .model    sigma2 log_lik   AIC   AICc   BIC ar_roots ma_roots
##   <chr>       <chr>     <dbl>   <dbl> <dbl> <dbl> <dbl> <list>   <list>
## 1 Madrid      arima_at1 0.0324   33.7 -53.5 -52.5 -34.1 <cpl [14]> <cpl [16]>
## 2 Madrid      arima_at2 0.0324   33.7 -53.5 -52.5 -34.1 <cpl [14]> <cpl [16]>
## 3 Madrid      Snaive    0.0971    NA     NA     NA    <NULL> <NULL>

# Good model >> Less Sigma / AICc
fit_model %>% pivot_longer(!sub_region_2,
                           names_to = "Model name",
                           values_to = "Orders")

## # A mable: 3 x 3
## # Key:      sub_region_2, Model name [3]
##   sub_region_2 `Model name`          Orders
##   <chr>       <chr>                  <model>
## 1 Madrid      arima_at1    <ARIMA(0,1,2)(2,1,2)[7]>
## 2 Madrid      arima_at2    <ARIMA(0,1,2)(2,1,2)[7]>
## 3 Madrid      Snaive      <SNAIVE>

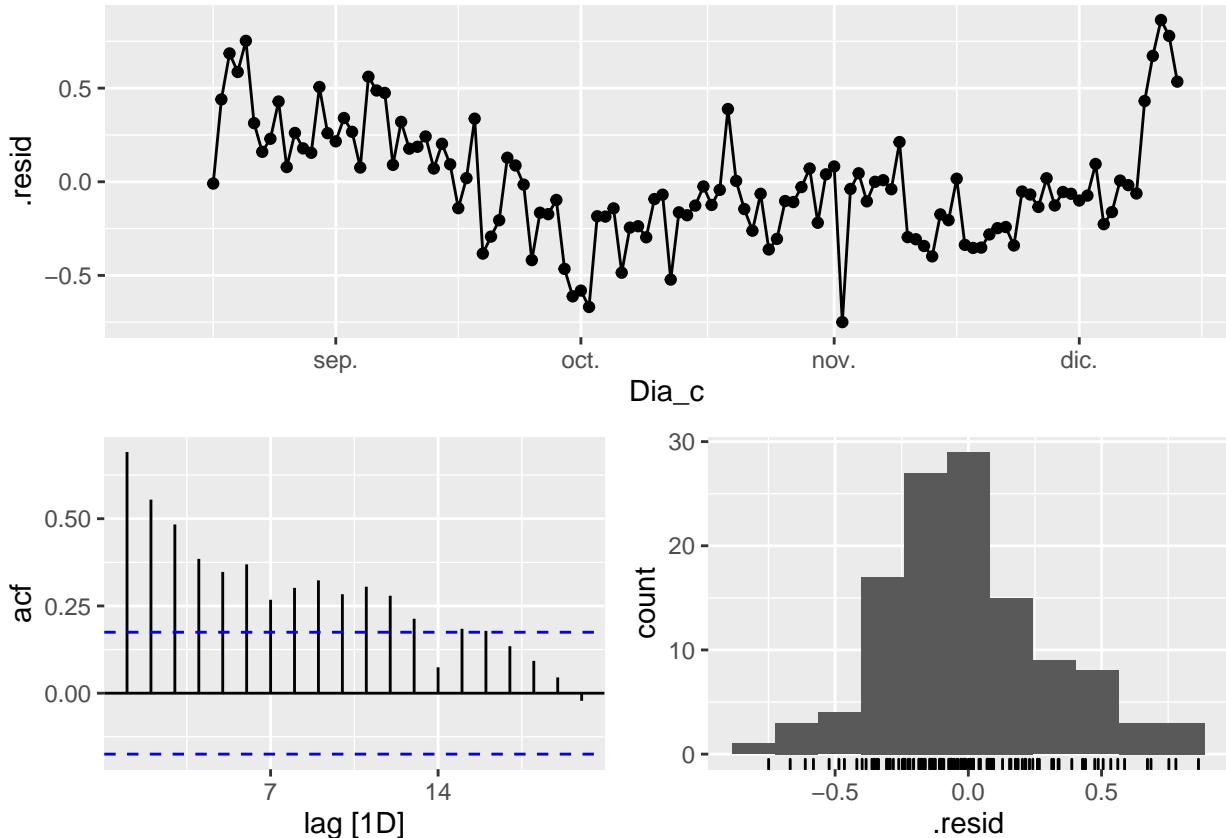
glance(fit_model) %>% arrange(AICc) %>% select(.model:BIC)

## # A tibble: 3 x 6
##   .model    sigma2 log_lik   AIC   AICc   BIC
##   <chr>     <dbl>   <dbl> <dbl> <dbl> <dbl>
## 1 arima_at1 0.0324   33.7 -53.5 -52.5 -34.1

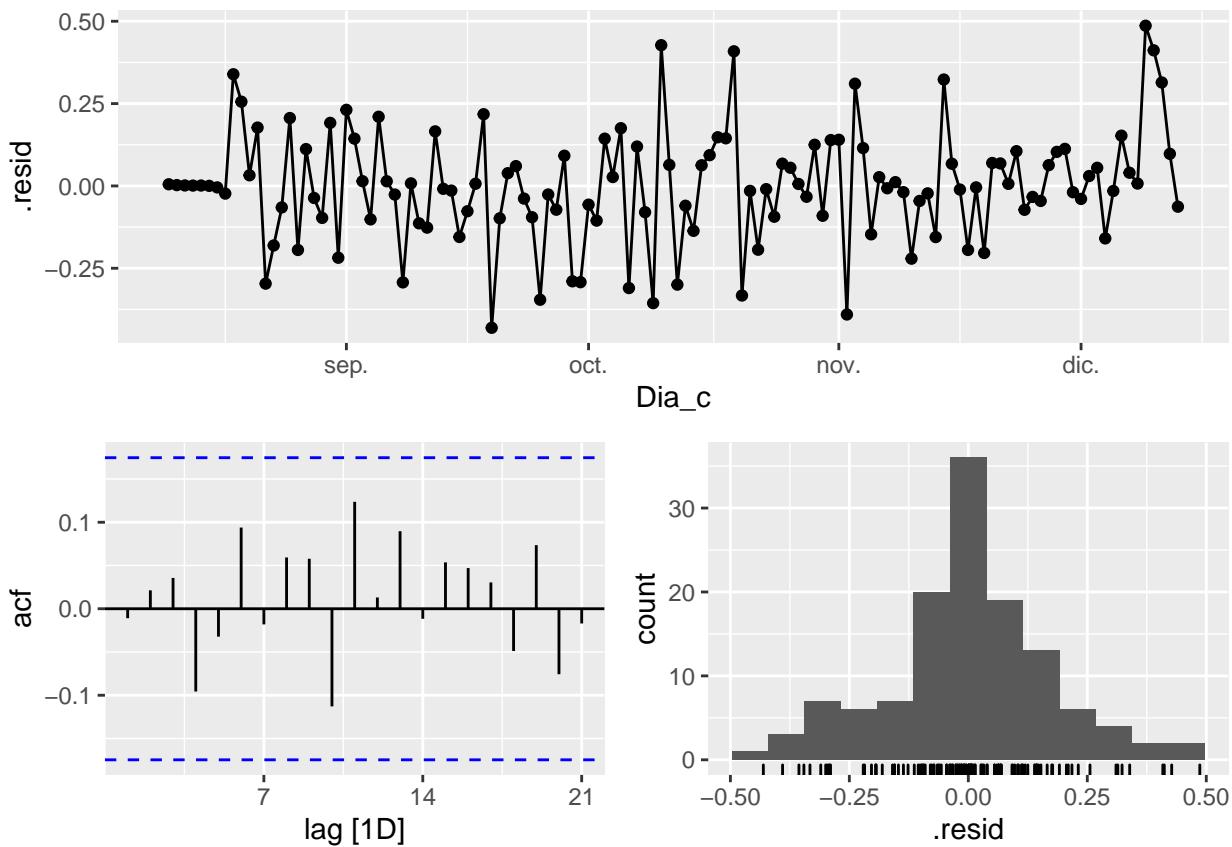
```

```
## 2 arima_at2 0.0324      33.7 -53.5 -52.5 -34.1
## 3 Snaive     0.0971      NA     NA     NA     NA
```

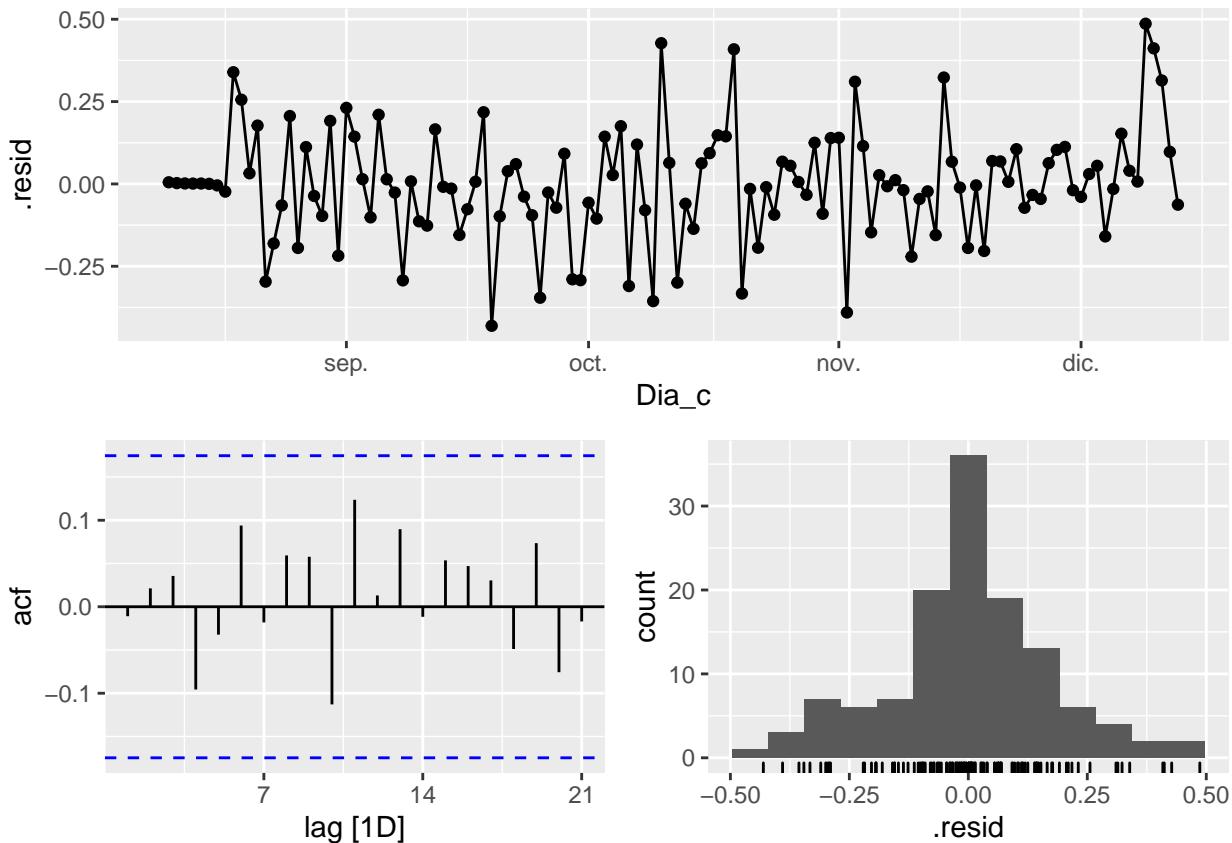
```
# We use a Ljung-Box test >> large p-value, confirms residuals are similar / considered to white noise
fit_model %>% select(Snaive) %>% gg_tsresiduals()
```



```
fit_model %>% select(arima_at1) %>% gg_tsresiduals()
```



```
fit_model %>% select(arima_at2) %>% gg_tsresiduals()
```



```

augment(fit_model) %>%
  features(.innov, ljung_box, lag=7)

## # A tibble: 3 x 4
##   sub_region_2 .model    lb_stat lb_pvalue
##   <chr>        <chr>      <dbl>    <dbl>
## 1 Madrid       arima_at1  2.82     0.901
## 2 Madrid       arima_at2  2.82     0.901
## 3 Madrid       Snaive     186.      0

augment(fit_model) %>%
  features(.innov, ljung_box, lag=14)

## # A tibble: 3 x 4
##   sub_region_2 .model    lb_stat lb_pvalue
##   <chr>        <chr>      <dbl>    <dbl>
## 1 Madrid       arima_at1  8.86     0.840
## 2 Madrid       arima_at2  8.86     0.840
## 3 Madrid       Snaive     251.     0

augment(fit_model) %>%
  features(.innov, ljung_box, lag=21)

## # A tibble: 3 x 4
##   sub_region_2 .model    lb_stat lb_pvalue
##   <chr>        <chr>      <dbl>    <dbl>
## 1 Madrid       arima_at1  11.8     0.944
## 2 Madrid       arima_at2  11.8     0.944

```

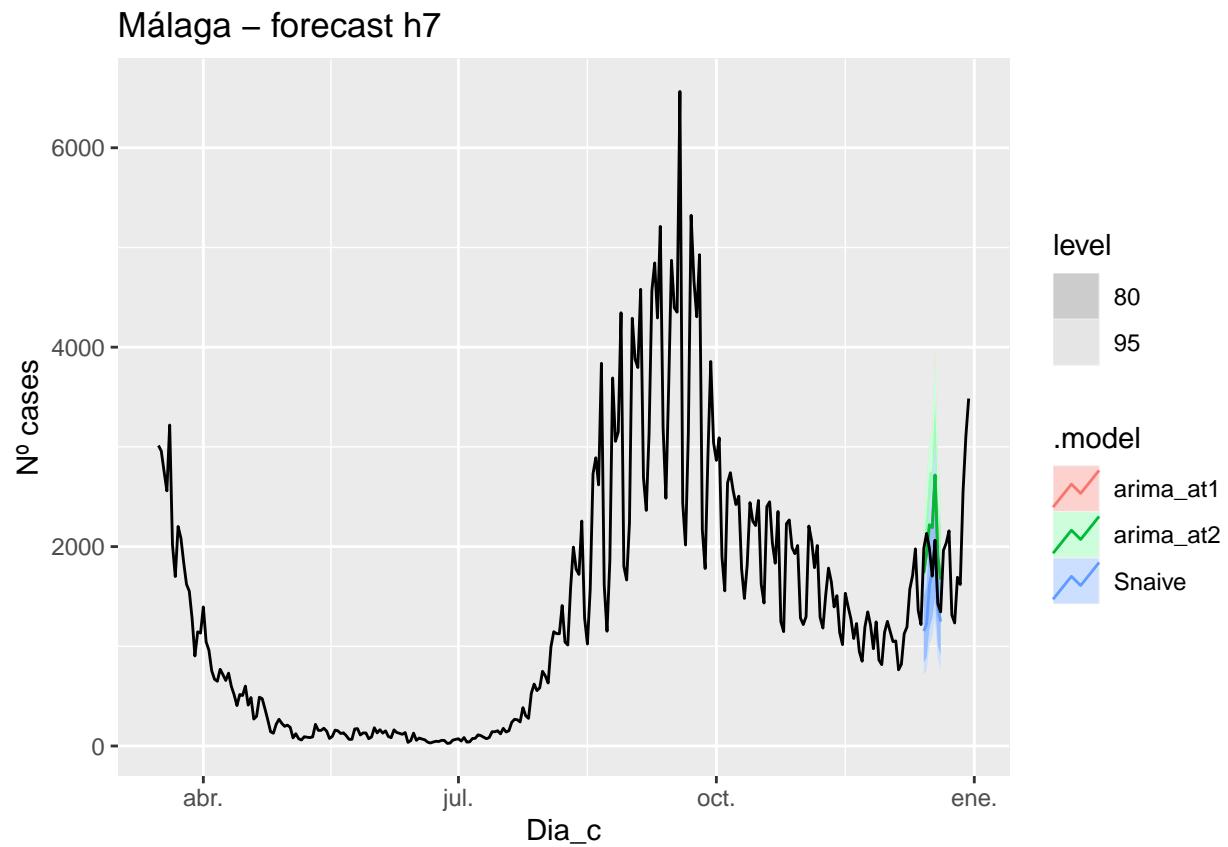
```

## 3 Madrid      Snaive     265.      0
# Significant spikes out of 30 is still consistent with white noise
# To be sure, use a Ljung-Box test, which has a large p-value, confirming that
# the residuals are similar to white noise.
# Note that the alternative models also passed this test.

# Forecast
fc_h7<-fabletools::forecast(fit_model, h=7)
fc_h14<-fabletools::forecast(fit_model, h=14)
fc_h17<-fabletools::forecast(fit_model, h=17)

# Plots
fc_h7 %>%
  autoplot(Mad_N_cases) +
  labs(y = "Nº cases", title = "Málaga - forecast h7")

```

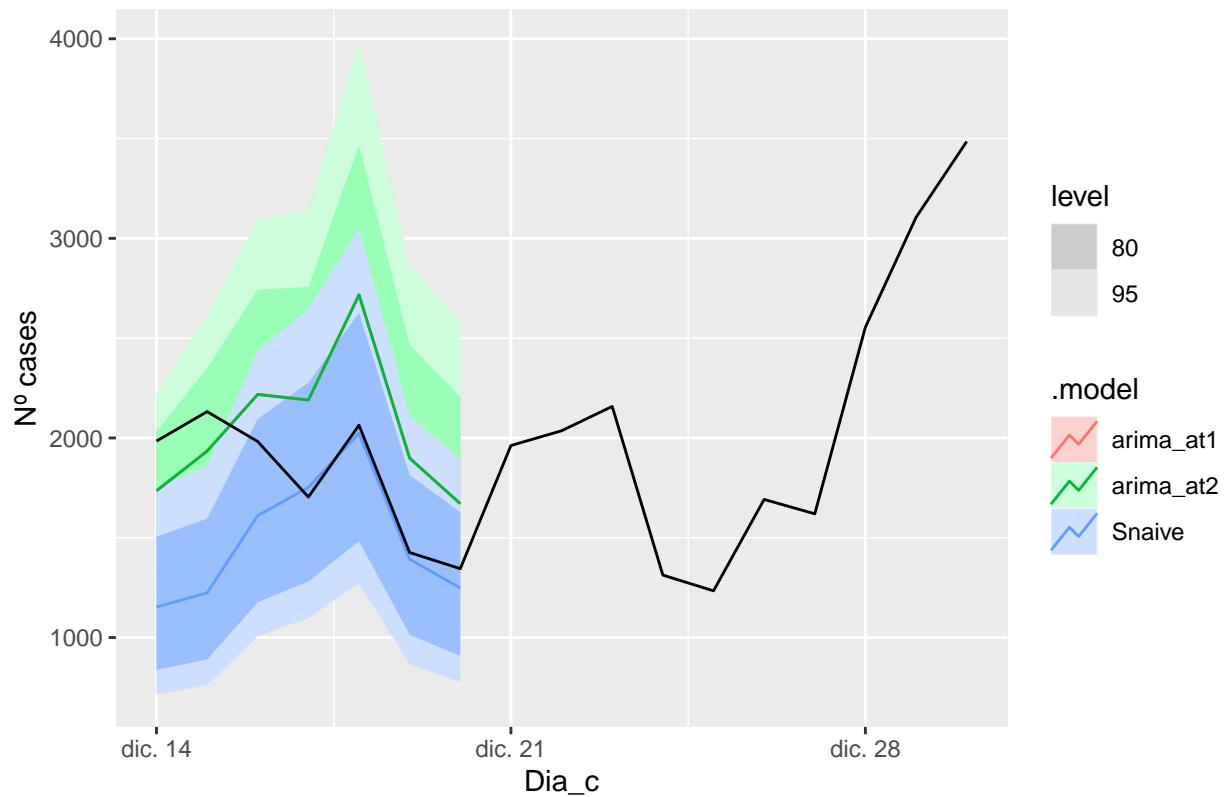


```

fc_h7 %>%
  autoplot(Mad_N_cases_tt) +
  labs(y = "Nº cases", title = "Málaga - forecast h7")

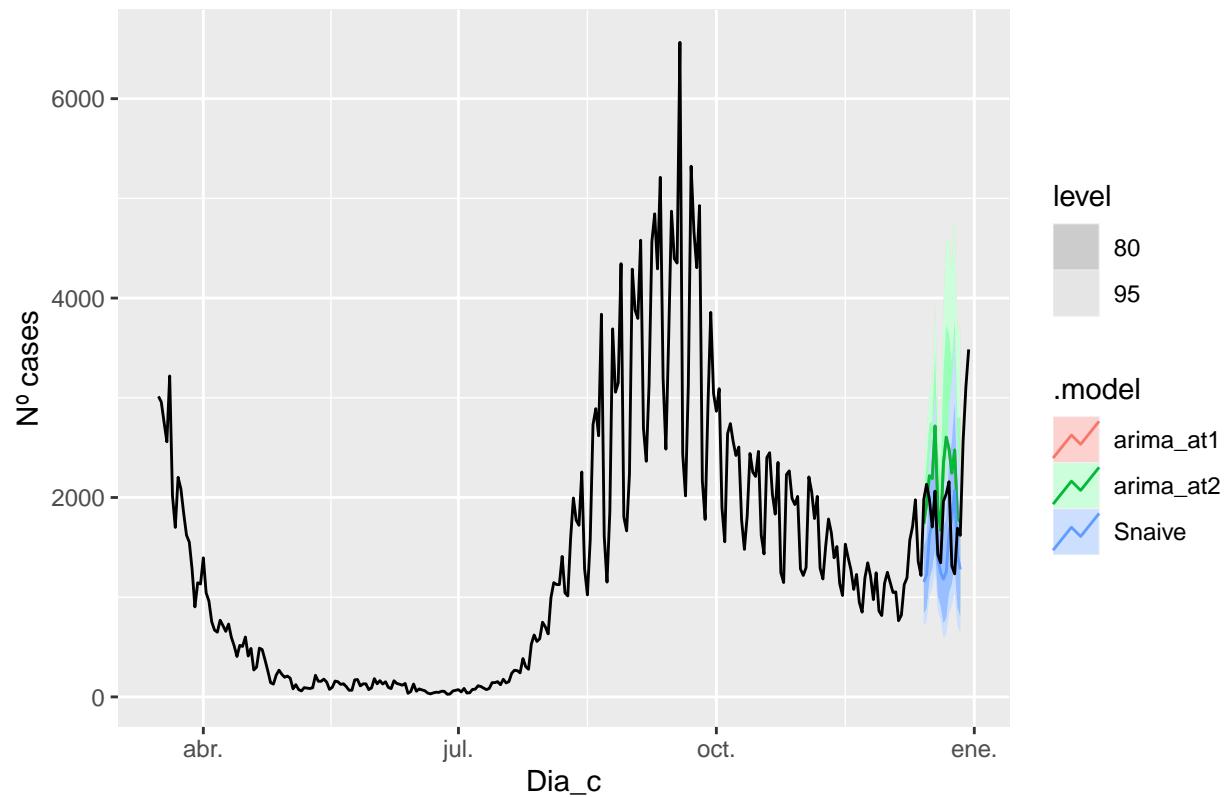
```

Málaga – forecast h7



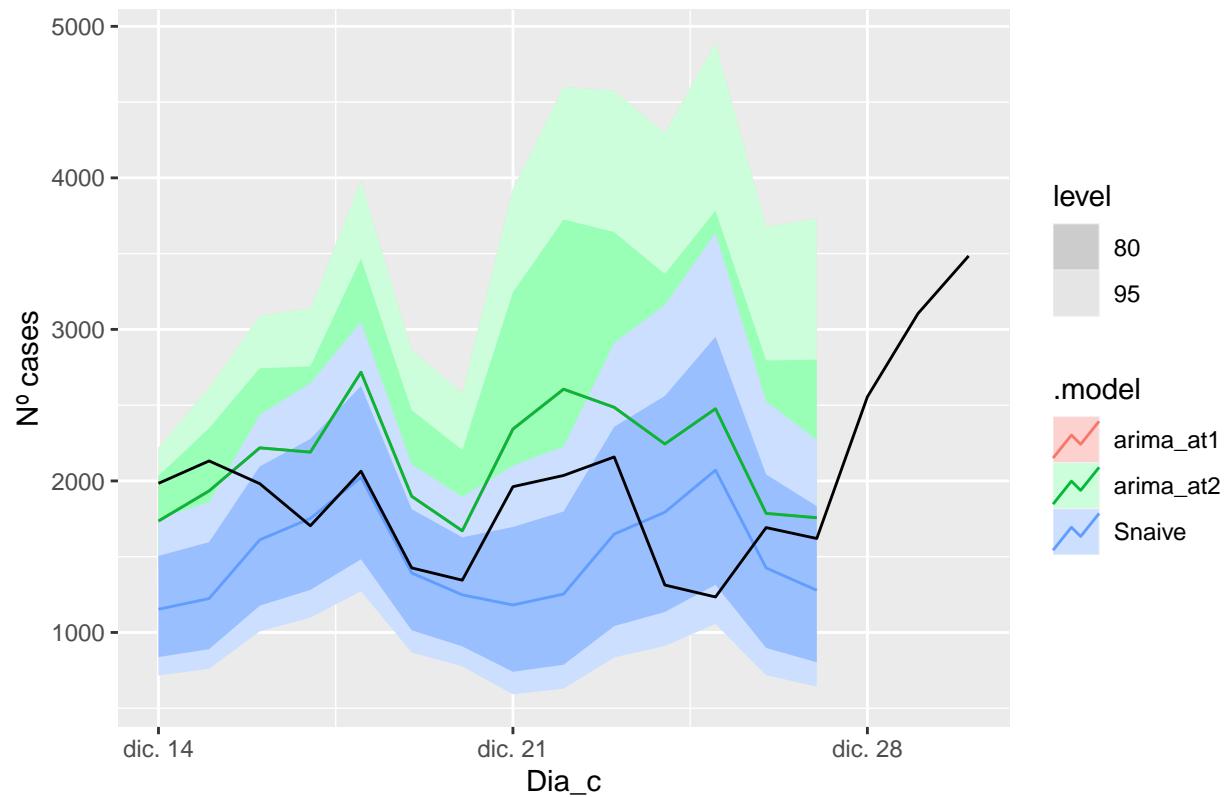
```
fc_h14 %>%
  autoplot(Mad_N_cases) +
  labs(y = "Nº cases", title = "Málaga - forecast h14")
```

Málaga – forecast h14



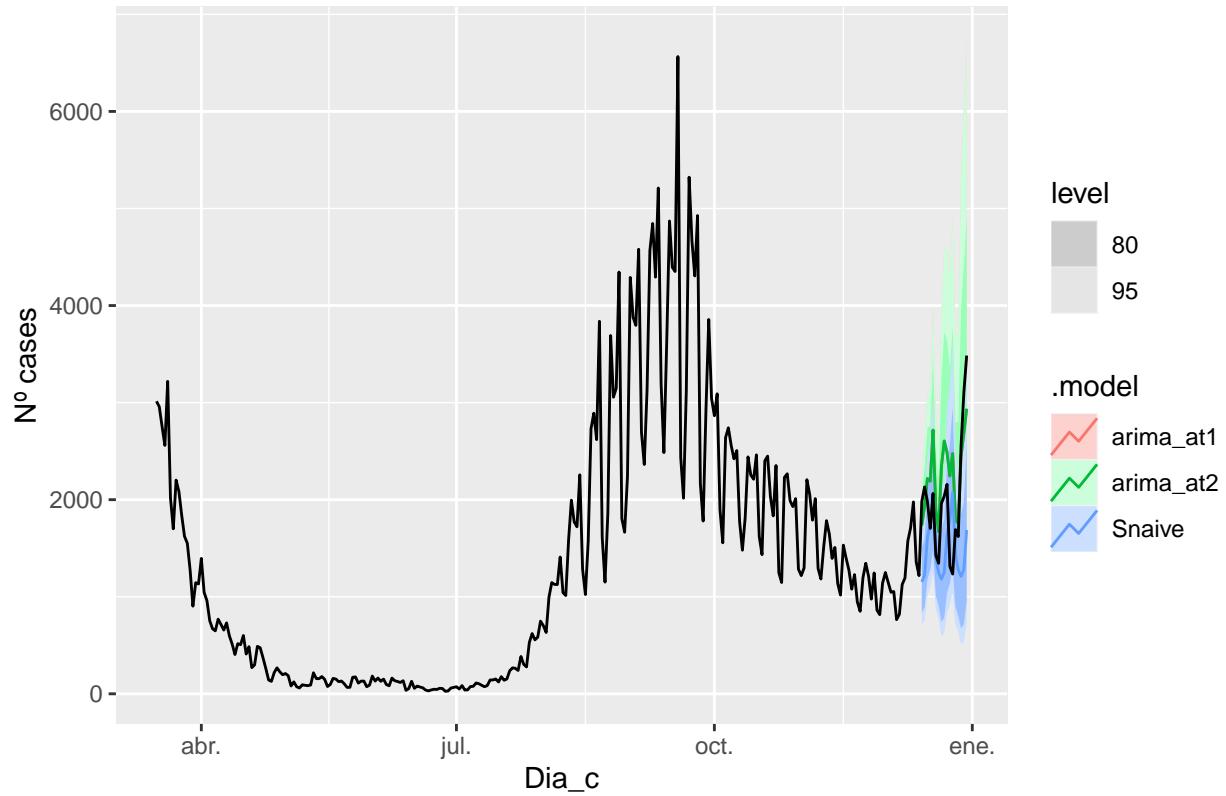
```
fc_h14 %>%
  autoplot(Mad_N_cases_tt) +
  labs(y = "Nº cases", title = "Málaga - forecast h14")
```

Málaga – forecast h14



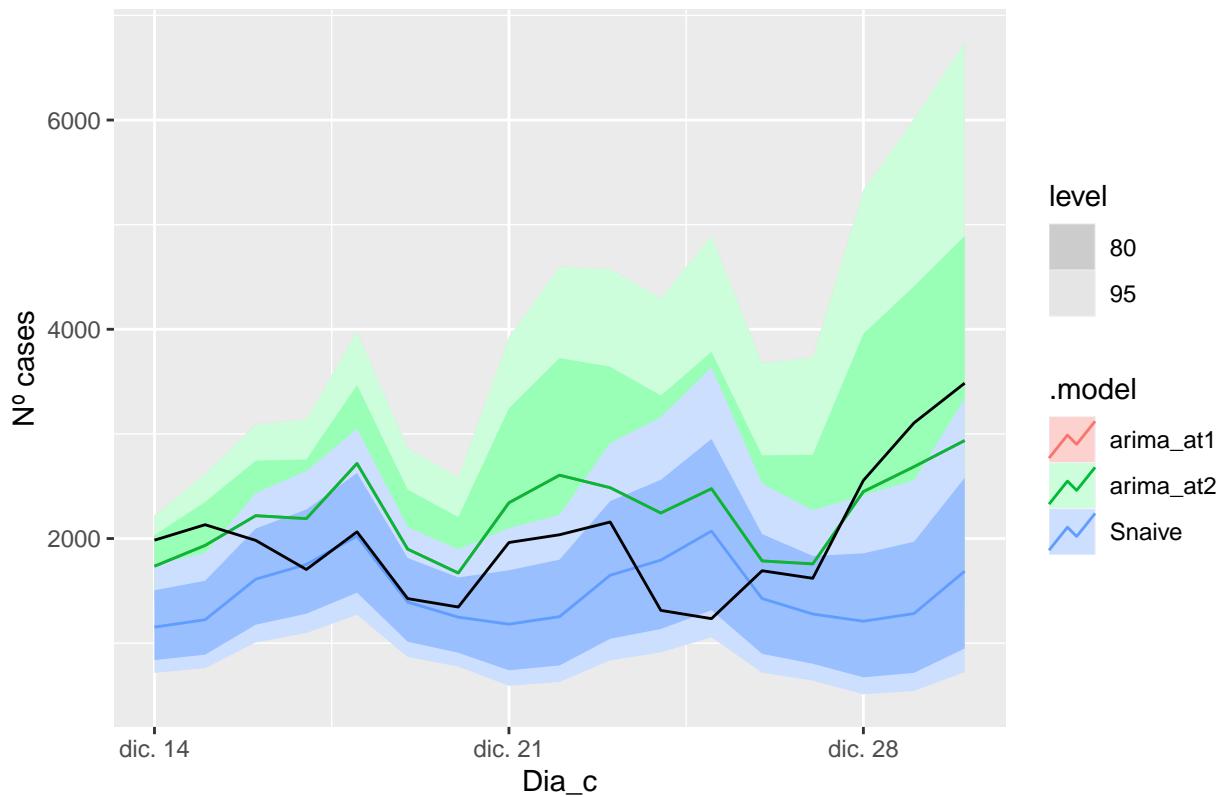
```
fc_h17 %>%
  autoplot(Mad_N_cases) +
  labs(y = "Nº cases", title = "Málaga - forecast h17")
```

Málaga – forecast h17



```
fc_h17 %>%
  autoplot(Mad_N_cases_tt) +
  labs(y = "Nº cases", title = "Málaga - forecast h17")
```

Málaga – forecast h17



```
# Accuracy
fabletools::accuracy(fc_h7, Mad_N_cases)

## # A tibble: 3 x 11
##   .model    sub_region_2 .type     ME   RMSE    MAE    MPE    MAPE    MASE   RMSSE   ACF1
##   <chr>      <chr>     <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 arima_at1 Madrid     Test  -247.  405.  374. -15.4  21.6  1.27  0.884  0.594
## 2 arima_at2 Madrid     Test  -247.  405.  374. -15.4  21.6  1.27  0.884  0.594
## 3 Snaive      Madrid     Test   319.  488.  333.  16.0  16.8  1.13  1.06  0.585

fabletools::accuracy(fc_h14, Mad_N_cases)

## # A tibble: 3 x 11
##   .model    sub_region_2 .type     ME   RMSE    MAE    MPE    MAPE    MASE   RMSSE   ACF1
##   <chr>      <chr>     <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 arima_at1 Madrid     Test  -386.  545.  450. -25.4  28.5  1.52  1.19  0.354
## 2 arima_at2 Madrid     Test  -386.  545.  450. -25.4  28.5  1.52  1.19  0.354
## 3 Snaive      Madrid     Test   257.  552.  452.  10.4  25.8  1.53  1.20  0.448

fabletools::accuracy(fc_h17, Mad_N_cases)

## # A tibble: 3 x 11
##   .model    sub_region_2 .type     ME   RMSE    MAE    MPE    MAPE    MASE   RMSSE   ACF1
##   <chr>      <chr>     <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 arima_at1 Madrid     Test  -255.  523.  434. -19.0  25.5  1.47  1.14  0.536
## 2 arima_at2 Madrid     Test  -255.  523.  434. -19.0  25.5  1.47  1.14  0.536
## 3 Snaive      Madrid     Test   504.  862.  664.  18.2  30.8  2.25  1.88  0.617
```

```

# Train data set creation
# Train
Mal_N_cases_tr <- Mal_N_cases %>%
  filter_index("2020-08-10" ~ "2020-12-13")

Mal_N_cases_tt <- Mal_N_cases %>%
  filter_index("2020-12-14" ~ "2020-12-31")

# Modeling over train
Mal_N_cases_tr %>%
  model(arima_at1=ARIMA(box_cox(num_casos.x,lambda_mal)),
        arima_at2=ARIMA(box_cox(num_casos.x,lambda_mal),stepwise = FALSE, approx = FALSE),
        Snaive=SNAIVE(box_cox(num_casos.x,lambda_mal))) -> fit_model

# Show and report model
fit_model

```

3.3.1.3 Málaga

```

## # A mable: 1 x 4
## # Key:   sub_region_2 [1]
##   sub_region_2           arima_at1           arima_at2     Snaive
##   <chr>                 <model>             <model>     <model>
## 1 Málaga    <ARIMA(1,0,1)(2,1,0)[7]> <ARIMA(1,0,1)(0,1,1)[7]> <SNAIVE>
report(fit_model)

## # A tibble: 3 x 9
##   sub_region_2 .model   sigma2 logLik   AIC   AICc   BIC ar_roots ma_roots
##   <chr>       <chr>     <dbl> <dbl>   <dbl> <dbl> <dbl> <list>   <list>
## 1 Málaga      arima_at1  0.982  -167.  345.  345.  359. <cpl [15]> <cpl [1]>
## 2 Málaga      arima_at2  0.943  -166.  340.  341.  351. <cpl [1]> <cpl [8]>
## 3 Málaga      Snaive    2.10    NA     NA     NA     NA  <NULL>   <NULL>

# Good model >> Less Sigma / AICc
fit_model %>% pivot_longer(!sub_region_2,
                           names_to = "Model name",
                           values_to = "Orders")

```

```

## # A mable: 3 x 3
## # Key:   sub_region_2, Model name [3]
##   sub_region_2 `Model name`          Orders
##   <chr>       <chr>                <model>
## 1 Málaga      arima_at1    <ARIMA(1,0,1)(2,1,0)[7]>
## 2 Málaga      arima_at2    <ARIMA(1,0,1)(0,1,1)[7]>
## 3 Málaga      Snaive      <SNAIVE>

glance(fit_model) %>% arrange(AICc) %>% select(.model:BIC)

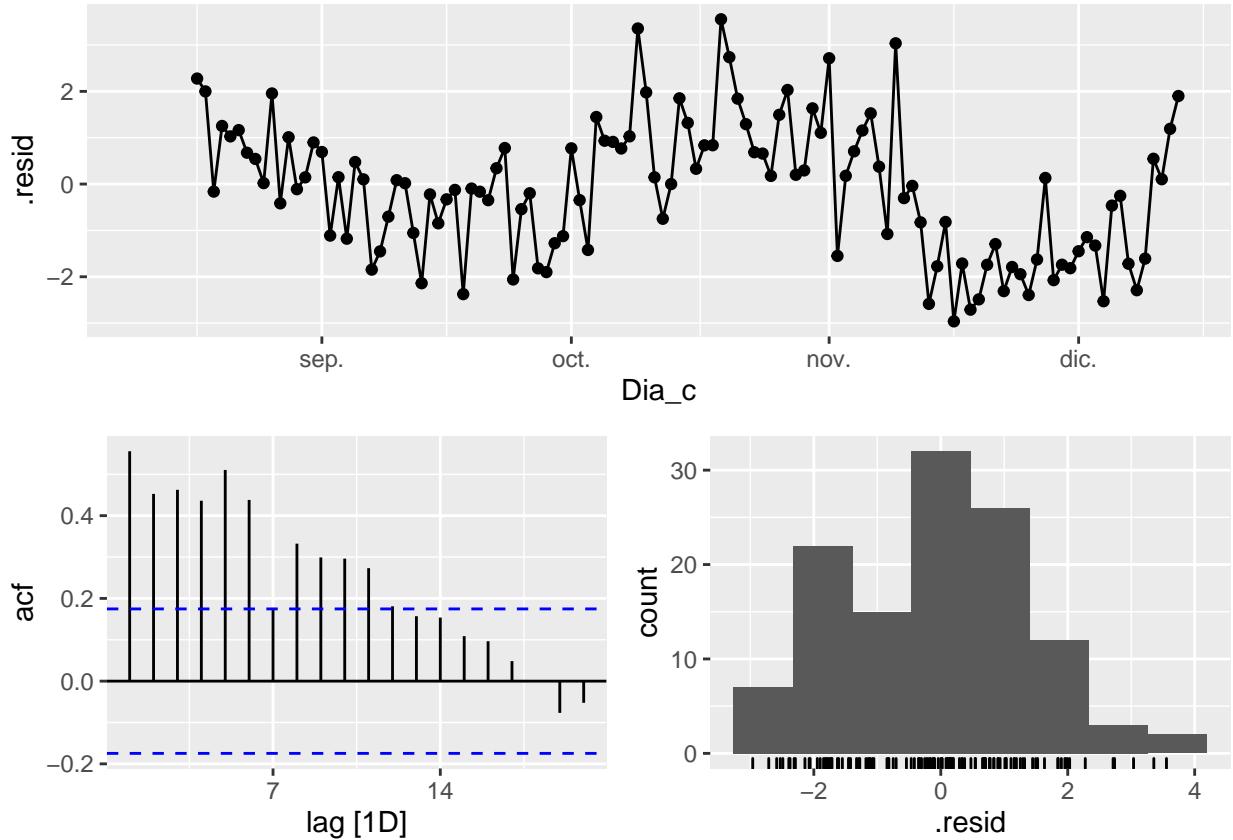
```

```

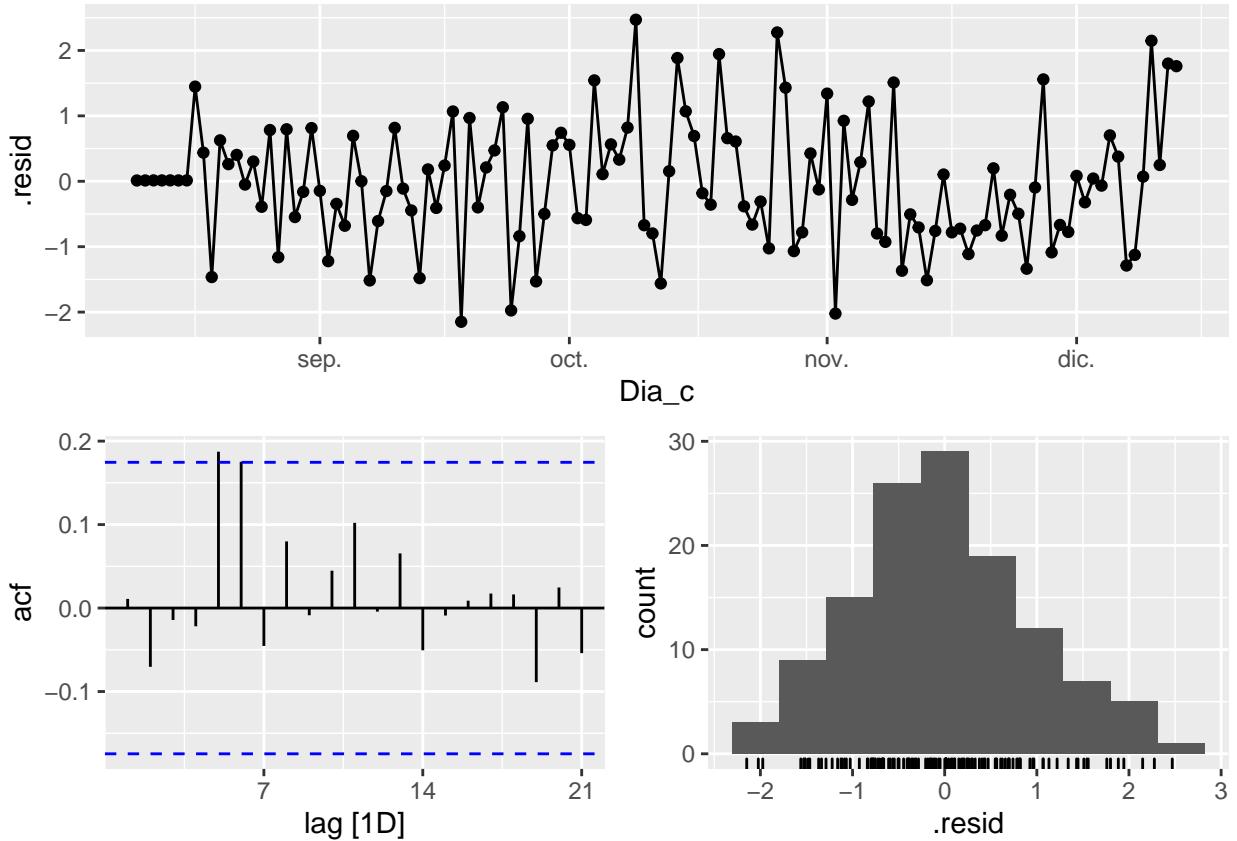
## # A tibble: 3 x 6
##   .model   sigma2 logLik   AIC   AICc   BIC
##   <chr>     <dbl> <dbl>   <dbl> <dbl> <dbl>
## 1 arima_at2  0.943  -166.  340.  341.  351.
## 2 arima_at1  0.982  -167.  345.  345.  359.
## 3 Snaive     2.10    NA     NA     NA     NA

```

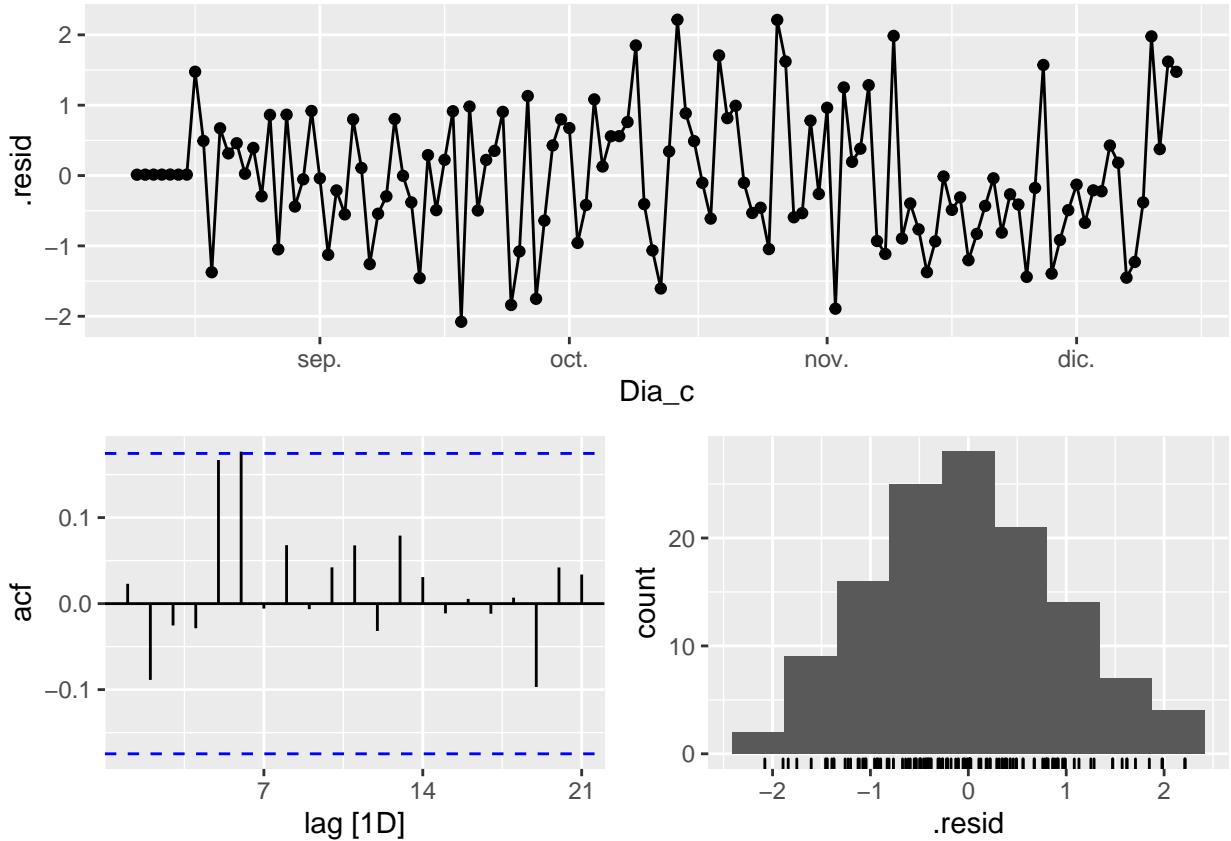
```
# We use a Ljung-Box test >> large p-value, confirms residuals are similar / considered to white noise
fit_model %>% select(Snaive) %>% gg_tsresiduals()
```



```
fit_model %>% select(arima_at1) %>% gg_tsresiduals()
```



```
fit_model %>% select(arima_at2) %>% gg_tsresiduals()
```



```

augment(fit_model) %>%
  features(.innov, ljung_box, lag=7)

## # A tibble: 3 x 4
##   sub_region_2 .model    lb_stat lb_pvalue
##   <chr>        <chr>      <dbl>     <dbl>
## 1 Málaga      arima_at1  9.83    0.198
## 2 Málaga      arima_at2  9.19    0.239
## 3 Málaga      Snaive     175.     0

augment(fit_model) %>%
  features(.innov, ljung_box, lag=14)

## # A tibble: 3 x 4
##   sub_region_2 .model    lb_stat lb_pvalue
##   <chr>        <chr>      <dbl>     <dbl>
## 1 Málaga      arima_at1  13.4    0.493
## 2 Málaga      arima_at2  11.9    0.615
## 3 Málaga      Snaive     233.     0

augment(fit_model) %>%
  features(.innov, ljung_box, lag=21)

## # A tibble: 3 x 4
##   sub_region_2 .model    lb_stat lb_pvalue
##   <chr>        <chr>      <dbl>     <dbl>
## 1 Málaga      arima_at1  15.3    0.809
## 2 Málaga      arima_at2  13.8    0.878

```

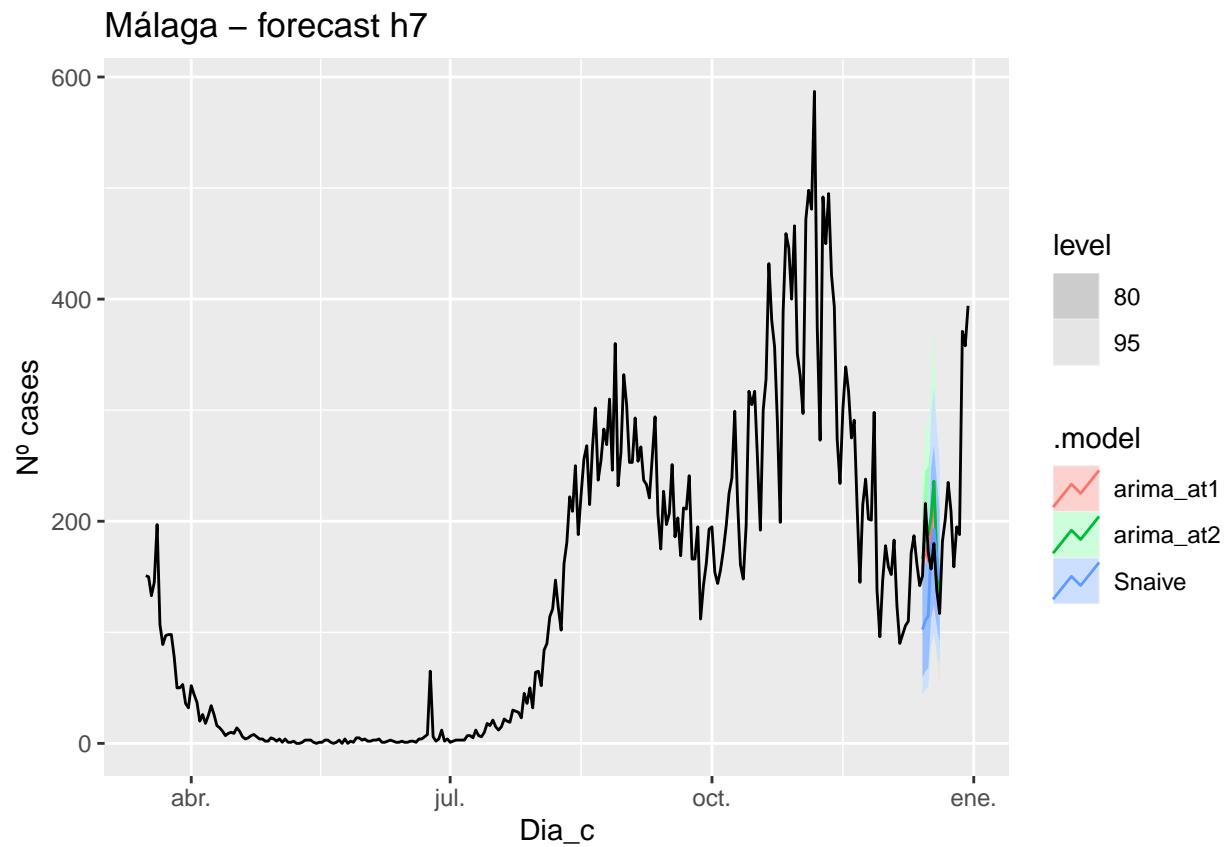
```

## 3 Málaga      Snaive     239.      0
# Significant spikes out of 30 is still consistent with white noise
# To be sure, use a Ljung-Box test, which has a large p-value, confirming that
# the residuals are similar to white noise.
# Note that the alternative models also passed this test.

# Forecast
fc_h7<-fabletools::forecast(fit_model, h=7)
fc_h14<-fabletools::forecast(fit_model, h=14)
fc_h17<-fabletools::forecast(fit_model, h=17)

# Plots
fc_h7 %>%
  autoplot(Mal_N_cases) +
  labs(y = "Nº cases", title = "Málaga - forecast h7")

```

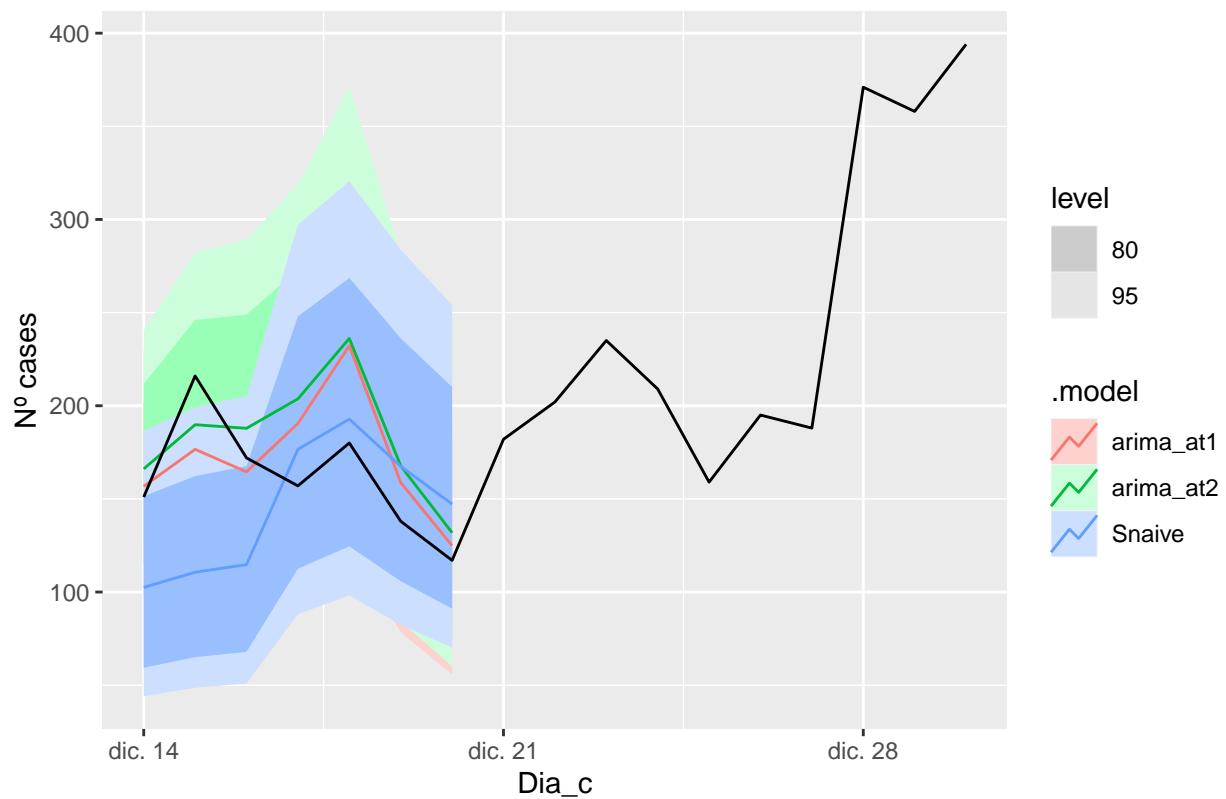


```

fc_h7 %>%
  autoplot(Mal_N_cases_tt) +
  labs(y = "Nº cases", title = "Málaga - forecast h7")

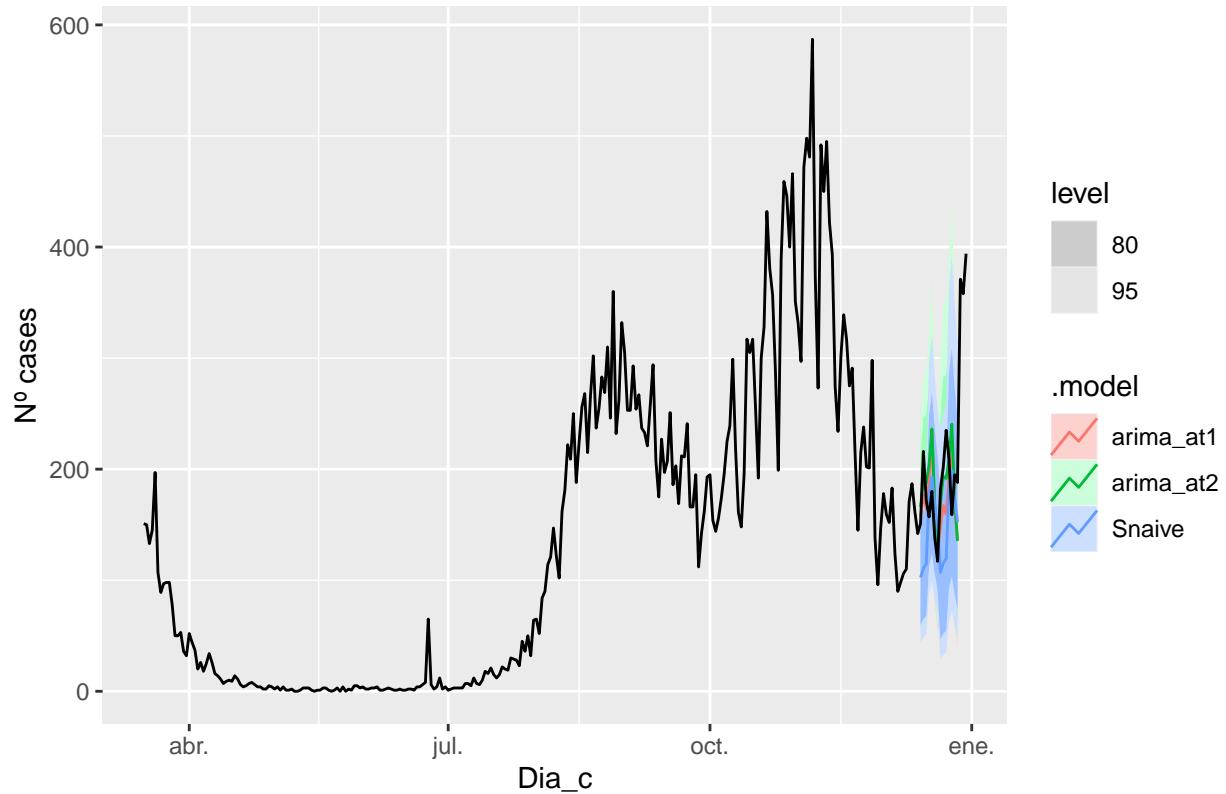
```

Málaga – forecast h7



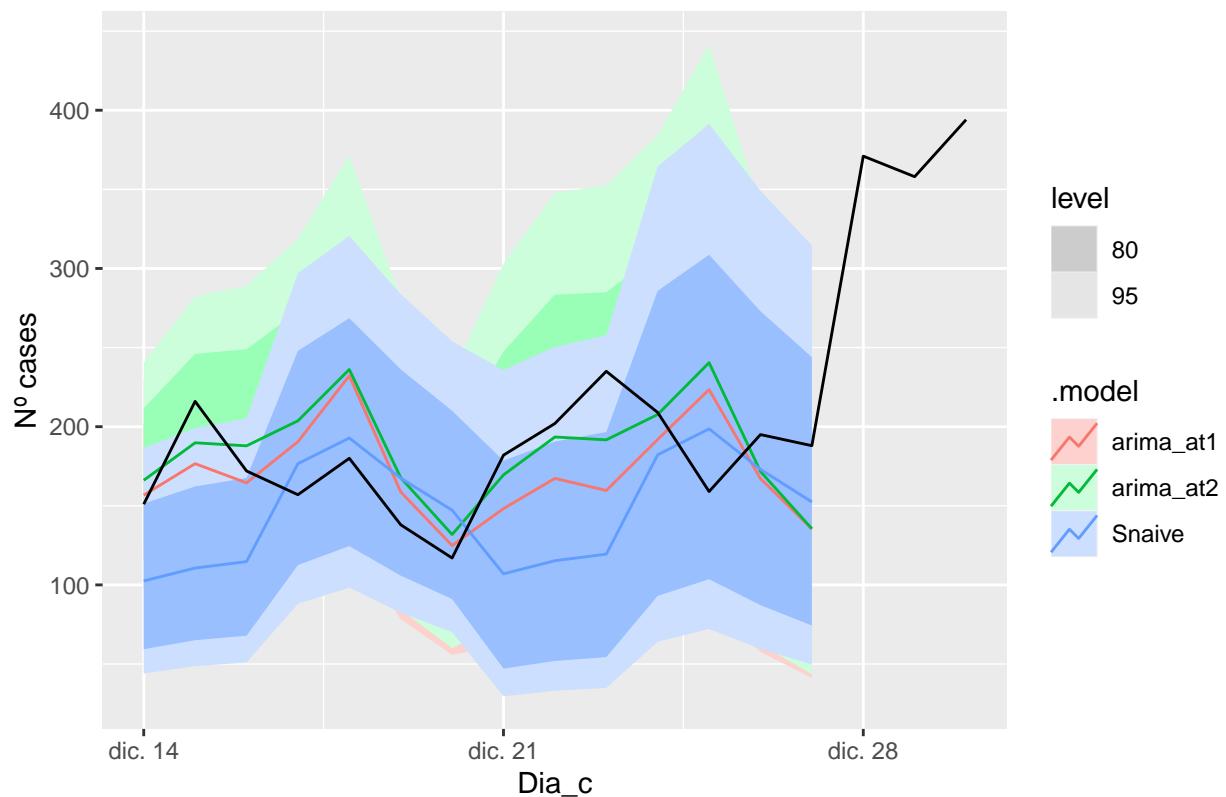
```
fc_h14 %>%
  autoplot(Mal_N_cases) +
  labs(y = "Nº cases", title = "Málaga - forecast h14")
```

Málaga – forecast h14



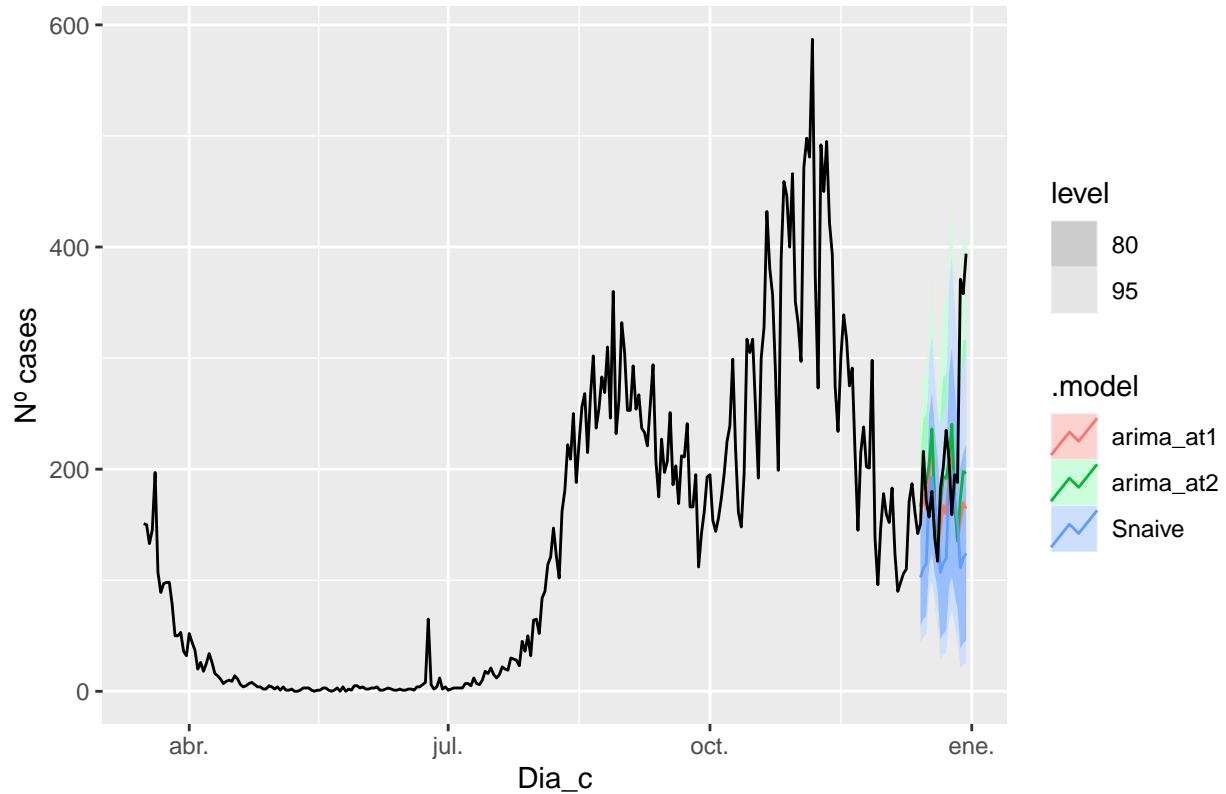
```
fc_h14 %>%
  autoplot(Mal_N_cases_tt) +
  labs(y = "Nº cases", title = "Málaga - forecast h14")
```

Málaga – forecast h14



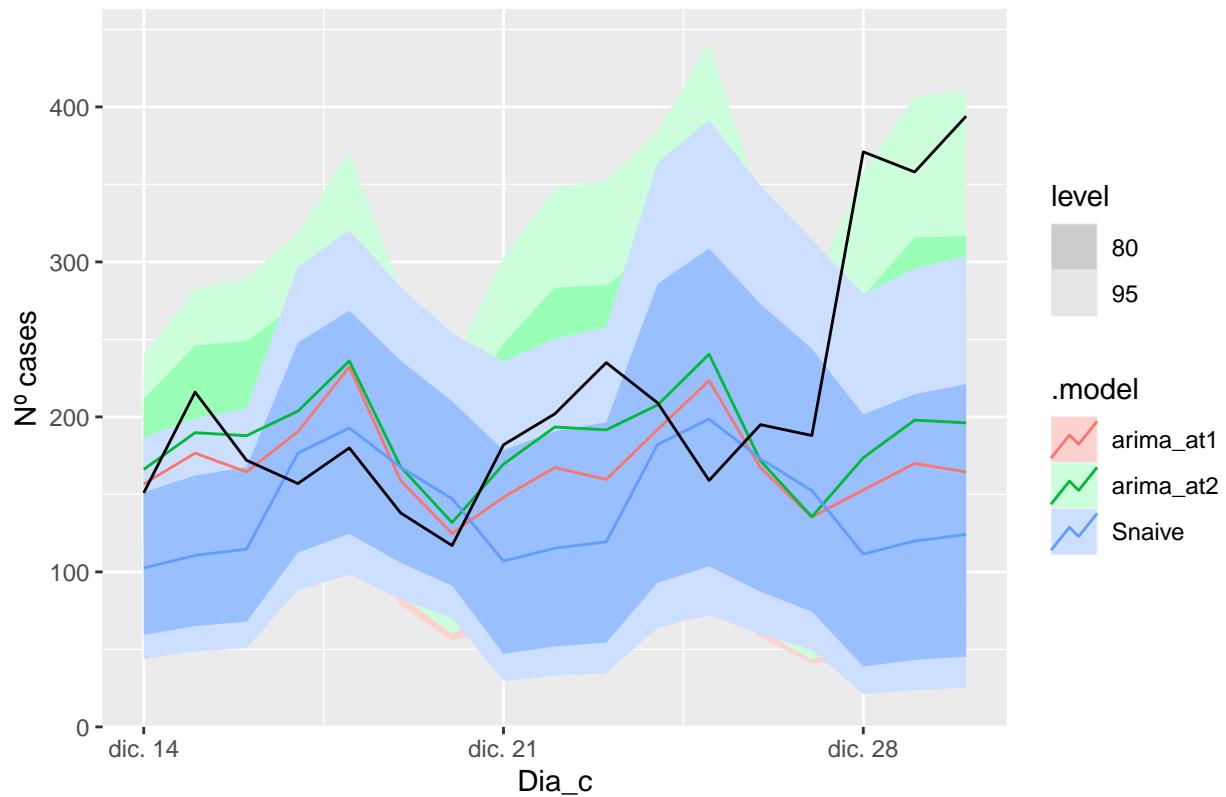
```
fc_h17 %>%
  autoplot(Mal_N_cases) +
  labs(y = "Nº cases", title = "Málaga - forecast h17")
```

Málaga – forecast h17



```
fc_h17 %>%
  autoplot(Mal_N_cases_tt) +
  labs(y = "Nº cases", title = "Málaga - forecast h17")
```

Málaga – forecast h17



```
# Accuracy
fabletools::accuracy(fc_h7, Mal_N_cases)

## # A tibble: 3 x 11
##   .model    sub_region_2 .type     ME   RMSE    MAE    MPE    MAPE    MASE   RMSSE   ACF1
##   <chr>      <chr>    <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 arima_at1 Málaga     Test  -10.4  29.2  23.8 -7.61  14.1  0.680  0.538  0.399
## 2 arima_at2 Málaga     Test  -21.7  33.0  29.2 -14.6   18.1  0.833  0.607  0.357
## 3 Snaive     Málaga     Test   17.0  52.2  43.3  6.77  25.9  1.23   0.961  0.560

fabletools::accuracy(fc_h14, Mal_N_cases)

## # A tibble: 3 x 11
##   .model    sub_region_2 .type     ME   RMSE    MAE    MPE    MAPE    MASE   RMSSE   ACF1
##   <chr>      <chr>    <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 arima_at1 Málaga     Test   7.41  39.6  33.8  1.75  18.4  0.962  0.729  0.269
## 2 arima_at2 Málaga     Test  -6.59  37.5  30.5 -5.96  17.7  0.870  0.689  0.175
## 3 Snaive     Málaga     Test  31.5  59.5  50.3 14.2   27.3  1.43   1.10   0.434

fabletools::accuracy(fc_h17, Mal_N_cases)

## # A tibble: 3 x 11
##   .model    sub_region_2 .type     ME   RMSE    MAE    MPE    MAPE    MASE   RMSSE   ACF1
##   <chr>      <chr>    <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 arima_at1 Málaga     Test  43.5  96.3  65.2 11.4   25.1  1.86   1.77  0.618
## 2 arima_at2 Málaga     Test  27.2  85.2  57.8  3.80   23.3  1.65   1.57  0.627
## 3 Snaive     Málaga     Test  71.1 120.   86.6 23.7   34.5  2.47   2.22  0.590
```

```

# Train data set creation
# Train
Cad_N_cases_tr <- Cad_N_cases %>%
  filter_index("2020-08-10" ~ "2020-12-13")

Cad_N_cases_tt <- Cad_N_cases %>%
  filter_index("2020-12-14" ~ "2020-12-31")

# Modeling over train
Cad_N_cases_tr %>% model(
  arima_at1=ARIMA(box_cox(num_casos.x,lambda_cad)),
  arima_at2=ARIMA(box_cox(num_casos.x,lambda_cad),
                  stepwise = FALSE, approx = FALSE),
  Snaive=SNAIVE(box_cox(num_casos.x,lambda_cad))) -> fit_model

# Show and report model
fit_model

```

3.3.1.4 Cádiz

```

## # A mable: 1 x 4
## # Key:      sub_region_2 [1]
##   sub_region_2           arima_at1           arima_at2     Snaive
##   <chr>                <model>            <model>    <model>
## 1 Cádiz      <ARIMA(0,1,1)(2,0,0)[7]> <ARIMA(0,1,1)(2,0,0)[7]> <SNAIVE>
report(fit_model)

## # A tibble: 3 x 9
##   sub_region_2 .model    sigma2 log_lik   AIC   AICc     BIC ar_roots ma_roots
##   <chr>        <chr>    <dbl>  <dbl> <dbl> <dbl> <dbl> <list>   <list>
## 1 Cádiz       arima_at1  1.36  -197.  402.  403.  414. <cpl [14]> <cpl [1]>
## 2 Cádiz       arima_at2  1.36  -197.  402.  403.  414. <cpl [14]> <cpl [1]>
## 3 Cádiz       Snaive    3.63    NA     NA     NA    NA  <NULL>   <NULL>

# Good model >> Less Sigma / AICc
fit_model %>% pivot_longer(!sub_region_2,
                           names_to = "Model name",
                           values_to = "Orders")

## # A mable: 3 x 3
## # Key:      sub_region_2, Model name [3]
##   sub_region_2 `Model name`          Orders
##   <chr>        <chr>            <model>
## 1 Cádiz       arima_at1      <ARIMA(0,1,1)(2,0,0)[7]>
## 2 Cádiz       arima_at2      <ARIMA(0,1,1)(2,0,0)[7]>
## 3 Cádiz       Snaive        <SNAIVE>

glance(fit_model) %>% arrange(AICc) %>% select(.model:BIC)

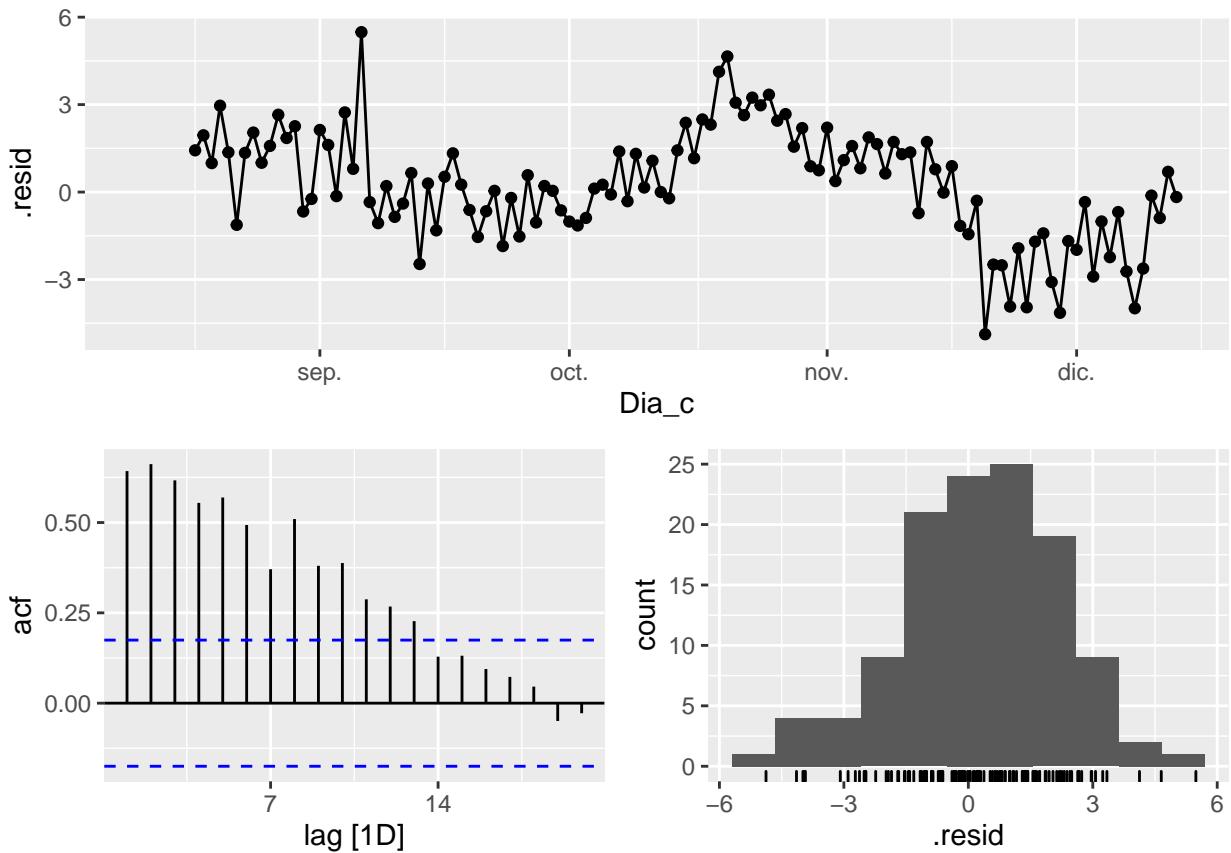
## # A tibble: 3 x 6
##   .model    sigma2 log_lik   AIC   AICc     BIC
##   <chr>    <dbl>  <dbl> <dbl> <dbl> <dbl>
## 1 arima_at1  1.36  -197.  402.  403.  414.
## 2 arima_at2  1.36  -197.  402.  403.  414.

```

```

## 3 Snaive      3.63      NA      NA      NA      NA
# We use a Ljung-Box test >> large p-value, confirms residuals are similar / considered to white noise
fit_model %>% select(Snaive) %>% gg_tsresiduals()

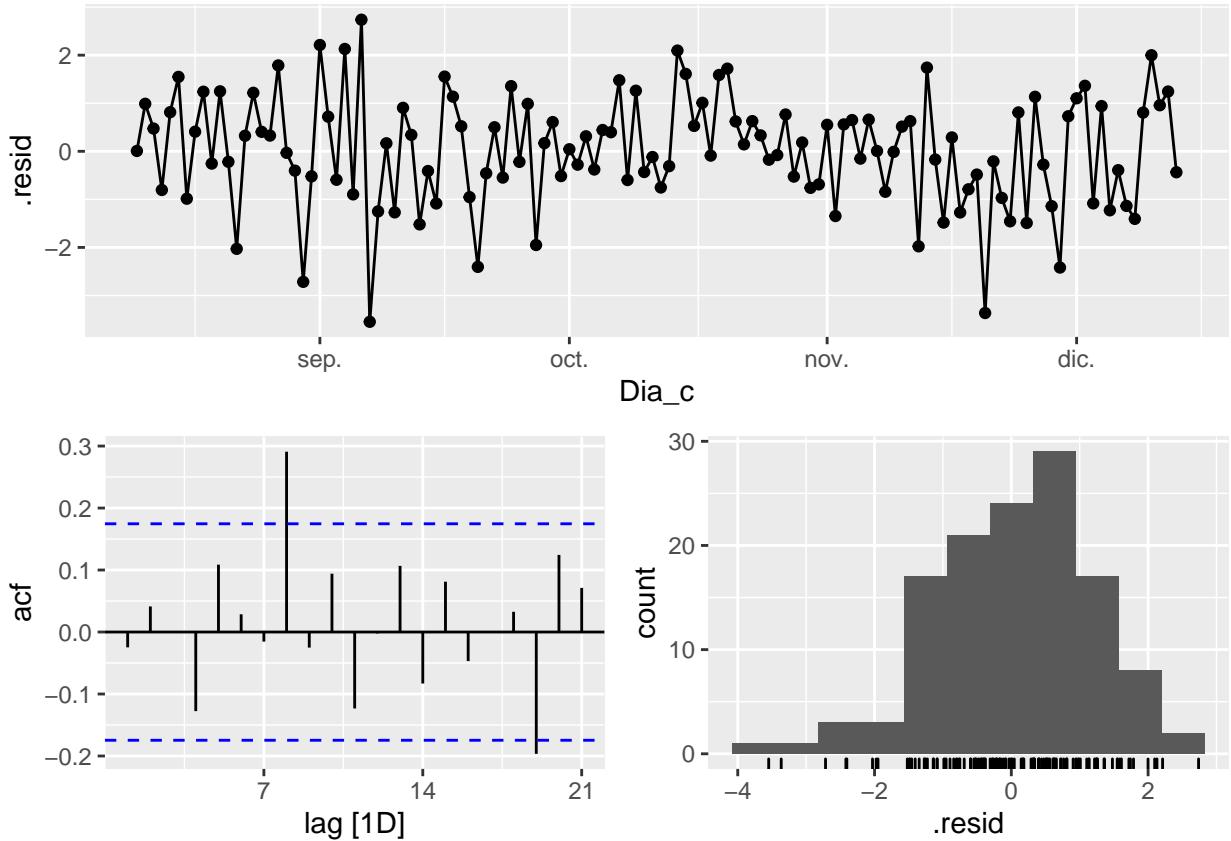
```



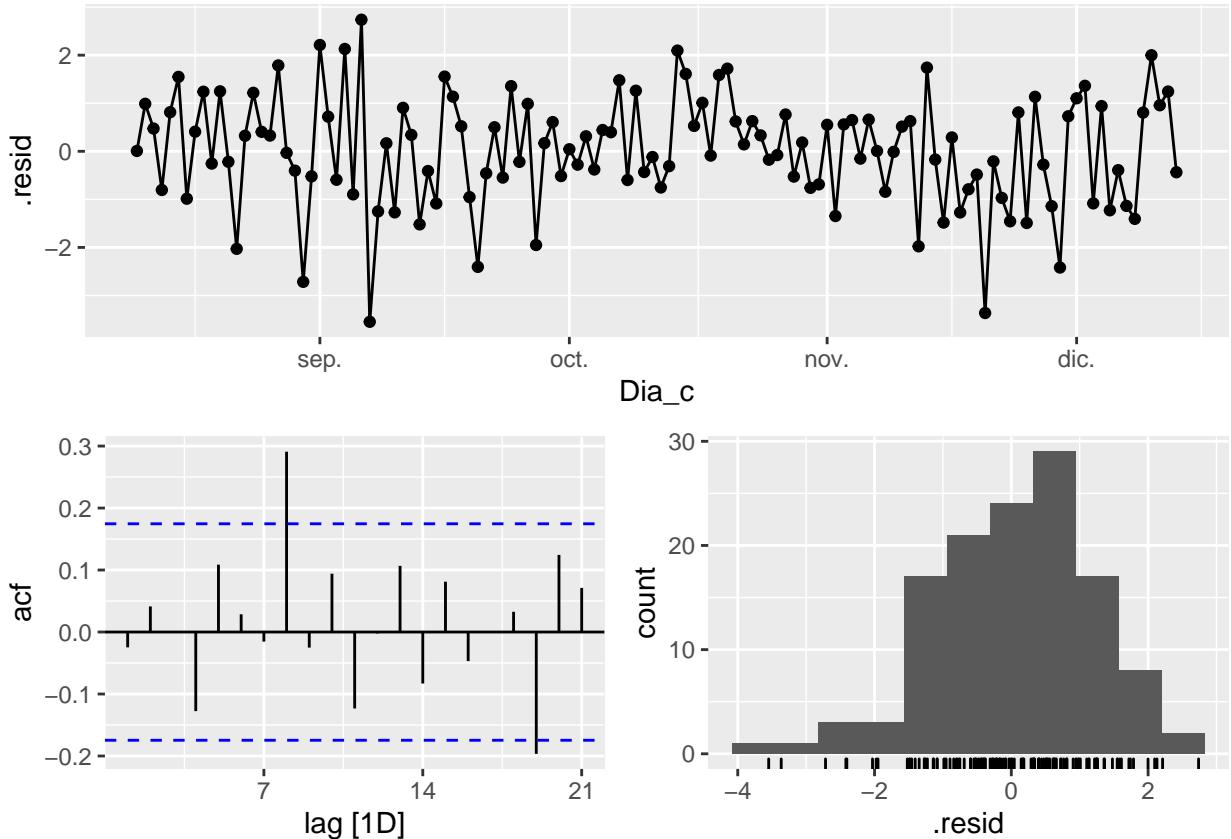
```

fit_model %>% select(arima_at1) %>% gg_tsresiduals()

```



```
fit_model %>% select(arima_at2) %>% gg_tsresiduals()
```



```

augment(fit_model) %>%
  features(.innov, ljung_box, lag=7)

## # A tibble: 3 x 4
##   sub_region_2 .model    lb_stat lb_pvalue
##   <chr>        <chr>      <dbl>    <dbl>
## 1 Cádiz       arima_at1  4.17    0.760
## 2 Cádiz       arima_at2  4.17    0.760
## 3 Cádiz       Snaive     280.     0

augment(fit_model) %>%
  features(.innov, ljung_box, lag=14)

## # A tibble: 3 x 4
##   sub_region_2 .model    lb_stat lb_pvalue
##   <chr>        <chr>      <dbl>    <dbl>
## 1 Cádiz       arima_at1  21.8    0.0828
## 2 Cádiz       arima_at2  21.8    0.0828
## 3 Cádiz       Snaive     382.     0

augment(fit_model) %>%
  features(.innov, ljung_box, lag=21)

## # A tibble: 3 x 4
##   sub_region_2 .model    lb_stat lb_pvalue
##   <chr>        <chr>      <dbl>    <dbl>
## 1 Cádiz       arima_at1  32.2    0.0560
## 2 Cádiz       arima_at2  32.2    0.0560

```

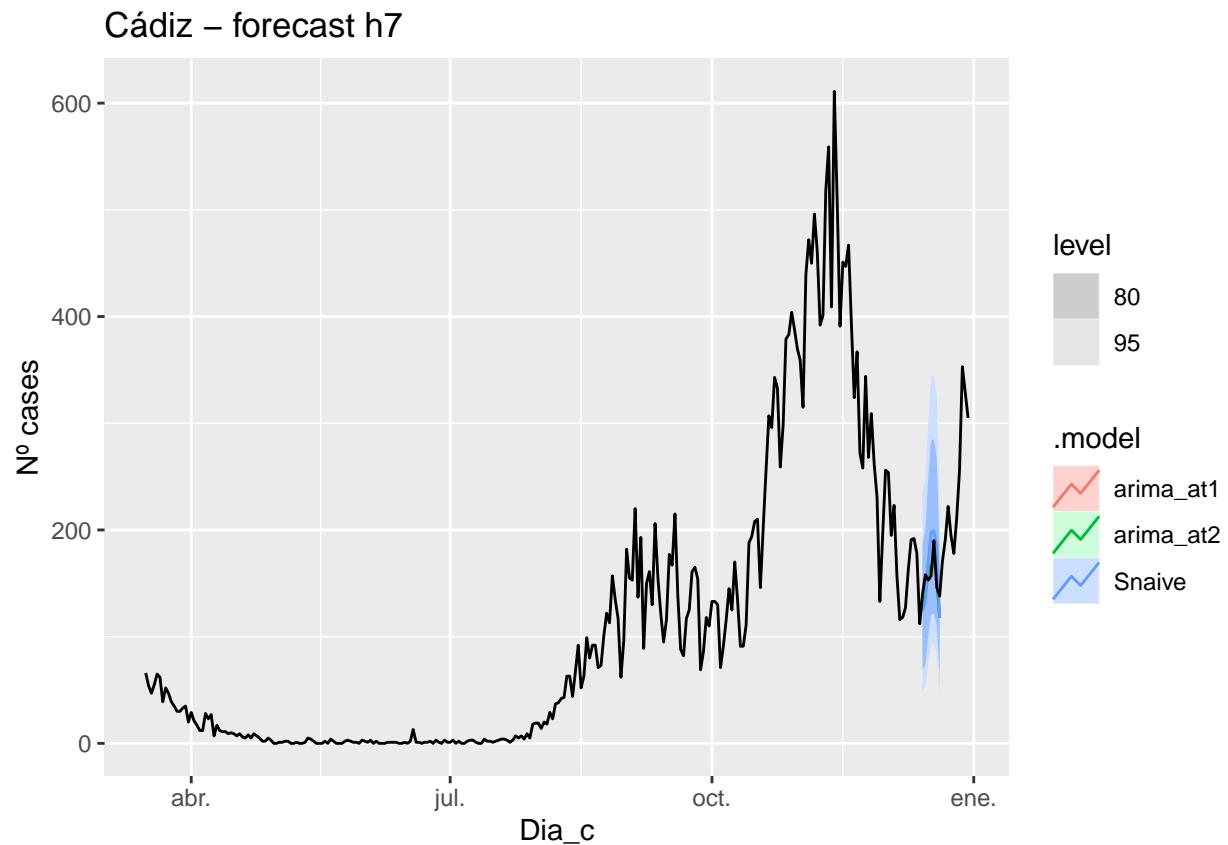
```

## 3 Cádiz      Snaive      388.      0
# Significant spikes out of 30 is still consistent with white noise
# To be sure, use a Ljung-Box test, which has a large p-value, confirming that
# the residuals are similar to white noise.
# Note that the alternative models also passed this test.

# Forecast
fc_h7<-fabletools::forecast(fit_model, h=7)
fc_h14<-fabletools::forecast(fit_model, h=14)
fc_h17<-fabletools::forecast(fit_model, h=17)

# Plots
fc_h7 %>%
  autoplot(Cad_N_cases) +
  labs(y = "Nº cases", title = "Cádiz - forecast h7")

```

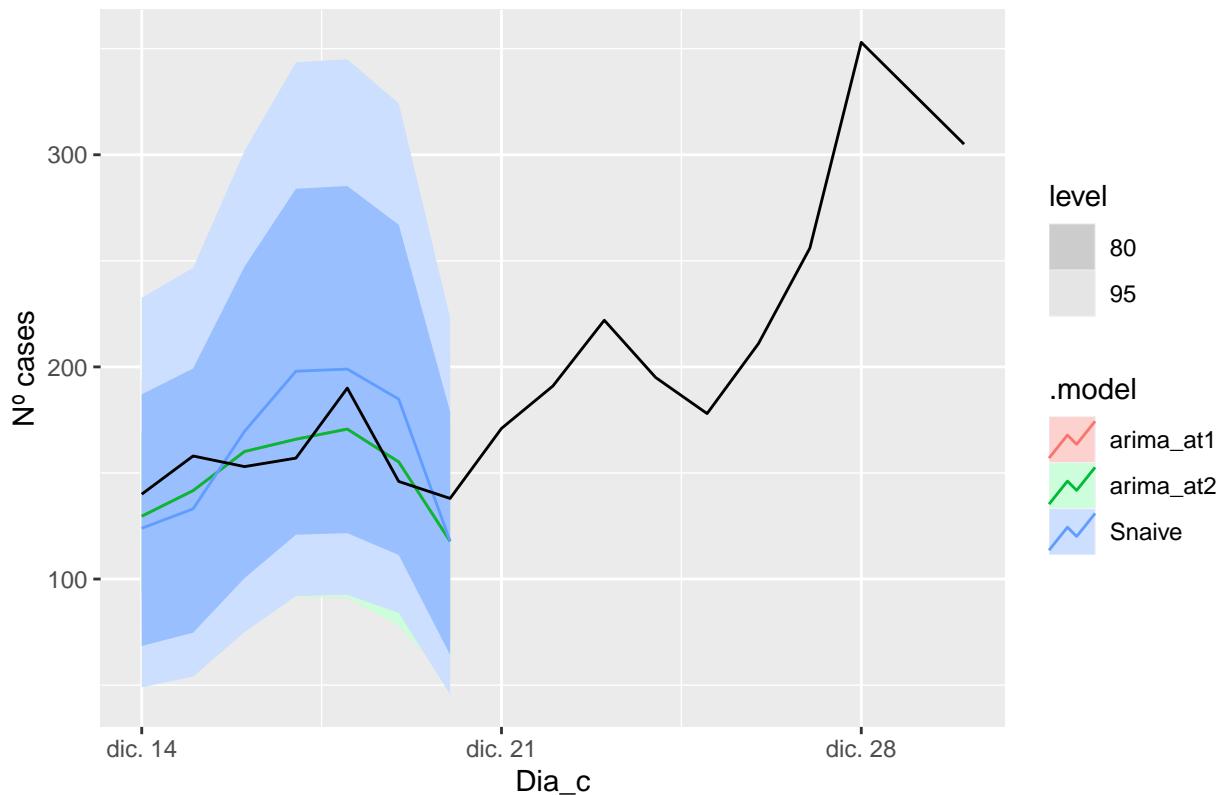


```

fc_h7 %>%
  autoplot(Cad_N_cases_tt) +
  labs(y = "Nº cases", title = "Cádiz - forecast h7")

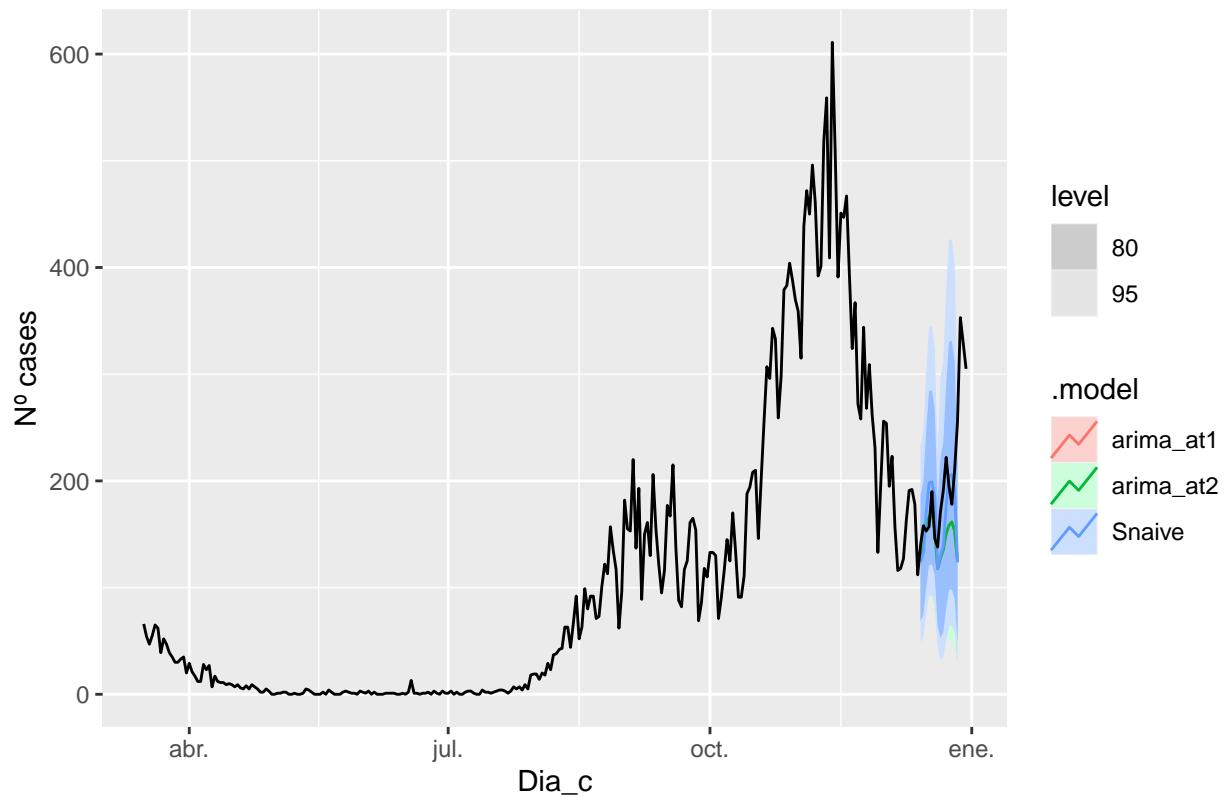
```

Cádiz – forecast h7

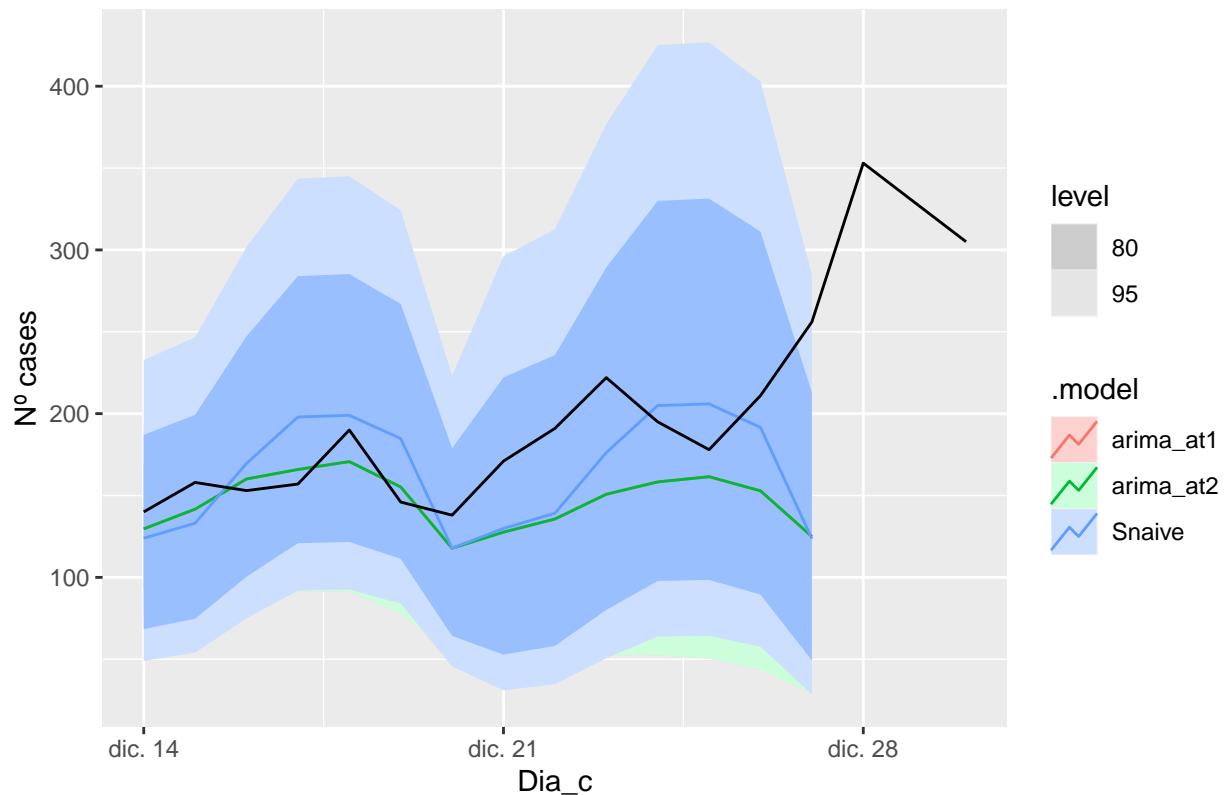


```
fc_h14 %>%
  autoplot(Cad_N_cases) +
  labs(y = "Nº cases", title = "Cádiz - forecast h14")
```

Cádiz – forecast h14

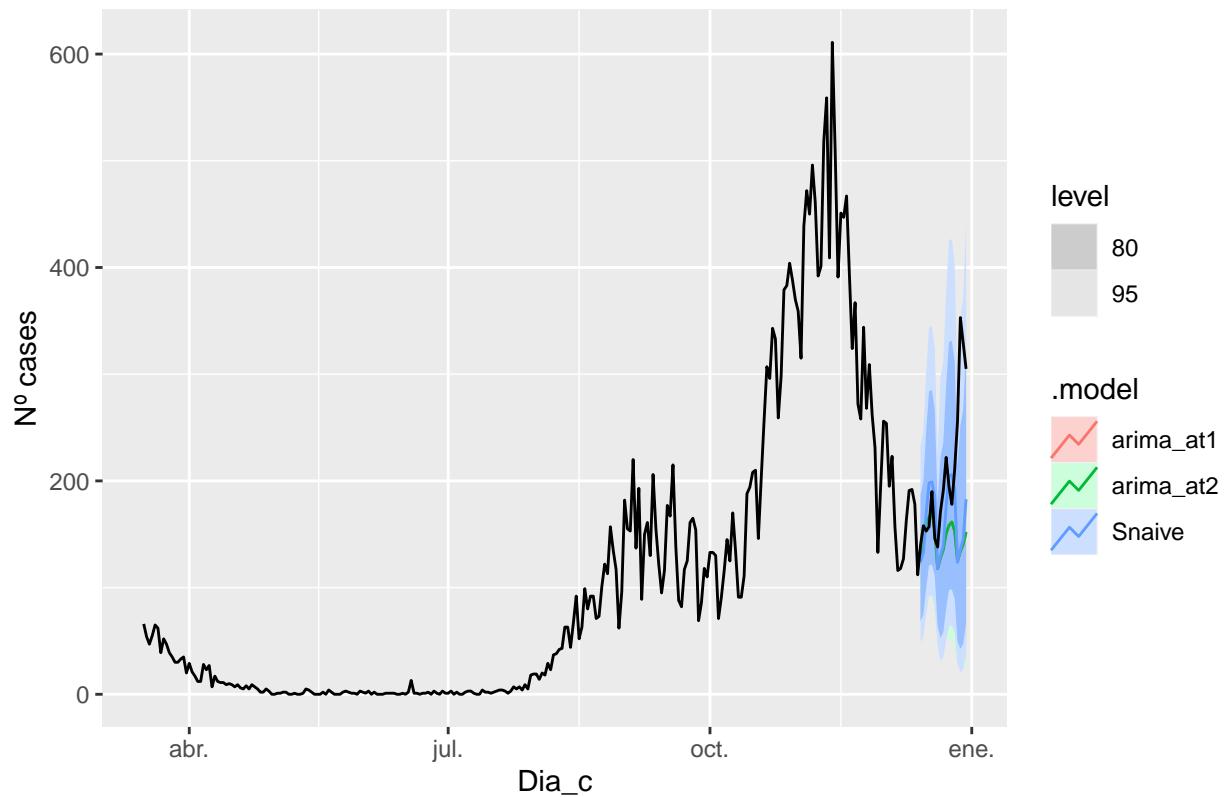


Cádiz – forecast h14

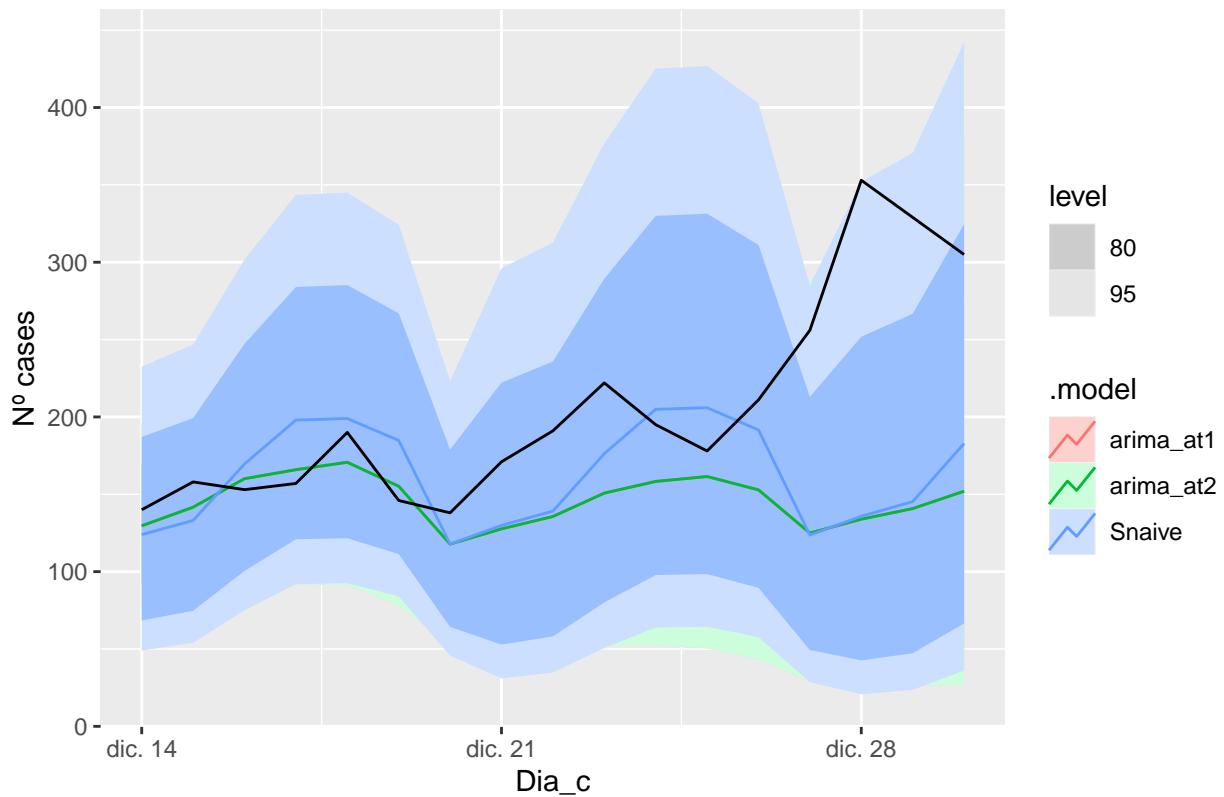


```
fc_h17 %>%
  autoplot(Cad_N_cases) +
  labs(y = "Nº cases", title = "Cádiz - forecast h17")
```

Cádiz – forecast h17



Cádiz – forecast h17



```
# Accuracy
fabletools::accuracy(fc_h7, Cad_N_cases)

## # A tibble: 3 x 11
##   .model    sub_region_2 .type     ME   RMSE    MAE    MPE    MAPE    MASE   RMSSE   ACF1
##   <chr>      <chr>     <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 arima_at1 Cádiz       Test    5.85  14.0  13.1  3.70  8.45  0.464  0.278 -0.455
## 2 arima_at2 Cádiz       Test    5.85  14.0  13.1  3.70  8.45  0.464  0.278 -0.455
## 3 Snaive     Cádiz       Test   -6.31  26.3  23.8 -3.77 15.7   0.846  0.522  0.0115

fabletools::accuracy(fc_h14, Cad_N_cases)

## # A tibble: 3 x 11
##   .model    sub_region_2 .type     ME   RMSE    MAE    MPE    MAPE    MASE   RMSSE   ACF1
##   <chr>      <chr>     <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 arima_at1 Cádiz       Test   32.4  49.0  36.0 15.7   18.0   1.28  0.974  0.395
## 2 arima_at2 Cádiz       Test   32.4  49.0  36.0 15.7   18.0   1.28  0.974  0.395
## 3 Snaive     Cádiz       Test   14.9  46.3  35.3  6.11   18.8   1.26  0.922  0.247

fabletools::accuracy(fc_h17, Cad_N_cases)

## # A tibble: 3 x 11
##   .model    sub_region_2 .type     ME   RMSE    MAE    MPE    MAPE    MASE   RMSSE   ACF1
##   <chr>      <chr>     <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 arima_at1 Cádiz       Test   59.6  90.9  62.6 22.9   24.8   2.23  1.81  0.789
## 2 arima_at2 Cádiz       Test   59.6  90.9  62.6 22.9   24.8   2.23  1.81  0.789
## 3 Snaive     Cádiz       Test   43.0  86.1  59.9 14.3   24.8   2.13  1.71  0.757
```

```

# Train data set creation
# Train
Sev_N_cases_tr <- Sev_N_cases %>%
  filter_index("2020-08-10" ~ "2020-12-13")

Sev_N_cases_tt <- Sev_N_cases %>%
  filter_index("2020-12-14" ~ "2020-12-31")

# Modeling over train
Sev_N_cases_tr %>% model(
  arima_at1=ARIMA(box_cox(num_casos.x,lambda_sev)),
  arima_at2=ARIMA(box_cox(num_casos.x,lambda_sev),
                  stepwise = FALSE, approx = FALSE),
  Snaive=SNAIVE(box_cox(num_casos.x,lambda_sev))) -> fit_model

# Show and report model
fit_model

```

3.3.1.5 Sevilla

```

## # A mable: 1 x 4
## # Key:      sub_region_2 [1]
##   sub_region_2           arima_at1           arima_at2     Snaive
##   <chr>                <model>            <model>    <model>
## 1 Sevilla    <ARIMA(0,1,1)(0,1,1)[7]> <ARIMA(0,1,1)(0,1,1)[7]> <SNAIVE>
report(fit_model)

## # A tibble: 3 x 9
##   sub_region_2 .model    sigma2 log_lik   AIC   AICc     BIC ar_roots ma_roots
##   <chr>        <chr>    <dbl>  <dbl> <dbl> <dbl> <dbl> <list>   <list>
## 1 Sevilla      arima_at1  0.813  -157.  319.  320.  328. <cpl [0]> <cpl [8]>
## 2 Sevilla      arima_at2  0.813  -157.  319.  320.  328. <cpl [0]> <cpl [8]>
## 3 Sevilla      Snaive    2.94    NA     NA     NA    NA  <NULL>  <NULL>

# Good model >> Less Sigma / AICc
fit_model %>% pivot_longer(!sub_region_2,
                           names_to = "Model name",
                           values_to = "Orders")

## # A mable: 3 x 3
## # Key:      sub_region_2, Model name [3]
##   sub_region_2 `Model name`          Orders
##   <chr>        <chr>            <model>
## 1 Sevilla      arima_at1    <ARIMA(0,1,1)(0,1,1)[7]>
## 2 Sevilla      arima_at2    <ARIMA(0,1,1)(0,1,1)[7]>
## 3 Sevilla      Snaive       <SNAIVE>

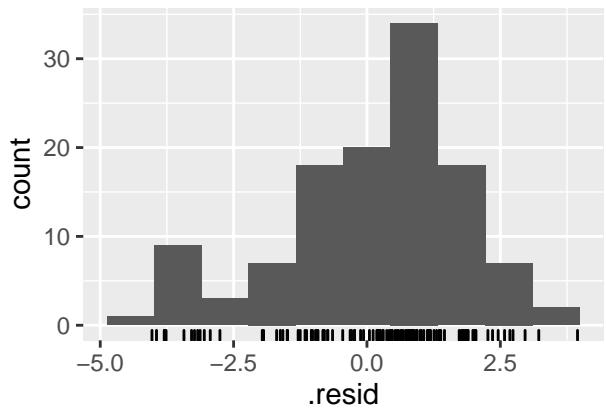
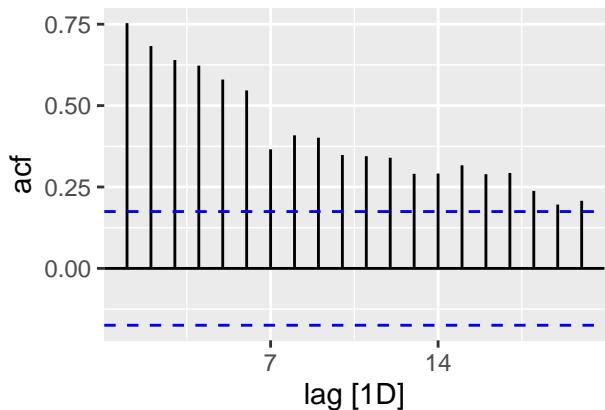
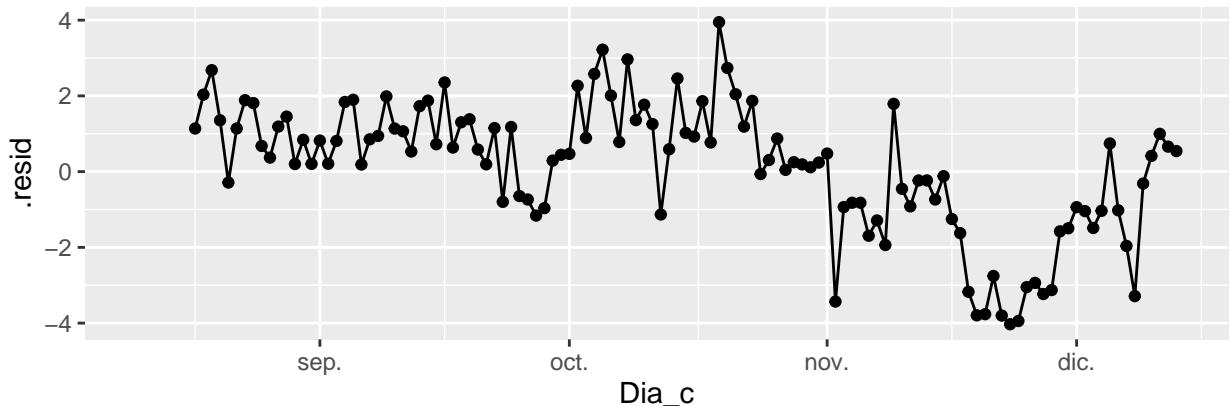
glance(fit_model) %>% arrange(AICc) %>% select(.model:BIC)

## # A tibble: 3 x 6
##   .model    sigma2 log_lik   AIC   AICc     BIC
##   <chr>    <dbl>  <dbl> <dbl> <dbl> <dbl>
## 1 arima_at1  0.813  -157.  319.  320.  328.
## 2 arima_at2  0.813  -157.  319.  320.  328.

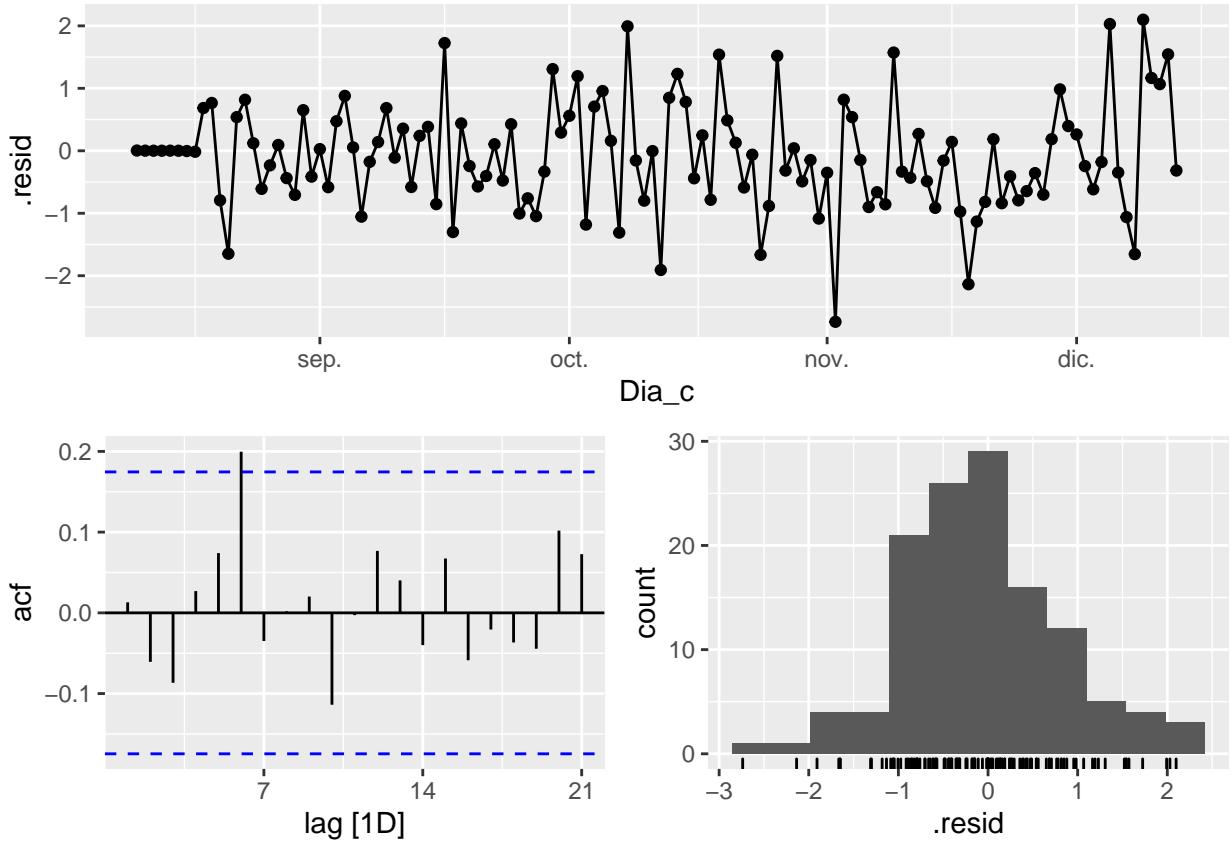
```

```
## 3 Snaive      2.94      NA      NA      NA      NA
```

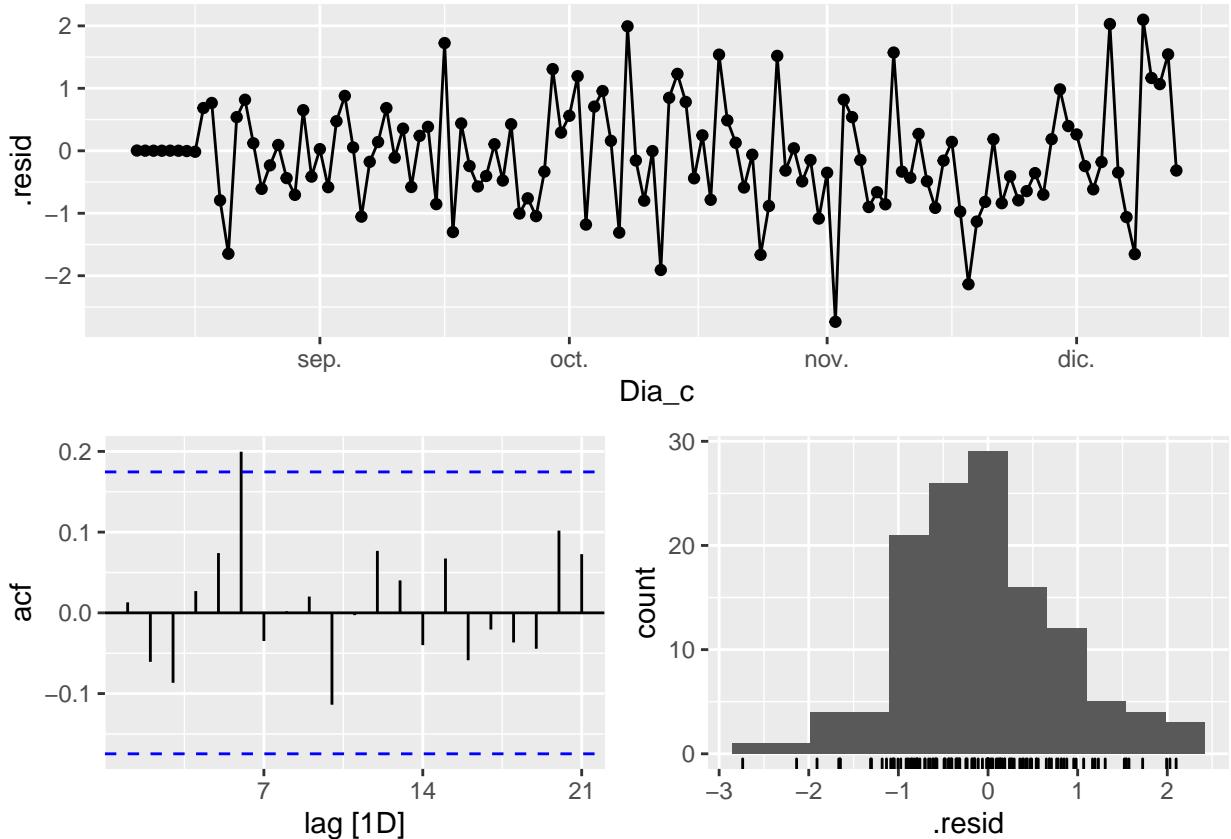
```
# We use a Ljung-Box test >> large p-value, confirms residuals are similar / considered to white noise
fit_model %>% select(Snaive) %>% gg_tsresiduals()
```



```
fit_model %>% select(arima_at1) %>% gg_tsresiduals()
```



```
fit_model %>% select(arima_at2) %>% gg_tsresiduals()
```



```

augment(fit_model) %>%
  features(.innov, ljung_box, lag=7)

## # A tibble: 3 x 4
##   sub_region_2 .model    lb_stat lb_pvalue
##   <chr>        <chr>      <dbl>     <dbl>
## 1 Sevilla      arima_at1  7.83     0.348
## 2 Sevilla      arima_at2  7.83     0.348
## 3 Sevilla      Snaive     324.      0

augment(fit_model) %>%
  features(.innov, ljung_box, lag=14)

## # A tibble: 3 x 4
##   sub_region_2 .model    lb_stat lb_pvalue
##   <chr>        <chr>      <dbl>     <dbl>
## 1 Sevilla      arima_at1 11.0      0.687
## 2 Sevilla      arima_at2 11.0      0.687
## 3 Sevilla      Snaive     437.      0

augment(fit_model) %>%
  features(.innov, ljung_box, lag=21)

## # A tibble: 3 x 4
##   sub_region_2 .model    lb_stat lb_pvalue
##   <chr>        <chr>      <dbl>     <dbl>
## 1 Sevilla      arima_at1 15.1      0.818
## 2 Sevilla      arima_at2 15.1      0.818

```

```

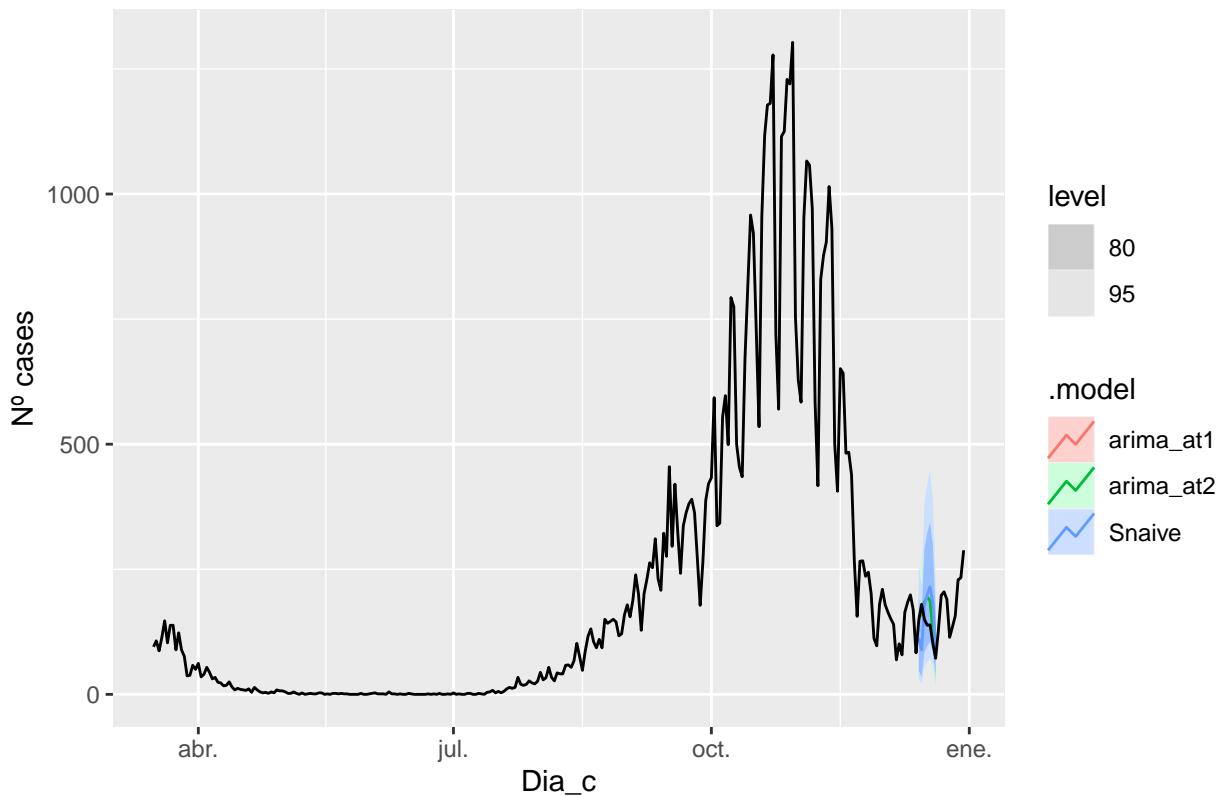
## 3 Sevilla      Snaive      501.      0
# Significant spikes out of 30 is still consistent with white noise
# To be sure, use a Ljung-Box test, which has a large p-value, confirming that
# the residuals are similar to white noise.
# Note that the alternative models also passed this test.

# Forecast
fc_h7<-fabletools::forecast(fit_model, h=7)
fc_h14<-fabletools::forecast(fit_model, h=14)
fc_h17<-fabletools::forecast(fit_model, h=17)

# Plots
fc_h7 %>%
  autoplot(Sev_N_cases) +
  labs(y = "Nº cases", title = "Sevilla - forecast h7")

```

Sevilla – forecast h7

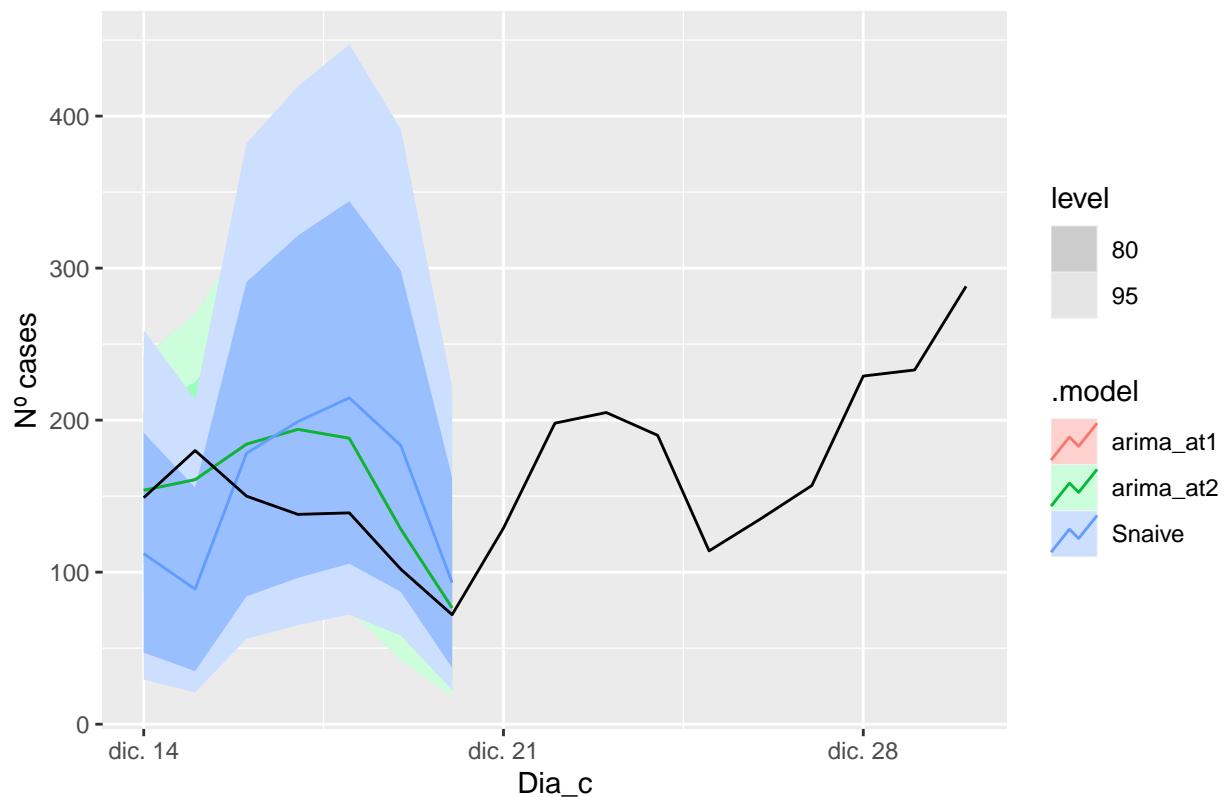


```

fc_h7 %>%
  autoplot(Sev_N_cases_tt) +
  labs(y = "Nº cases", title = "Sevilla - forecast h7")

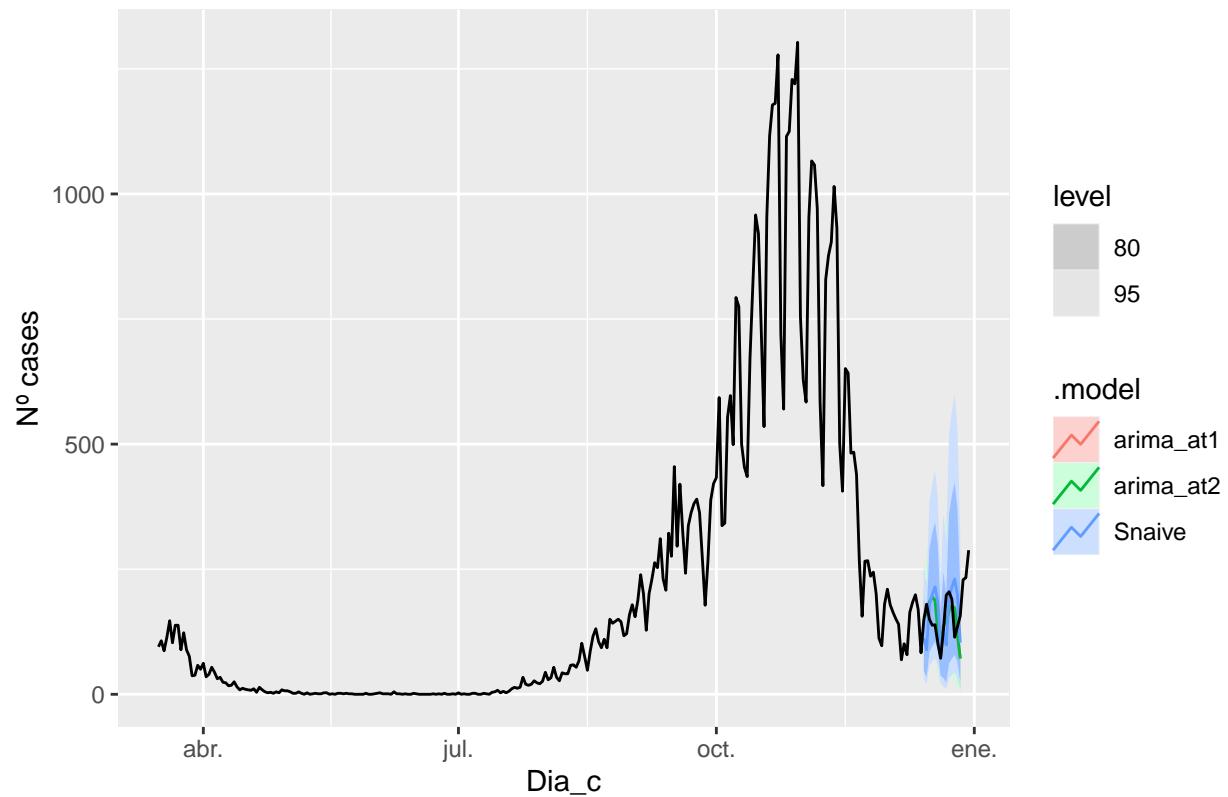
```

Sevilla – forecast h7



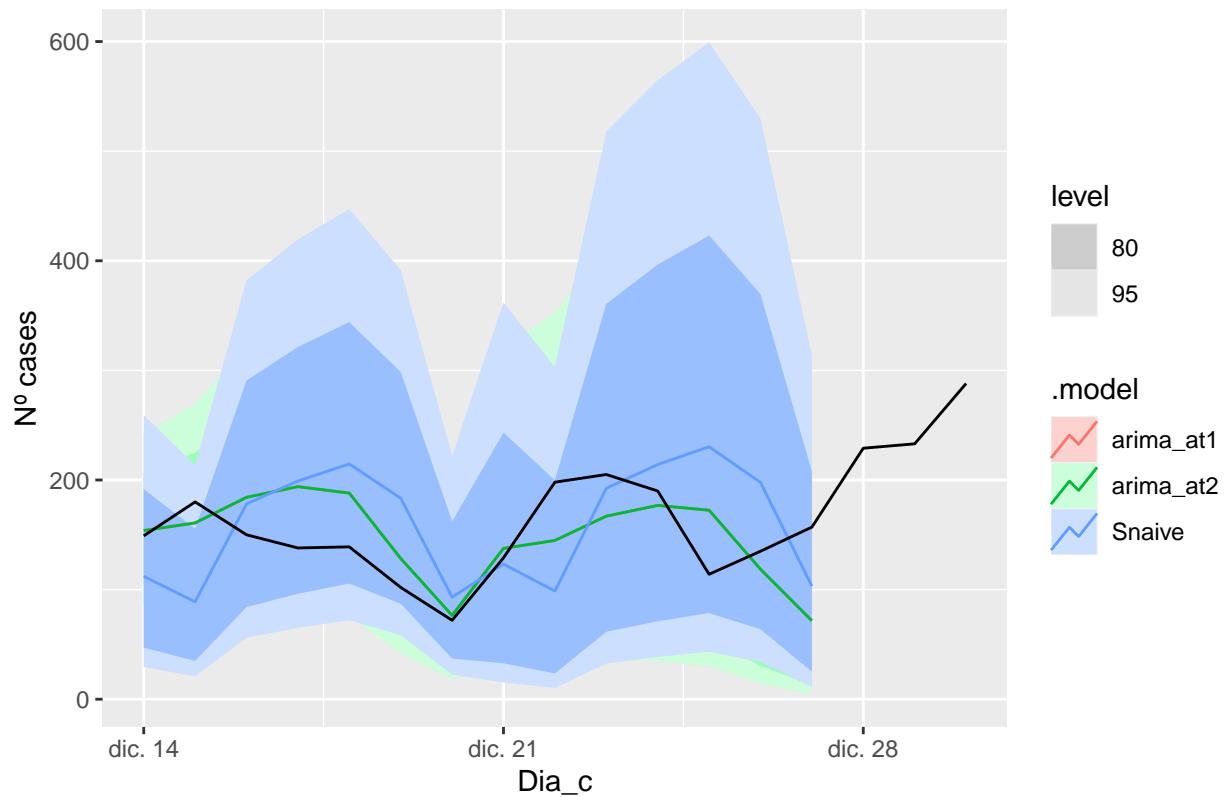
```
fc_h14 %>%
  autoplot(Sev_N_cases) +
  labs(y = "Nº cases", title = "Sevilla - forecast h14")
```

Sevilla – forecast h14



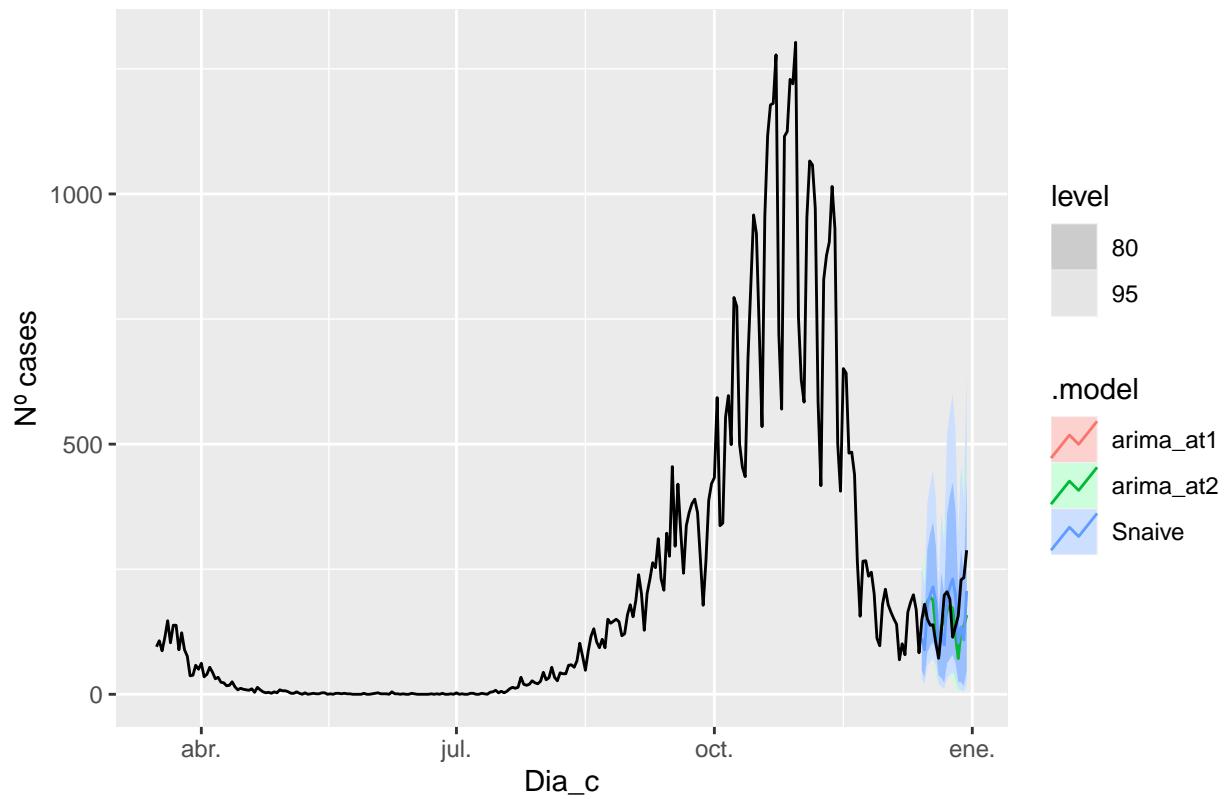
```
fc_h14 %>%
  autoplot(Sev_N_cases_tt) +
  labs(y = "Nº cases", title = "Sevilla - forecast h14")
```

Sevilla – forecast h14

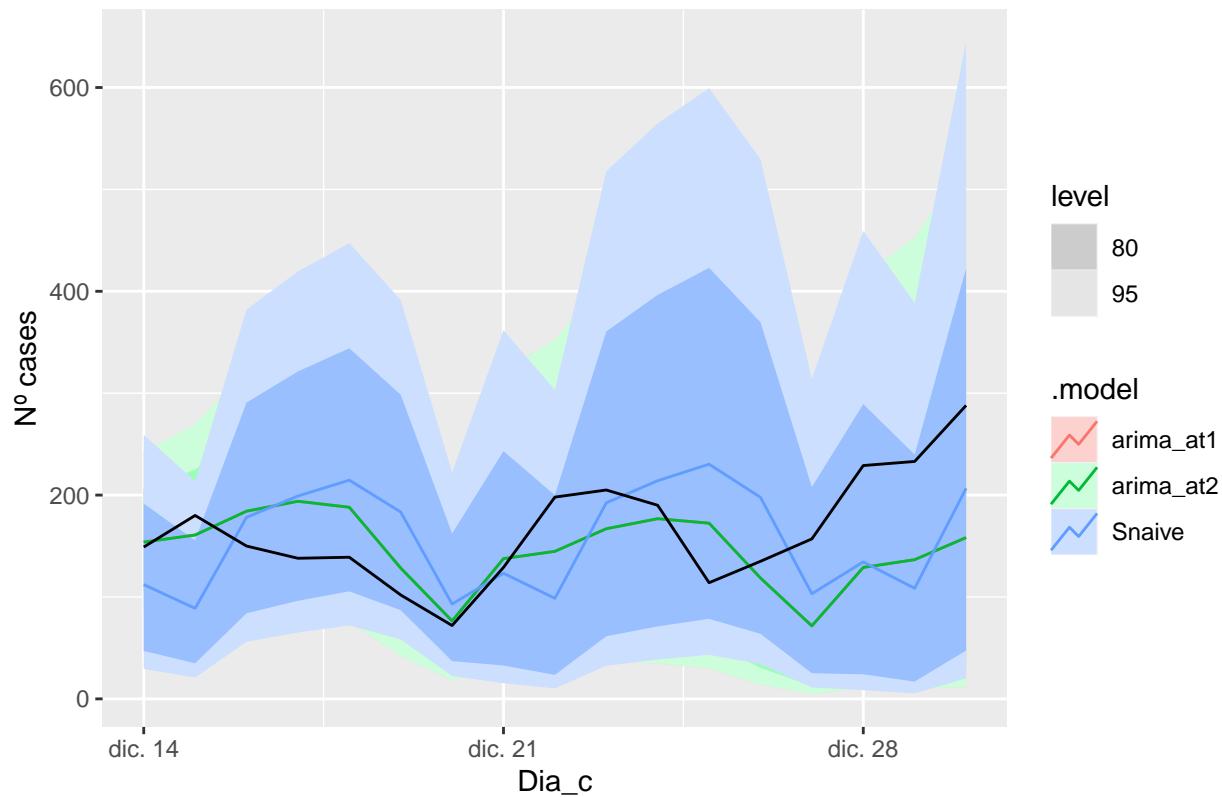


```
fc_h17 %>%
  autoplot(Sev_N_cases) +
  labs(y = "Nº cases", title = "Sevilla - forecast h17")
```

Sevilla – forecast h17



Sevilla – forecast h17



```
# Accuracy
fabletools::accuracy(fc_h7, Sev_N_cases)

## # A tibble: 3 x 11
##   .model    sub_region_2 .type     ME   RMSE    MAE    MPE    MAPE    MASE RMSSE   ACF1
##   <chr>      <chr>     <chr>  <dbl> <dbl>  <dbl>  <dbl>  <dbl>  <dbl> <dbl> <dbl>
## 1 arima_at1 Sevilla     Test  -22.3  33.4  27.7 -17.6  20.7  0.446  0.279  0.361
## 2 arima_at2 Sevilla     Test  -22.3  33.4  27.7 -17.6  20.7  0.446  0.279  0.361
## 3 Snaive      Sevilla     Test  -19.9  62.1  56.5 -21.6  43.1  0.909  0.519  0.476

fabletools::accuracy(fc_h14, Sev_N_cases)

## # A tibble: 3 x 11
##   .model    sub_region_2 .type     ME   RMSE    MAE    MPE    MAPE    MASE RMSSE   ACF1
##   <chr>      <chr>     <chr>  <dbl> <dbl>  <dbl>  <dbl>  <dbl>  <dbl> <dbl> <dbl>
## 1 arima_at1 Sevilla     Test  -1.20  40.8  33.4 -4.47  23.0  0.537  0.341  0.300
## 2 arima_at2 Sevilla     Test  -1.20  40.8  33.4 -4.47  23.0  0.537  0.341  0.300
## 3 Snaive      Sevilla     Test  -12.2  64.4  55.0 -15.5  39.9  0.885  0.538  0.354

fabletools::accuracy(fc_h17, Sev_N_cases)

## # A tibble: 3 x 11
##   .model    sub_region_2 .type     ME   RMSE    MAE    MPE    MAPE    MASE RMSSE   ACF1
##   <chr>      <chr>     <chr>  <dbl> <dbl>  <dbl>  <dbl>  <dbl>  <dbl> <dbl> <dbl>
## 1 arima_at1 Sevilla     Test  18.2   59.1  46.7  3.97  26.6  0.751  0.494  0.634
## 2 arima_at2 Sevilla     Test  18.2   59.1  46.7  3.97  26.6  0.751  0.494  0.634
## 3 Snaive      Sevilla     Test  7.60   72.4  63.0 -5.56  40.1  1.01   0.605  0.568
```

3.3.2 Multivariate (7, 14, 17 days) + residential_percent_change (Google) + Total (INE - EM3)

```

# Opt A
# We have added "residential_percent_change" and "Total" variables to models
lambda_bar_b <- Bar_N_cases %>%
  features(residential_percent_change_from_baseline, features = guererro) %>%
  pull(lambda_guererro)
lambda_bar_h <- Bar_N_cases %>%
  features(Total, features = guererro) %>%
  pull(lambda_guererro)

fit_model <- Bar_N_cases_tr %>%
  model(
    SNaive = SNAIVE(num_casos.x),
    arima_at1 = ARIMA(box_cox(num_casos.x,lambda_bar) ~
                      box_cox(residential_percent_change_from_baseline,lambda_bar_b)+
                      box_cox(Total,lambda_bar_h)),
    arima_at2 = ARIMA(box_cox(num_casos.x,lambda_bar) ~
                      box_cox(residential_percent_change_from_baseline,lambda_bar_b)+
                      box_cox(Total,lambda_bar_h) ,
                      stepwise = FALSE, approx = FALSE))

# Show and report model
fit_model

```

3.3.2.1 Barcelona

```

## # A mable: 1 x 4
## # Key:      sub_region_2 [1]
##   sub_region_2   SNaive                         arima_at1
##   <chr>          <model>                        <model>
## 1 Barcelona     <SNAIVE> <LM w/ ARIMA(0,1,2)(0,0,2)[7] errors>
## # ... with 1 more variable: arima_at2 <model>

fit_model %>% select(arima_at1) %>% report()

## Series: num_casos.x
## Model: LM w/ ARIMA(0,1,2)(0,0,2)[7] errors
## Transformation: box_cox(num_casos.x, lambda_bar)
##
## Coefficients:
##           ma1      ma2      sma1      sma2
##           -0.2940  -0.2605  0.6841  0.4128
## s.e.    0.1003   0.0933  0.1260  0.0877
##           box_cox(residential_percent_change_from_baseline, lambda_bar_b)
##                                         -0.0643
## s.e.                               0.0197
##           box_cox(Total, lambda_bar_h)
##                                         0.3839
## s.e.                               0.0939
## 
## sigma^2 estimated as 0.1744: log likelihood=-67.5
## AIC=148.99   AICc=149.95   BIC=168.79

```

```

fit_model %>% select(arima_at2) %>% report()

## Series: num_casos.x
## Model: LM w/ ARIMA(0,1,5)(1,0,0)[7] errors
## Transformation: box_cox(num_casos.x, lambda_bar)
##
## Coefficients:
##             ma1      ma2      ma3      ma4      ma5     sar1
##             -0.1981 -0.1154 -0.0259  0.1134 -0.0771  0.9318
## s.e.      0.0981  0.1440  0.1072  0.1253  0.1295  0.0301
##             box_cox(residential_percent_change_from_baseline, lambda_bar_b)
##                                         -0.1089
## s.e.                               0.0118
##             box_cox(Total, lambda_bar_h)
##                                         0.028
## s.e.                               0.117
##
## sigma^2 estimated as 0.09879: log likelihood=-35.7
## AIC=89.4   AICc=90.96   BIC=114.85

# Good model >> Less Sigma / AICc
fit_model %>% pivot_longer(!sub_region_2,
                           names_to = "Model name",
                           values_to = "Orders")

## # A mable: 3 x 3
## # Key:   sub_region_2, Model name [3]
##   sub_region_2 `Model name`          Orders
##   <chr>           <chr>           <model>
## 1 Barcelona       SNaive          <SNAIVE>
## 2 Barcelona       arima_at1      <LM w/ ARIMA(0,1,2)(0,0,2)[7] errors>
## 3 Barcelona       arima_at2      <LM w/ ARIMA(0,1,5)(1,0,0)[7] errors>

glance(fit_model) %>% arrange(AICc) %>% select(.model:BIC)

## # A tibble: 3 x 6
##   .model      sigma2 logLik  AIC  AICc    BIC
##   <chr>      <dbl>  <dbl> <dbl> <dbl> <dbl>
## 1 arima_at2  0.0988 -35.7  89.4  91.0  115.
## 2 arima_at1  0.174   -67.5 149.  150.  169.
## 3 SNaive     299977.     NA    NA    NA    NA

# We use a Ljung-Box test >> large p-value, confirms residuals are similar to white noise.
augment(fit_model) %>%
  features(.innov, ljung_box, lag=7)

## # A tibble: 3 x 4
##   sub_region_2 .model    lb_stat lb_pvalue
##   <chr>        <chr>      <dbl>     <dbl>
## 1 Barcelona    arima_at1  17.5     0.0146
## 2 Barcelona    arima_at2  16.2     0.0231
## 3 Barcelona    SNaive     378.      0

augment(fit_model) %>%
  features(.innov, ljung_box, lag=14)

## # A tibble: 3 x 4

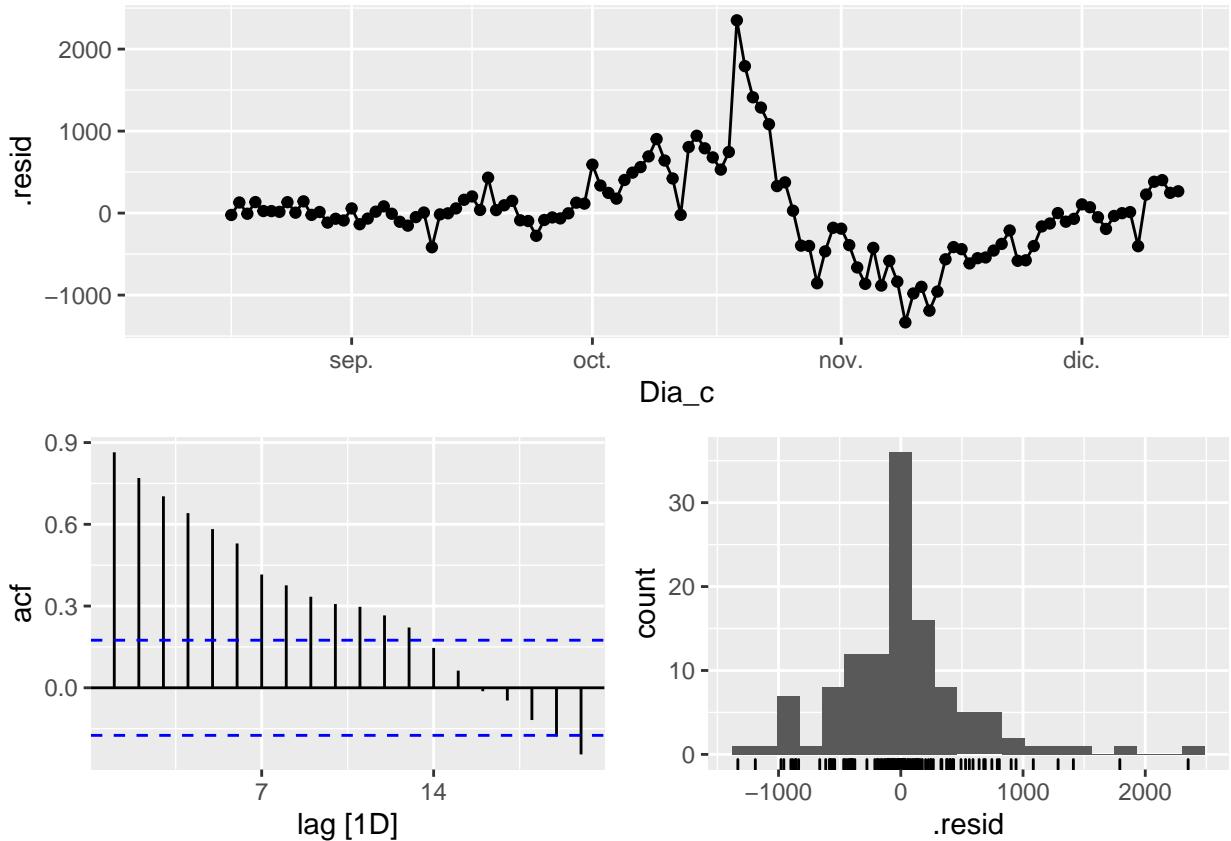
```

```

##   sub_region_2 .model    lb_stat lb_pvalue
##   <chr>        <chr>      <dbl>     <dbl>
## 1 Barcelona    arima_at1    45.3  0.0000361
## 2 Barcelona    arima_at2    24.3  0.0417
## 3 Barcelona    SNaive      454.   0
augment(fit_model) %>%
  features(.innov, ljung_box, lag=21)

## # A tibble: 3 x 4
##   sub_region_2 .model    lb_stat lb_pvalue
##   <chr>        <chr>      <dbl>     <dbl>
## 1 Barcelona    arima_at1    91.4  9.18e-11
## 2 Barcelona    arima_at2    29.3  1.07e- 1
## 3 Barcelona    SNaive      482.   0
fit_model %>% select(SNaive) %>% gg_tsresiduals()

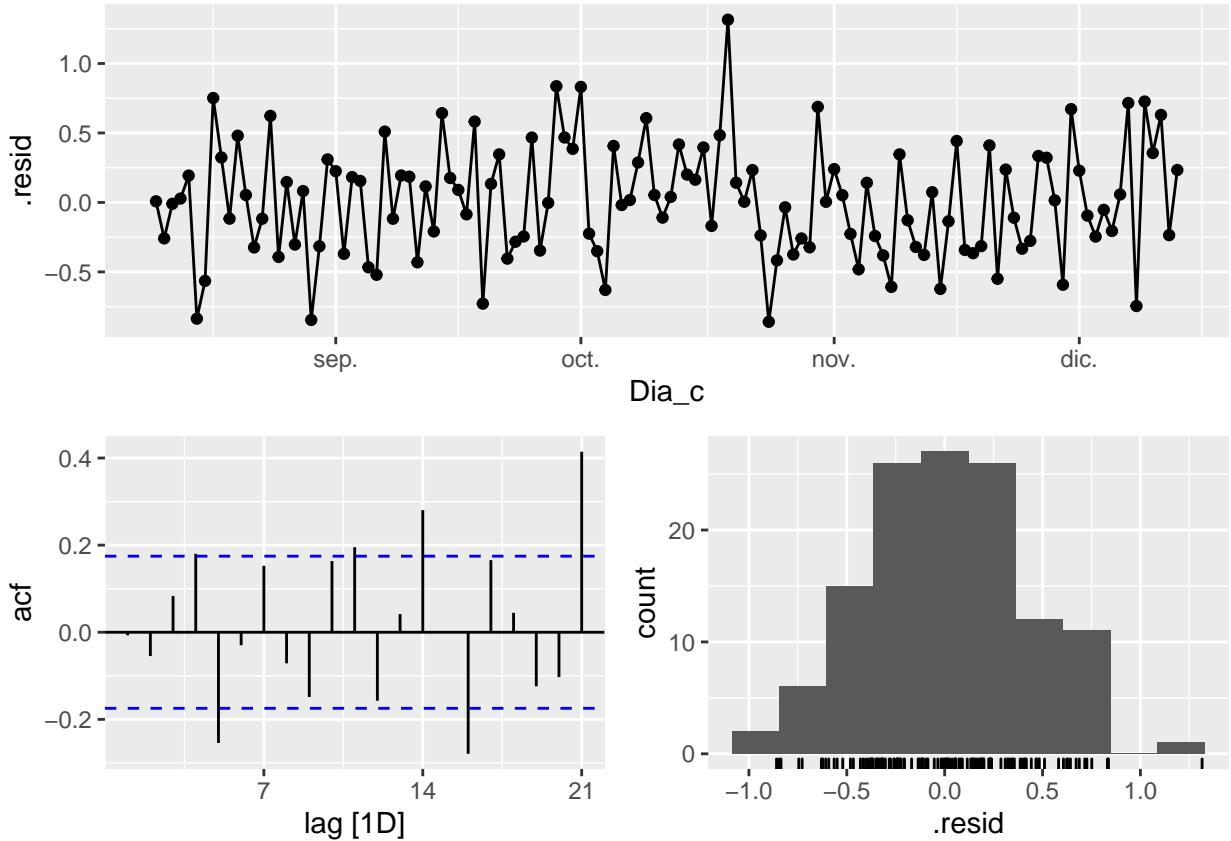
```



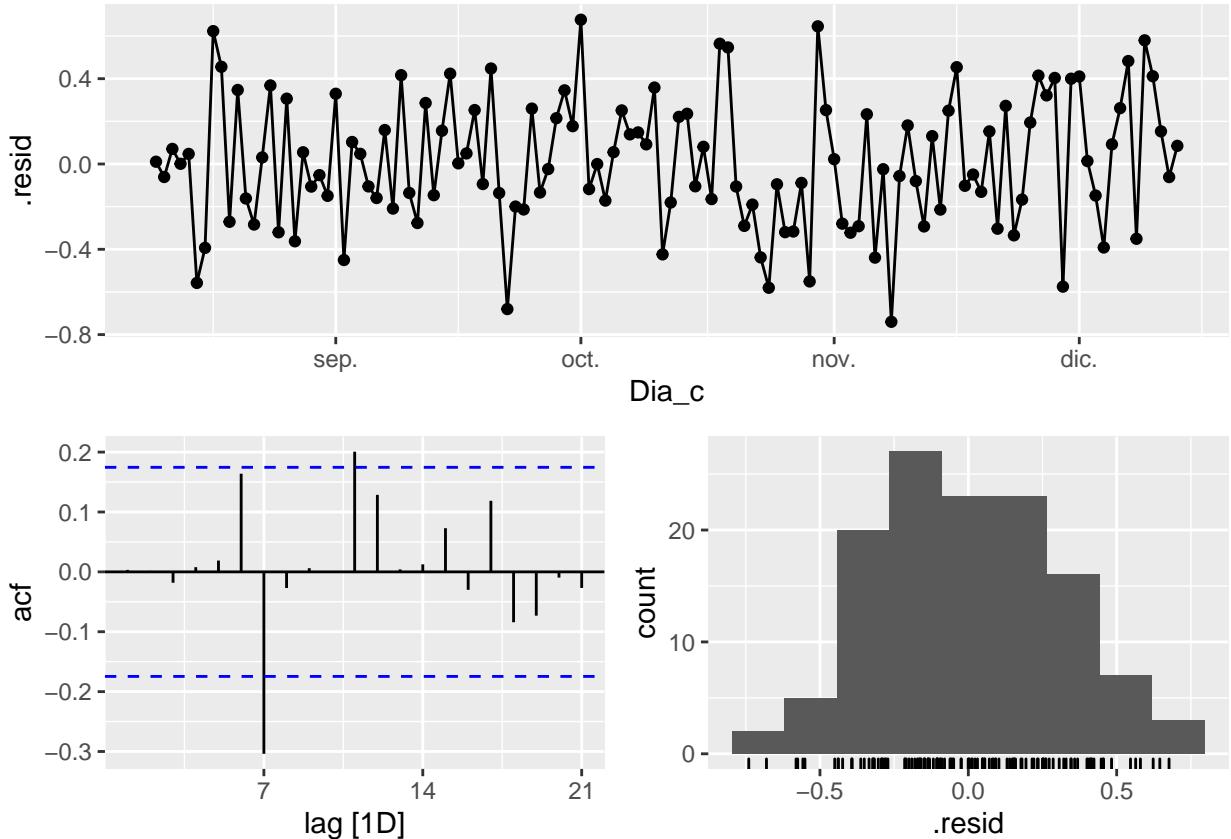
```

fit_model %>% select(arima_at1) %>% gg_tsresiduals()

```



```
fit_model %>% select(arima_at2) %>% gg_tsresiduals()
```



```

# Significant spikes out of 30 is still consistent with white noise.
# To be sure, use a Ljung-Box test, which has a large p-value, confirming that the
# residuals are similar to white noise.
# Note that the alternative models also pass this test.

# New data (dynamic regression)
# Here it is needed generate future values for the exogenous variables
# For simplicity we select a rand number included into the 2nd
# and 3rd quantile for the variable

# h7
Bar_N_cases_fr7 <- new_data(Bar_N_cases_tr, 7) %>%
  mutate(residential_percent_change_from_baseline =
    runif(7,quantile(Bar_N_cases_tt$residential_percent_change_from_baseline,
      0.25),
    quantile(Bar_N_cases_tt$residential_percent_change_from_baseline,
      0.75)),
  Total = runif(7,quantile(Bar_N_cases_tt$Total,0.25),
    quantile(Bar_N_cases_tt$Total,0.75)))

# h14
Bar_N_cases_fr14 <- new_data(Bar_N_cases_tr, 14) %>%
  mutate(residential_percent_change_from_baseline =
    runif(14,quantile(Bar_N_cases_tt$residential_percent_change_from_baseline,
      0.25),
    quantile(Bar_N_cases_tt$residential_percent_change_from_baseline,
      0.75)))
  
```

```

        0.75)),
Total = runif(14,quantile(Bar_N_cases_tt$Total,0.25),
               quantile(Bar_N_cases_tt$Total,0.75)))

# h17
Bar_N_cases_fr17 <- new_data(Bar_N_cases_tr, 17) %>%
  mutate(residential_percent_change_from_baseline =
    runif(17,quantile(Bar_N_cases_tt$residential_percent_change_from_baseline,
                      0.25),
           quantile(Bar_N_cases_tt$residential_percent_change_from_baseline,
                     0.75)),
  Total = runif(17,quantile(Bar_N_cases_tt$Total,0.25),
                quantile(Bar_N_cases_tt$Total,0.75)))

# Forecast
fc_fh7<-fabletools::forecast(fit_model, new_data = Bar_N_cases_fr7)
fc_fh14<-fabletools::forecast(fit_model, new_data = Bar_N_cases_fr14)
fc_fh17<-fabletools::forecast(fit_model, new_data = Bar_N_cases_fr17)

# Accuracy
fabletools::accuracy(fc_fh7, Bar_N_cases)

## # A tibble: 3 x 11
##   .model   sub_region_2 .type     ME   RMSE   MAE     MPE   MAPE   MASE RMSSE   ACF1
##   <chr>    <chr>     <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 arima_at1 Barcelona Test  65.4   232.  172.  3.10  13.7  0.759  0.605  0.211
## 2 arima_at2 Barcelona Test -6.98   197.  139. -0.574  9.80  0.612  0.512  0.278
## 3 SNaive    Barcelona Test 377.   442.  377.  28.5   28.5  1.66  1.15  0.113
fabletools::accuracy(fc_fh14, Bar_N_cases)

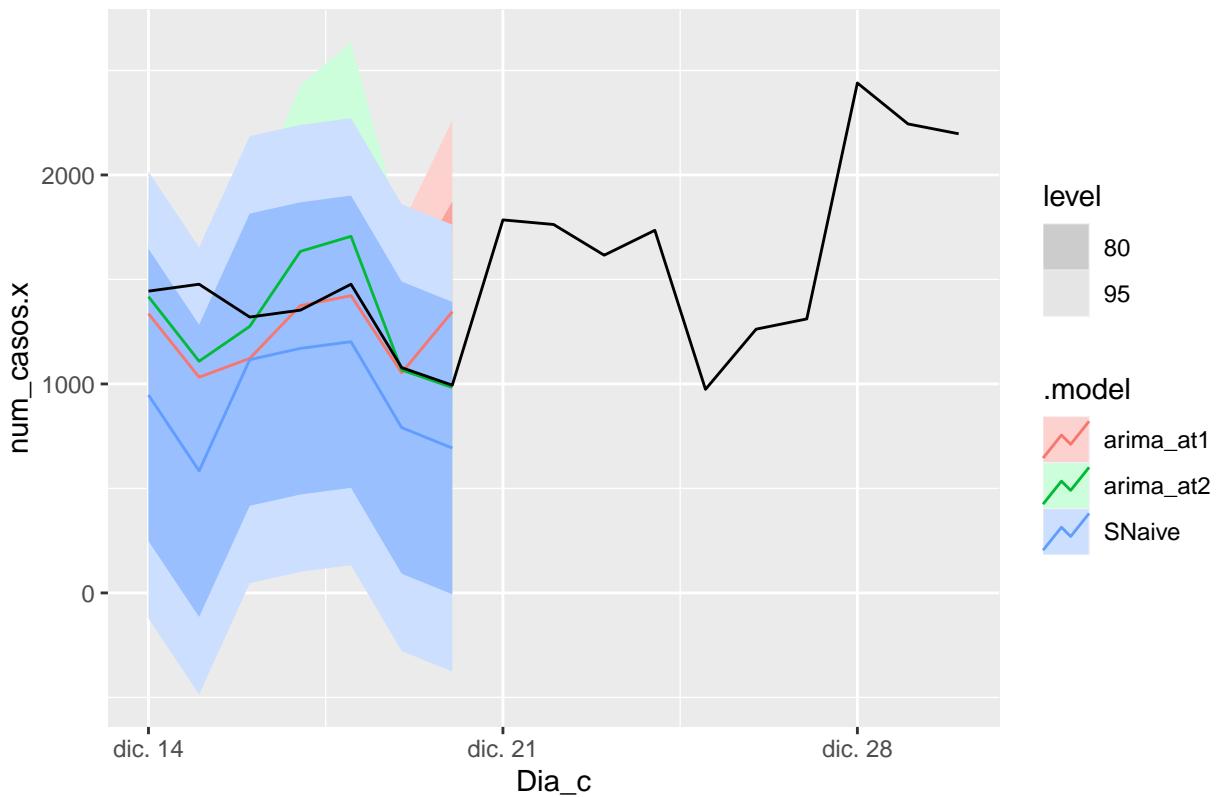
## # A tibble: 3 x 11
##   .model   sub_region_2 .type     ME   RMSE   MAE     MPE   MAPE   MASE RMSSE   ACF1
##   <chr>    <chr>     <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 arima_~ Barcelona Test  58.0   332.  282.  0.0463 20.7  1.24  0.865  0.269
## 2 arima_~ Barcelona Test -220.   449.  259. -18.9   21.6  1.14  1.17  0.466
## 3 SNaive    Barcelona Test 470.   579.  503.  31.3   34.6  2.22  1.51  0.160
fabletools::accuracy(fc_fh17, Bar_N_cases)

## # A tibble: 3 x 11
##   .model   sub_region_2 .type     ME   RMSE   MAE     MPE   MAPE   MASE RMSSE   ACF1
##   <chr>    <chr>     <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 arima_at1 Barcelona Test 143.   444.  341.  3.05  22.0  1.50  1.16  0.303
## 2 arima_at2 Barcelona Test -316.   517.  354. -23.3  25.7  1.56  1.35  0.276
## 3 SNaive    Barcelona Test 636.   798.  663.  36.6   39.3  2.92  2.08  0.495

# Plots
fc_fh7 %>%
  autoplot(Bar_N_cases_tt) +
  labs(title="Barcelona - forecast h7")

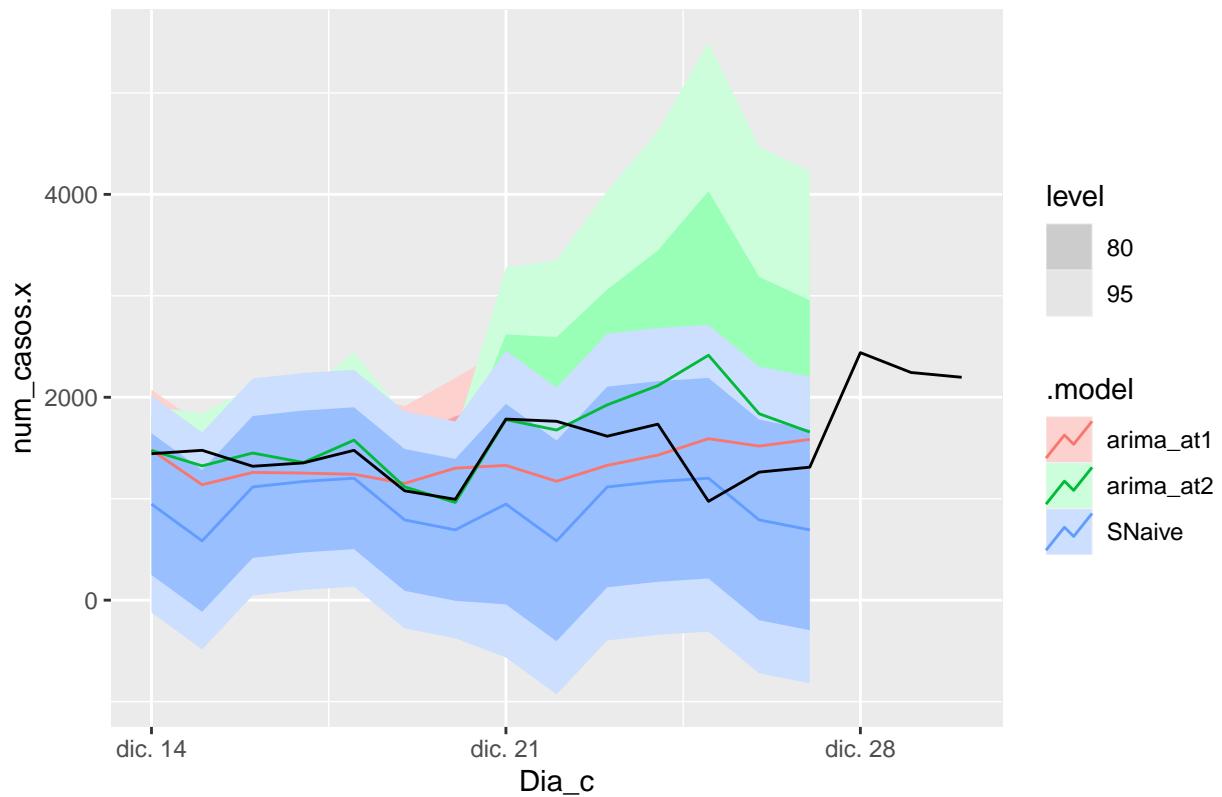
```

Barcelona – forecast h7



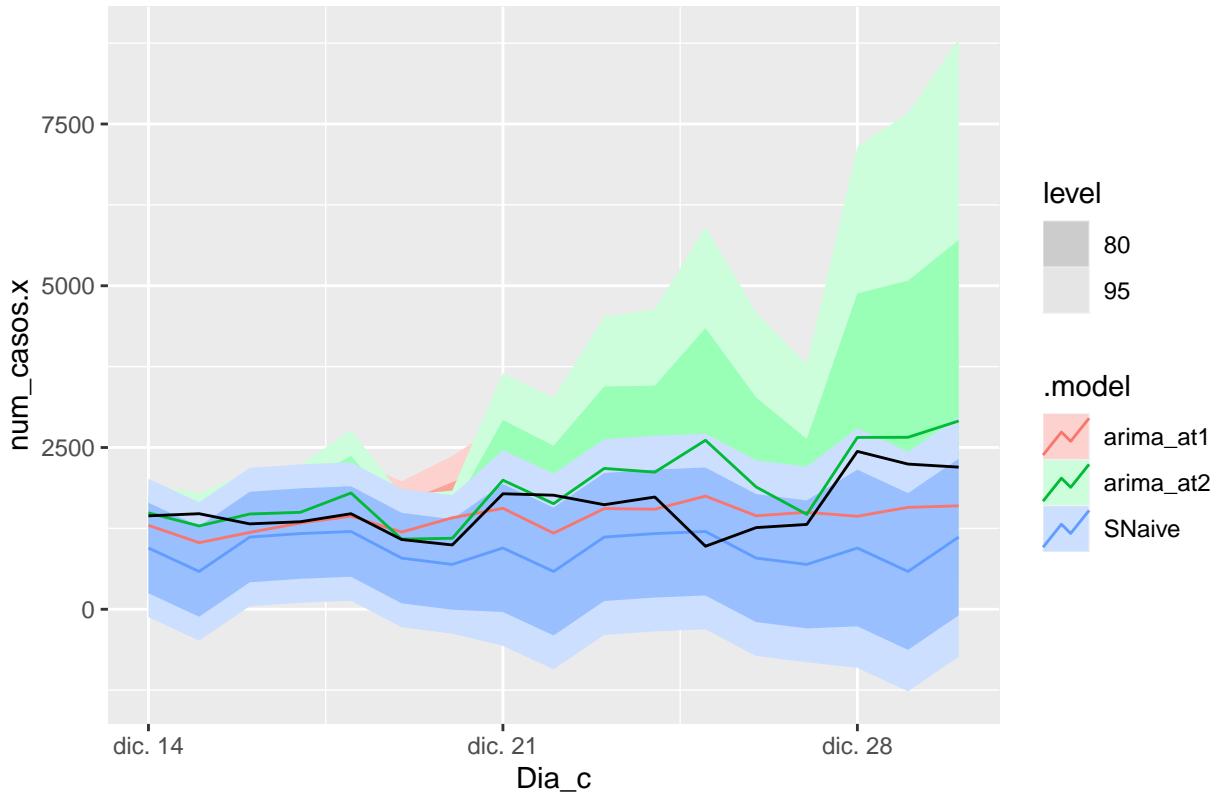
```
fc_fh14 %>%
  autoplot(Bar_N_cases_tt) +
  labs(title="Barcelona - forecast h14")
```

Barcelona – forecast h14



```
fc_fh17 %>%
  autoplot(Bar_N_cases_tt) +
  labs(title="Barcelona - forecast h17")
```

Barcelona – forecast h17



As stated by (Hyndman and Athanasopoulos 2021) (Sec 10)... “There is clear heteroscedasticity in the residuals... and lower variance in May. The model also has some significant autocorrelation in the residuals, and the histogram of the residuals shows long tails. All of these issues with the residuals may affect the coverage of the prediction intervals, but the point forecasts should still be ok.”

```
# Opt A
# We have added "residential_percent_change" and "Total" variables to models
lambda_mad_b <- Mad_N_cases %>%
  features(residential_percent_change_from_baseline, features = guerero) %>%
  pull(lambda_guerero)
lambda_mad_h <- Mad_N_cases %>%
  features(Total, features = guerero) %>%
  pull(lambda_guerero)

fit_model <- Mad_N_cases_tr %>%
  model(
    SNaive = SNAIVE(num_casos.x),
    arima_at1 = ARIMA(box_cox(num_casos.x,lambda_bar) ~
      box_cox(residential_percent_change_from_baseline,lambda_mad_b)+ 
      box_cox(Total,lambda_mad_h)),
    arima_at2 = ARIMA(box_cox(num_casos.x,lambda_bar) ~
      box_cox(residential_percent_change_from_baseline,lambda_mad_b)+ 
      box_cox(Total,lambda_mad_h) ,
      stepwise = FALSE, approx = FALSE))
```

```

# Show and report model
fit_model



### 3.3.2.2 Madrid



## # A mable: 1 x 4
## # Key:      sub_region_2 [1]
##   sub_region_2   SNaive                               arima_at1
##   <chr>          <model>                             <model>
## 1 Madrid        <SNAIVE> <LM w/ ARIMA(0,1,2)(2,1,0)[7] errors>
## # ... with 1 more variable: arima_at2 <model>

fit_model %>% select(arima_at1) %>% report()

## Series: num_casos.x
## Model: LM w/ ARIMA(0,1,2)(2,1,0)[7] errors
## Transformation: box_cox(num_casos.x, lambda_bar)
##
## Coefficients:
##       ma1      ma2      sar1      sar2
##       -0.2286  -0.1866  -0.2749  -0.4835
## s.e.  0.0980  0.0959  0.0920  0.0903
##       box_cox(residential_percent_change_from_baseline, lambda_mad_b)
##                                         -0.0496
##                                         0.0137
## s.e.
##       box_cox(Total, lambda_mad_h)
##                                         0.0836
## s.e.                      0.3188
##
## sigma^2 estimated as 0.09479: log likelihood=-27.41
## AIC=68.82  AICc=69.84  BIC=88.21
fit_model %>% select(arima_at2) %>% report()

## Series: num_casos.x
## Model: LM w/ ARIMA(0,1,2)(2,1,0)[7] errors
## Transformation: box_cox(num_casos.x, lambda_bar)
##
## Coefficients:
##       ma1      ma2      sar1      sar2
##       -0.2286  -0.1866  -0.2749  -0.4835
## s.e.  0.0980  0.0959  0.0920  0.0903
##       box_cox(residential_percent_change_from_baseline, lambda_mad_b)
##                                         -0.0496
##                                         0.0137
## s.e.
##       box_cox(Total, lambda_mad_h)
##                                         0.0836
## s.e.                      0.3188
##
## sigma^2 estimated as 0.09479: log likelihood=-27.41
## AIC=68.82  AICc=69.84  BIC=88.21

# Good model >> Less Sigma / AICc
fit_model %>% pivot_longer(!sub_region_2,
                           names_to = "Model name",
                           values_to = "Orders")

```

```

## # A mable: 3 x 3
## # Key:     sub_region_2, Model name [3]
##   sub_region_2 `Model name`          Orders
##   <chr>           <chr>            <model>
## 1 Madrid          SNaive           <SNAIVE>
## 2 Madrid          arima_at1      <LM w/ ARIMA(0,1,2)(2,1,0)[7] errors>
## 3 Madrid          arima_at2      <LM w/ ARIMA(0,1,2)(2,1,0)[7] errors>
glance(fit_model) %>% arrange(AICc) %>% select(.model:BIC)

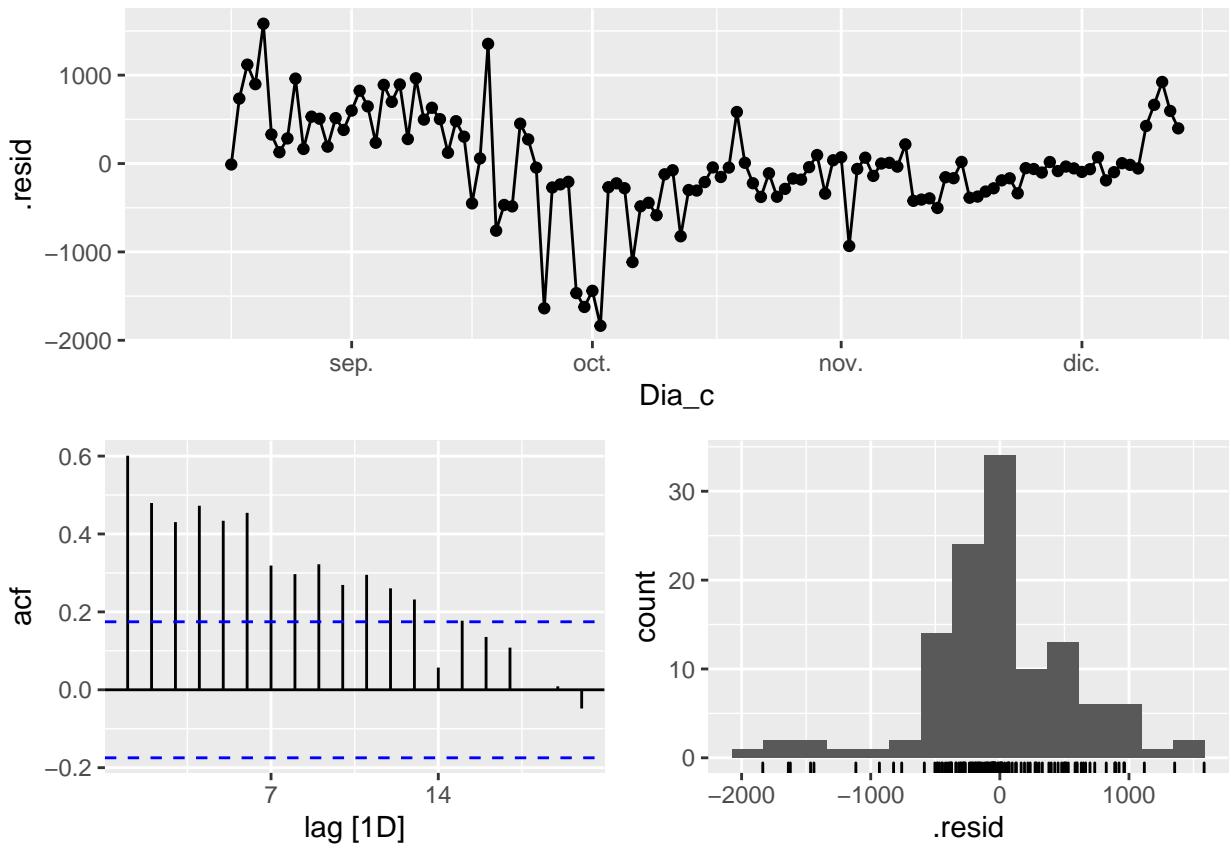
## # A tibble: 3 x 6
##   .model      sigma2 log_lik   AIC  AICc    BIC
##   <chr>      <dbl>   <dbl> <dbl> <dbl> <dbl>
## 1 arima_at1  0.0948 -27.4  68.8  69.8  88.2
## 2 arima_at2  0.0948 -27.4  68.8  69.8  88.2
## 3 SNaive    322923.     NA     NA     NA     NA
# We use a Ljung-Box test >> large p-value, confirms residuals are similar to white noise.
augment(fit_model) %>%
  features(.innov, lag=7)

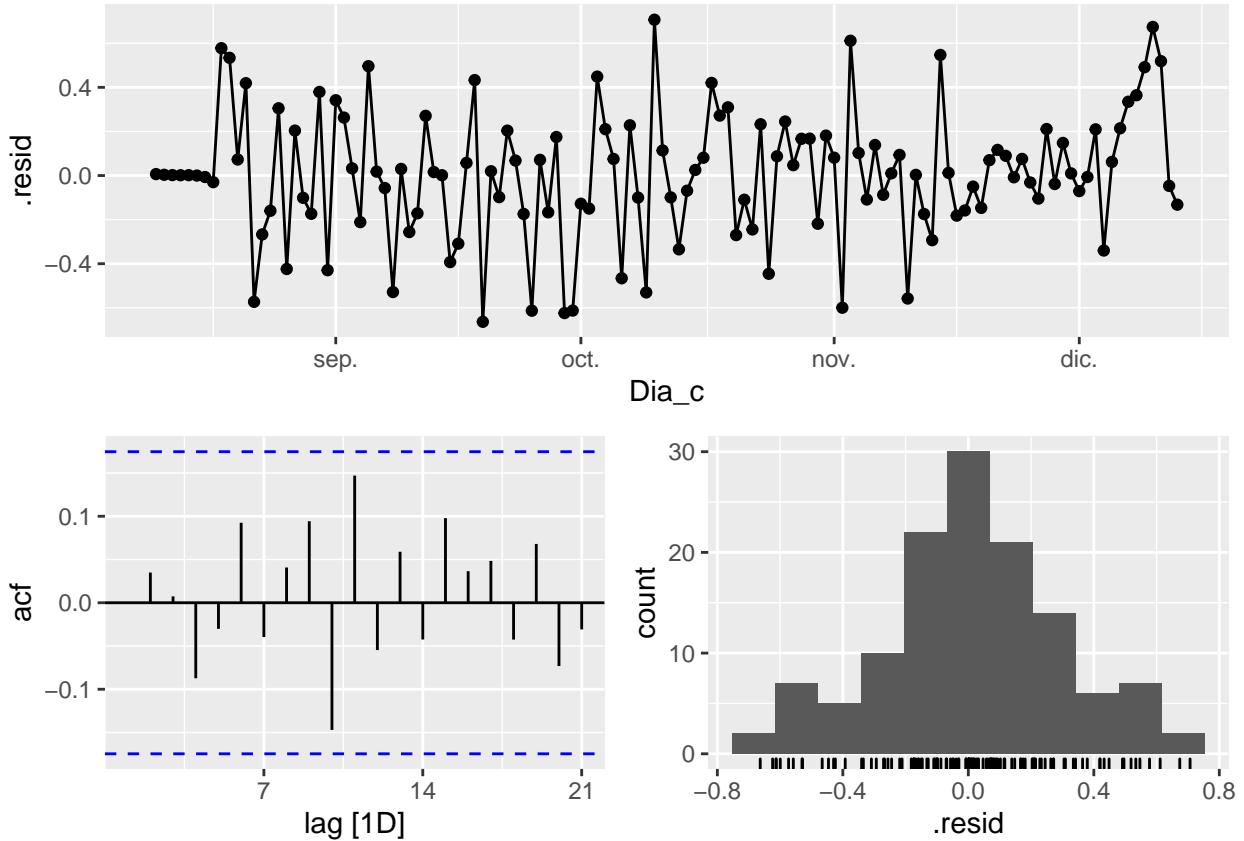
## # A tibble: 3 x 4
##   sub_region_2 .model    lb_stat lb_pvalue
##   <chr>        <chr>    <dbl>     <dbl>
## 1 Madrid       arima_at1  2.66     0.915
## 2 Madrid       arima_at2  2.66     0.915
## 3 Madrid       SNaive    186.      0
augment(fit_model) %>%
  features(.innov, ljung_box, lag=14)

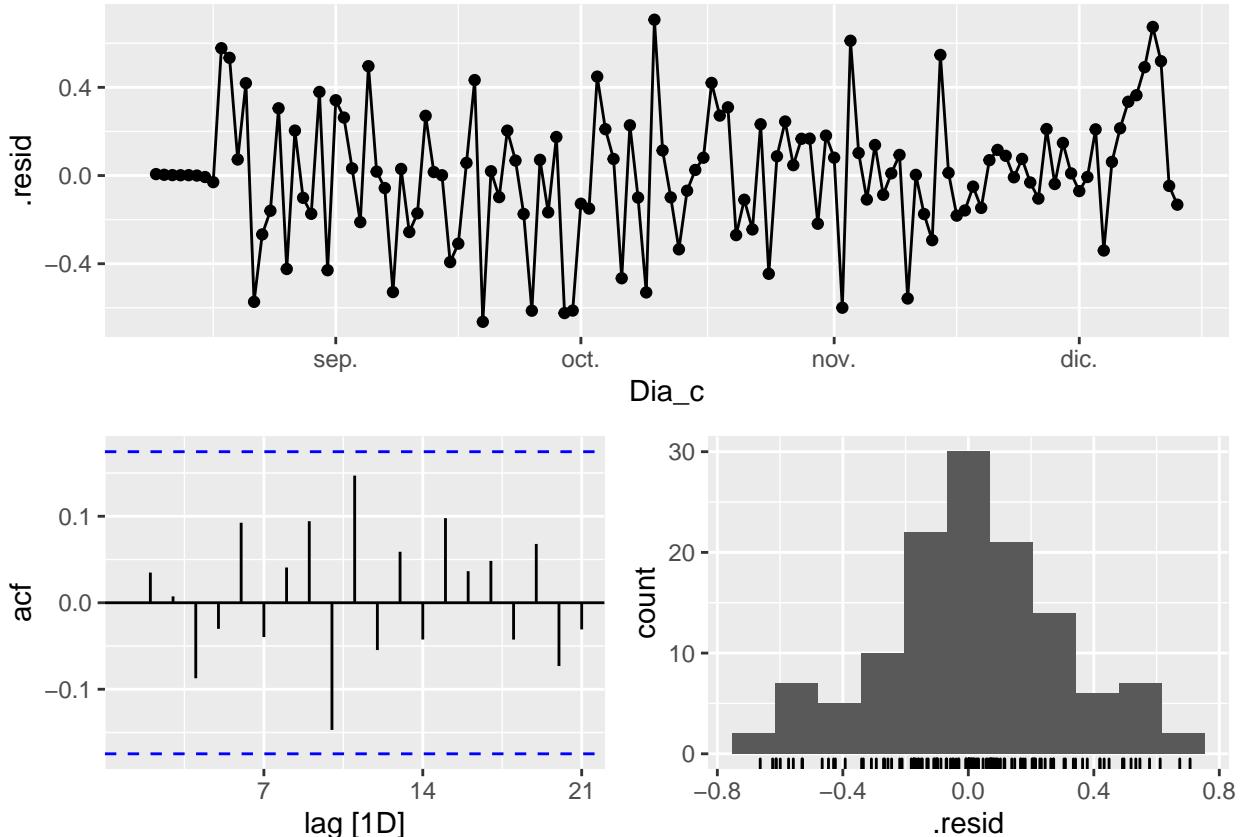
## # A tibble: 3 x 4
##   sub_region_2 .model    lb_stat lb_pvalue
##   <chr>        <chr>    <dbl>     <dbl>
## 1 Madrid       arima_at1  11.3     0.660
## 2 Madrid       arima_at2  11.3     0.660
## 3 Madrid       SNaive    250.      0
augment(fit_model) %>%
  features(.innov, ljung_box, lag=21)

## # A tibble: 3 x 4
##   sub_region_2 .model    lb_stat lb_pvalue
##   <chr>        <chr>    <dbl>     <dbl>
## 1 Madrid       arima_at1  15.2     0.814
## 2 Madrid       arima_at2  15.2     0.814
## 3 Madrid       SNaive    259.      0
fit_model %>% select(SNaive) %>% gg_tsresiduals()

```







```

# Significant spikes out of 30 is still consistent with white noise.
# To be sure, use a Ljung-Box test, which has a large p-value, confirming that the
# residuals are similar to white noise.
# Note that the alternative models also pass this test.

# New data (dynamic regression)
# Here it is needed generate future values for the exogenous variables
# For simplicity we select a rand number included into the 2nd
# and 3rd quantile for the variable

# h7
Mad_N_cases_fr7 <- new_data(Mad_N_cases_tr, 7) %>%
  mutate(residential_percent_change_from_baseline =
    runif(7,quantile(Mad_N_cases_tt$residential_percent_change_from_baseline,
                      0.25),
          quantile(Mad_N_cases_tt$residential_percent_change_from_baseline,
                      0.75)),
    Total = runif(7,quantile(Mad_N_cases_tt$Total,0.25),
                  quantile(Mad_N_cases_tt$Total,0.75)))

# h14
Mad_N_cases_fr14 <- new_data(Mad_N_cases_tr, 14) %>%
  mutate(residential_percent_change_from_baseline =
    runif(14,quantile(Mad_N_cases_tt$residential_percent_change_from_baseline,
                      0.25),
          quantile(Mad_N_cases_tt$residential_percent_change_from_baseline,
                      0.75)))

```

```

        0.75)),
Total = runif(14,quantile(Mad_N_cases_tt$Total,0.25),
               quantile(Mad_N_cases_tt$Total,0.75))

# h17
Mad_N_cases_fr17 <- new_data(Mad_N_cases_tr, 17) %>%
  mutate(residential_percent_change_from_baseline =
    runif(17,quantile(Mad_N_cases_tt$residential_percent_change_from_baseline,
                      0.25),
           quantile(Mad_N_cases_tt$residential_percent_change_from_baseline,
                     0.75)),
  Total = runif(17,quantile(Mad_N_cases_tt$Total,0.25),
                quantile(Mad_N_cases_tt$Total,0.75)))

# Forecast
fc_fh7<-fabletools::forecast(fit_model, new_data = Mad_N_cases_fr7)
fc_fh14<-fabletools::forecast(fit_model, new_data = Mad_N_cases_fr14)
fc_fh17<-fabletools::forecast(fit_model, new_data = Mad_N_cases_fr17)

# Accuracy
fabletools::accuracy(fc_fh7, Mad_N_cases)

## # A tibble: 3 x 11
##   .model   sub_region_2 .type     ME   RMSE    MAE    MPE    MAPE    MASE   RMSSE   ACF1
##   <chr>    <chr>     <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 arima_at1 Madrid     Test -282.  360.  300. -16.5  17.5  1.02  0.785  0.450
## 2 arima_at2 Madrid     Test -282.  360.  300. -16.5  17.5  1.02  0.785  0.450
## 3 SNaive    Madrid     Test  354.  509.  356.  18.0  18.1  1.21  1.11  0.584

fabletools::accuracy(fc_fh14, Mad_N_cases)

## # A tibble: 3 x 11
##   .model   sub_region_2 .type     ME   RMSE    MAE    MPE    MAPE    MASE   RMSSE   ACF1
##   <chr>    <chr>     <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 arima_at1 Madrid     Test -491.  649.  501. -31.8  32.3  1.69  1.41  0.388
## 2 arima_at2 Madrid     Test -491.  649.  501. -31.8  32.3  1.69  1.41  0.388
## 3 SNaive    Madrid     Test  310.  570.  474.  13.6  26.6  1.60  1.24  0.438

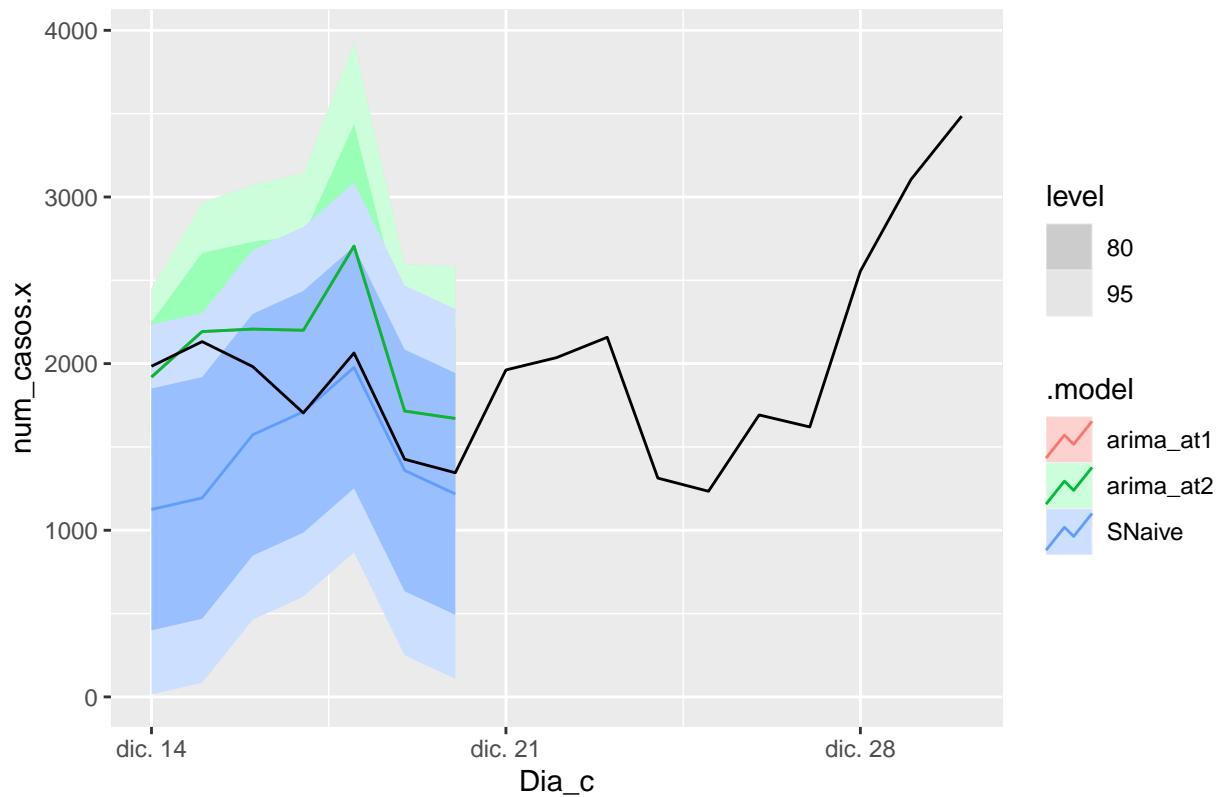
fabletools::accuracy(fc_fh17, Mad_N_cases)

## # A tibble: 3 x 11
##   .model   sub_region_2 .type     ME   RMSE    MAE    MPE    MAPE    MASE   RMSSE   ACF1
##   <chr>    <chr>     <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 arima_at1 Madrid     Test -421.  599.  463. -26.8  28.3  1.57  1.31  0.371
## 2 arima_at2 Madrid     Test -421.  599.  463. -26.8  28.3  1.57  1.31  0.371
## 3 SNaive    Madrid     Test  564.  904.  699.  21.4  32.1  2.37  1.97  0.618

# Plots
fc_fh7 %>%
  autoplot(Mad_N_cases_tt) +
  labs(title="Madrid - forecast h7")

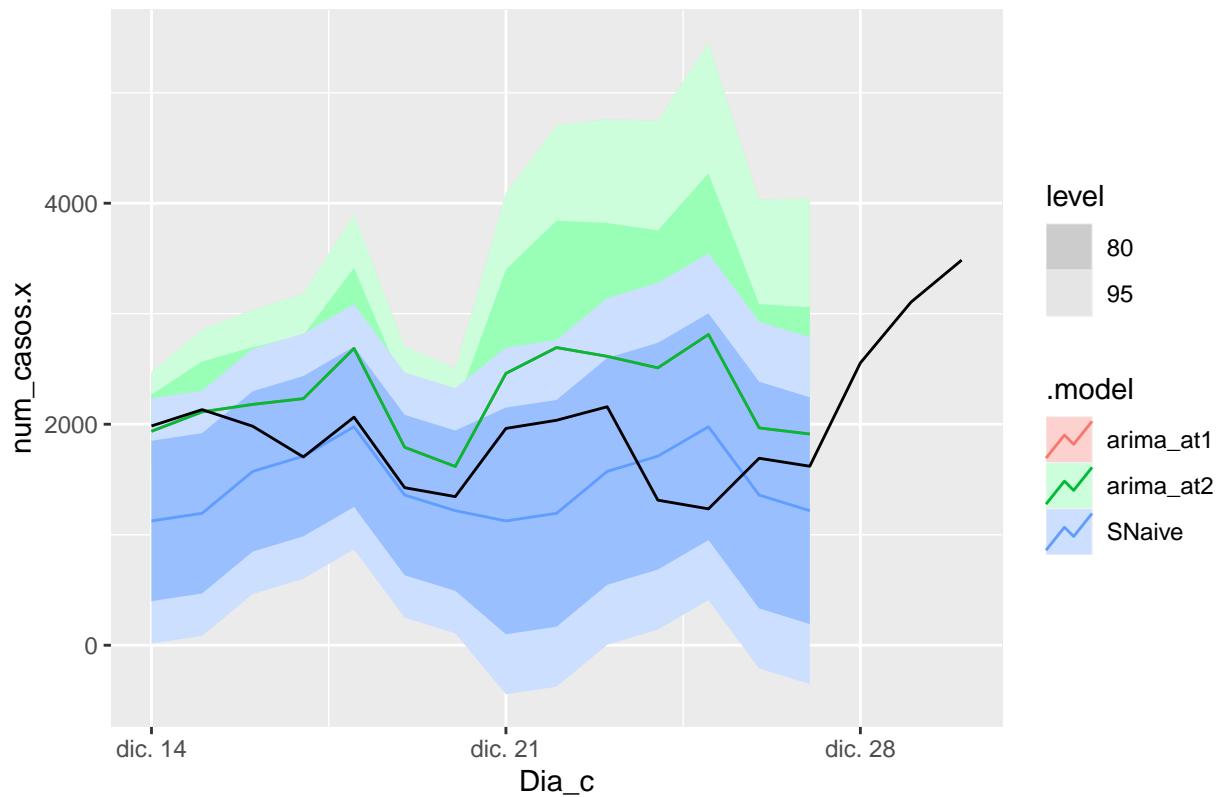
```

Madrid – forecast h7



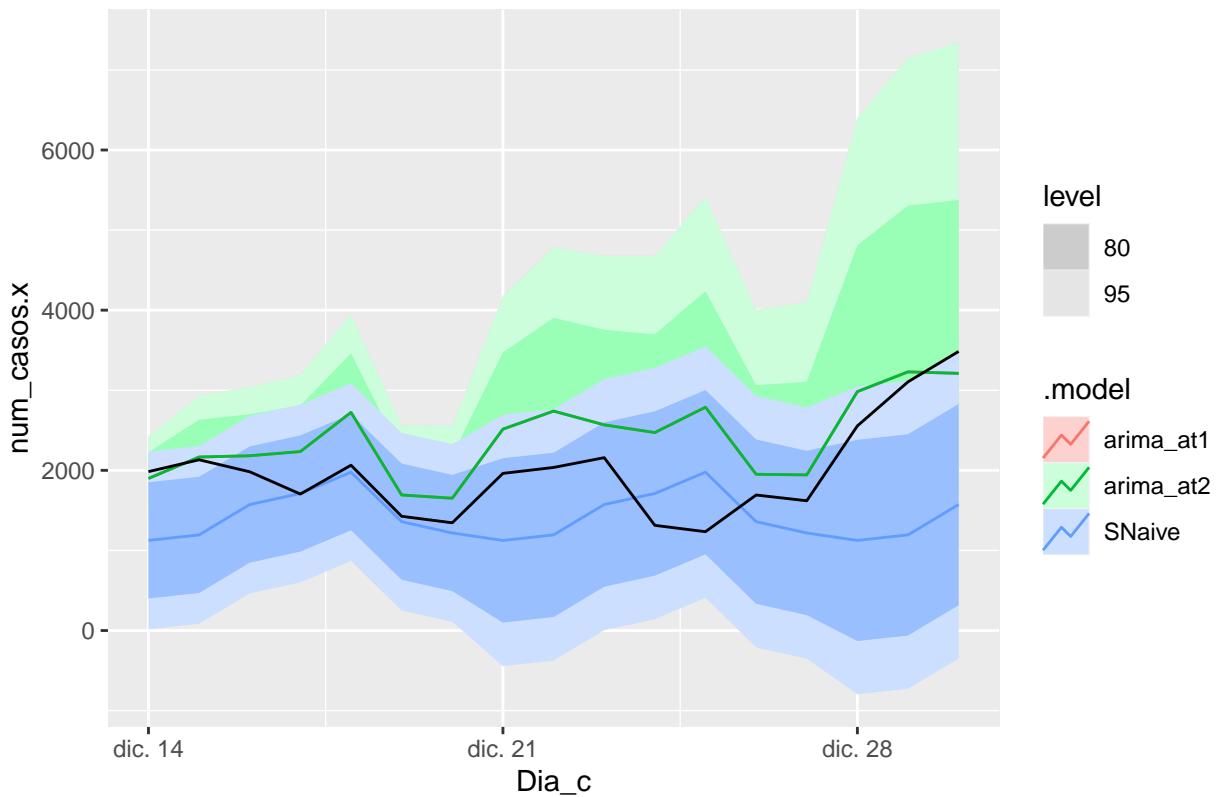
```
fc_fh14 %>%
  autoplot(Mad_N_cases_tt) +
  labs(title="Madrid - forecast h14")
```

Madrid – forecast h14



```
fc_fh17 %>%
  autoplot(Mad_N_cases_tt) +
  labs(title="Madrid - forecast h17")
```

Madrid – forecast h17



```
# Opt A
# We have added "residential_percent_change" variable to models
lambda_mal_b <- Mal_N_cases %>%
  features(residential_percent_change_from_baseline, features = guerero) %>%
  pull(lambda_guerero)
lambda_mal_h <- Mal_N_cases %>%
  features(Total, features = guerero) %>%
  pull(lambda_guerero)

fit_model <- Mal_N_cases_tr %>%
  model(
    SNaive = SNAIVE(num_casos.x),
    arima_at1 = ARIMA(box_cox(num_casos.x,lambda_mal) ~
      box_cox(residential_percent_change_from_baseline,lambda_mal_b)+
      box_cox(Total,lambda_mal_h)),
    arima_at2 = ARIMA(box_cox(num_casos.x,lambda_bar) ~
      box_cox(residential_percent_change_from_baseline,lambda_mal_b)+
      box_cox(Total,lambda_mal_h) ,
      stepwise = FALSE, approx = FALSE))

# Show and report model
fit_model
```

3.3.2.3 Málaga

```

## # A mable: 1 x 4
## # Key:      sub_region_2 [1]
##   sub_region_2    SNaive                               arima_at1
##   <chr>          <model>                               <model>
## 1 Málaga        <SNAIVE> <LM w/ ARIMA(1,0,1)(2,0,0)[7] errors>
## # ... with 1 more variable: arima_at2 <model>
fit_model %>% select(arima_at1) %>% report()

## Series: num_casos.x
## Model: LM w/ ARIMA(1,0,1)(2,0,0)[7] errors
## Transformation: box_cox(num_casos.x, lambda_mal)
##
## Coefficients:
##             ar1      ma1      sar1      sar2
##             0.9570 -0.5447  0.4181  0.2534
## s.e.     0.0267  0.0807  0.0928  0.0939
##             box_cox(residential_percent_change_from_baseline, lambda_mal_b)
##                                         -0.2010
## s.e.           0.0603
##             box_cox(Total, lambda_mal_h)  intercept
##                                         0.1975    9.9129
## s.e.           0.0613    2.3287
##
## sigma^2 estimated as 0.7886: log likelihood=-163.26
## AIC=342.53  AICc=343.76  BIC=365.22
fit_model %>% select(arima_at2) %>% report()

## Series: num_casos.x
## Model: LM w/ ARIMA(1,0,3)(2,0,0)[7] errors
## Transformation: box_cox(num_casos.x, lambda_bar)
##
## Coefficients:
##             ar1      ma1      ma2      ma3      sar1      sar2
##             0.9540 -0.5709 -0.1776  0.2524  0.4199  0.2233
## s.e.     0.0284  0.0976  0.0933  0.0954  0.0933  0.0944
##             box_cox(residential_percent_change_from_baseline, lambda_mal_b)
##                                         -0.0817
## s.e.           0.0198
##             box_cox(Total, lambda_mal_h)  intercept
##                                         0.0775    6.0554
## s.e.           0.0181    0.7513
##
## sigma^2 estimated as 0.09545: log likelihood=-29.11
## AIC=78.22  AICc=80.14  BIC=106.59
# Good model >> Less Sigma / AICc
fit_model %>% pivot_longer(!sub_region_2,
                           names_to = "Model name",
                           values_to = "Orders")

## # A mable: 3 x 3
## # Key:      sub_region_2, Model name [3]
##   sub_region_2 `Model name`                               Orders
##   <chr>          <chr>                               <model>

```

```

## 1 Málaga      SNaive          <SNAIVE>
## 2 Málaga      arima_at1     <LM w/ ARIMA(1,0,1)(2,0,0)[7] errors>
## 3 Málaga      arima_at2     <LM w/ ARIMA(1,0,3)(2,0,0)[7] errors>
glance(fit_model) %>% arrange(AICc) %>% select(.model:BIC)

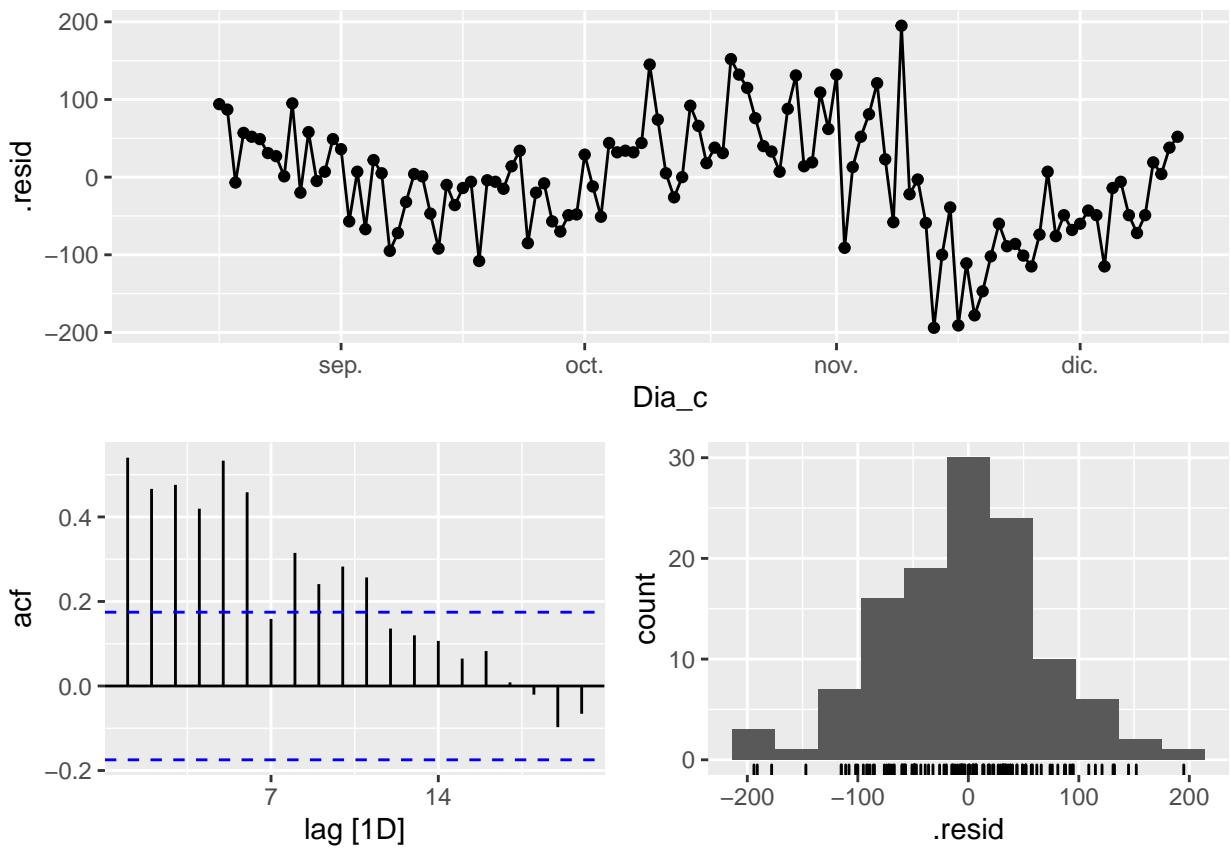
## # A tibble: 3 x 6
##   .model      sigma2 log_lik    AIC   AICc     BIC
##   <chr>      <dbl>    <dbl> <dbl> <dbl> <dbl>
## 1 arima_at2  0.0954   -29.1  78.2  80.1  107.
## 2 arima_at1  0.789    -163.  343.  344.  365.
## 3 SNaive     5268.      NA     NA     NA     NA
# We use a Ljung-Box test >> large p-value, confirms residuals are similar to white noise.
augment(fit_model) %>%
  features(.innov, lag=7)

## # A tibble: 3 x 4
##   sub_region_2 .model    lb_stat lb_pvalue
##   <chr>        <chr>      <dbl>     <dbl>
## 1 Málaga      arima_at1  12.0     0.102
## 2 Málaga      arima_at2  4.87     0.676
## 3 Málaga      SNaive     178.      0
augment(fit_model) %>%
  features(.innov, lag=14)

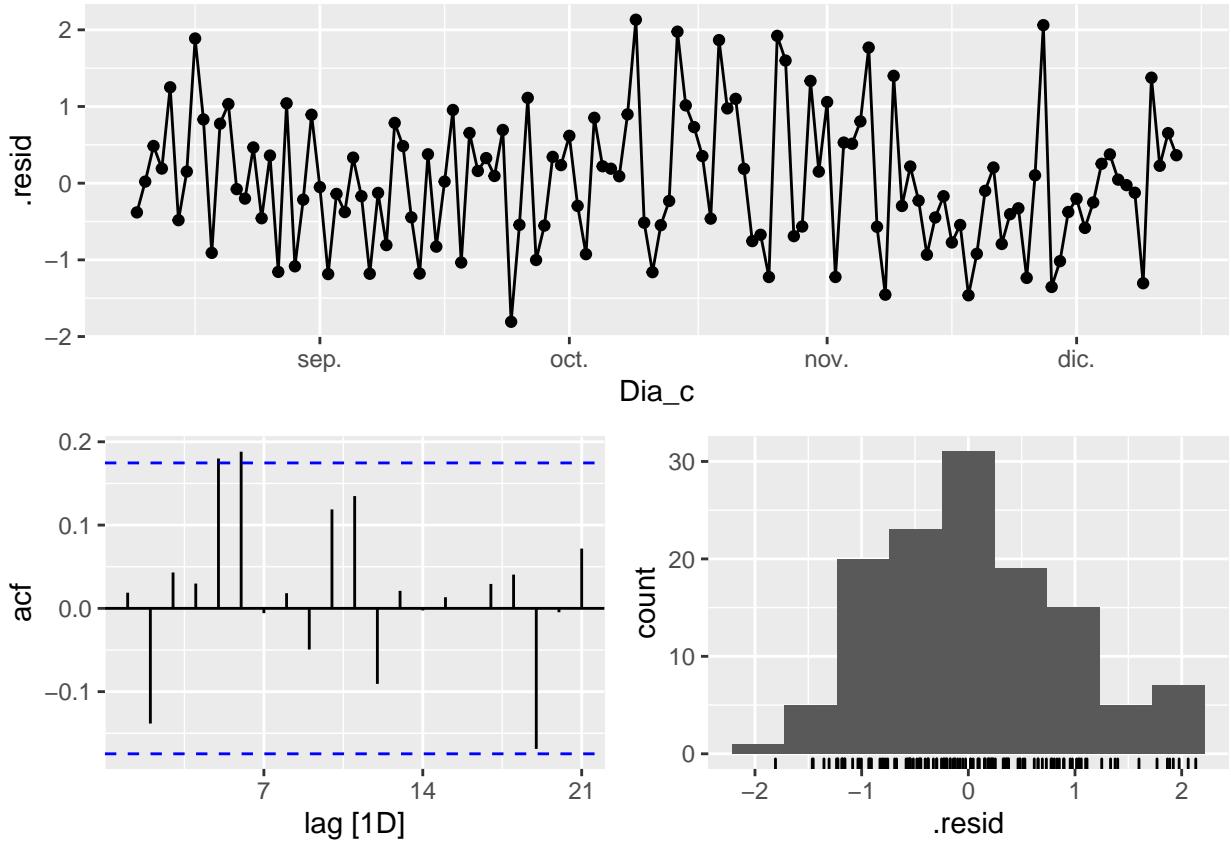
## # A tibble: 3 x 4
##   sub_region_2 .model    lb_stat lb_pvalue
##   <chr>        <chr>      <dbl>     <dbl>
## 1 Málaga      arima_at1  18.1     0.203
## 2 Málaga      arima_at2  9.49     0.799
## 3 Málaga      SNaive     224.      0
augment(fit_model) %>%
  features(.innov, lag=21)

## # A tibble: 3 x 4
##   sub_region_2 .model    lb_stat lb_pvalue
##   <chr>        <chr>      <dbl>     <dbl>
## 1 Málaga      arima_at1  23.6     0.314
## 2 Málaga      arima_at2  12.9     0.911
## 3 Málaga      SNaive     229.      0
fit_model %>% select(SNaive) %>% gg_tsresiduals()

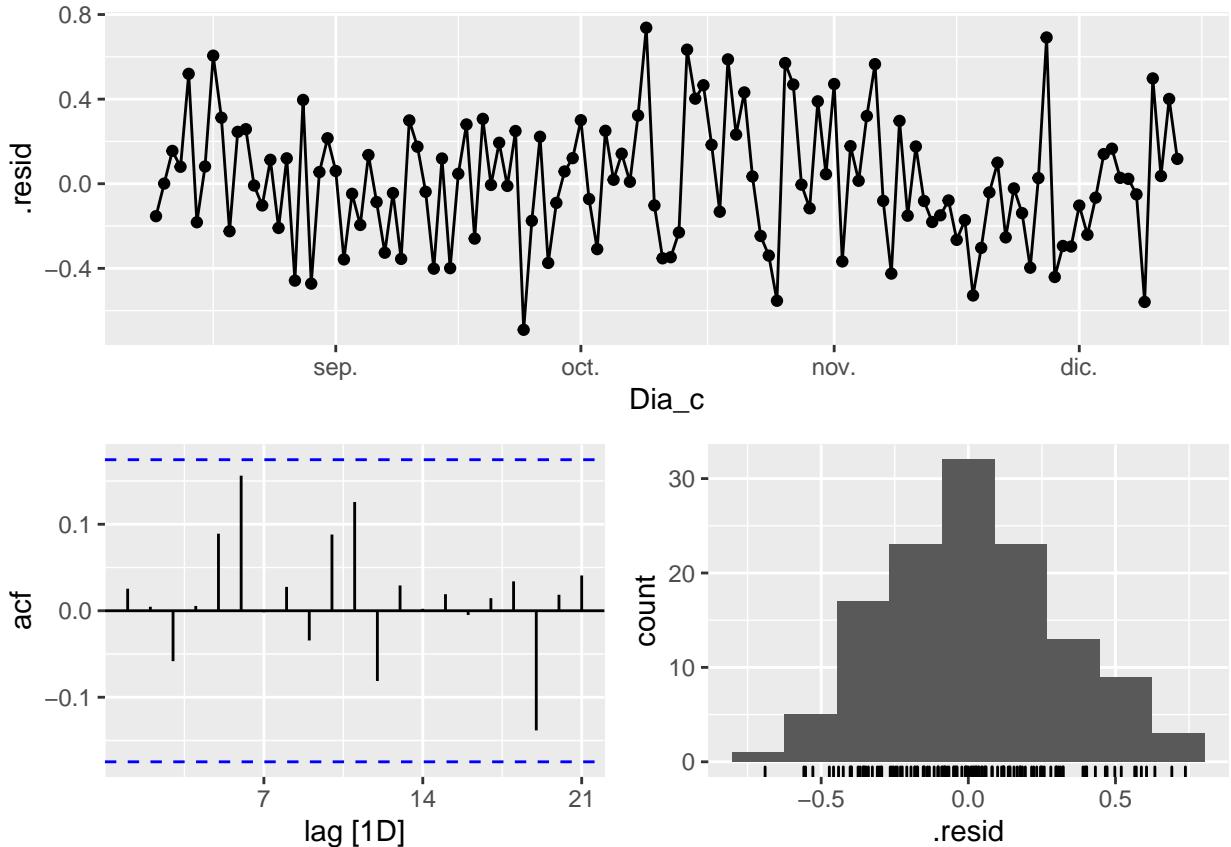
```



```
fit_model %>% select(arima_at1) %>% gg_tsresiduals()
```



```
fit_model %>% select(arima_at2) %>% gg_tsresiduals()
```



```
# Significant spikes out of 30 is still consistent with white noise.
# To be sure, use a Ljung-Box test, which has a large p-value, confirming that the
# residuals are similar to white noise.
# Note that the alternative models also pass this test.
```

```
# New data (dynamic regression)
# Here it is needed generate future values for the exogenous variables
# For simplicity we select a rand number included into the 2nd
# and 3rd quantile for the variable

# h7
Mal_N_cases_fr7 <- new_data(Mal_N_cases_tr, 7) %>%
  mutate(residential_percent_change_from_baseline =
    runif(7,quantile(Mal_N_cases_tt$residential_percent_change_from_baseline,
    0.25),
    quantile(Mal_N_cases_tt$residential_percent_change_from_baseline,
    0.75)),
  Total = runif(7,quantile(Mal_N_cases_tt$Total,0.25),
  quantile(Mal_N_cases_tt$Total,0.75)))

# h14
Mal_N_cases_fr14 <- new_data(Mal_N_cases_tr, 14) %>%
  mutate(residential_percent_change_from_baseline =
    runif(14,quantile(Mal_N_cases_tt$residential_percent_change_from_baseline,
    0.25),
    quantile(Mal_N_cases_tt$residential_percent_change_from_baseline,
```

```

          0.75)),
Total = runif(14,quantile(Mal_N_cases_tt$Total,0.25),
               quantile(Mal_N_cases_tt$Total,0.75))

# h17
Mal_N_cases_fr17 <- new_data(Mal_N_cases_tr, 17) %>%
  mutate(residential_percent_change_from_baseline =
    runif(17,quantile(Mal_N_cases_tt$residential_percent_change_from_baseline,
                      0.25),
           quantile(Mal_N_cases_tt$residential_percent_change_from_baseline,
                     0.75)),
  Total = runif(17,quantile(Mal_N_cases_tt$Total,0.25),
                quantile(Mal_N_cases_tt$Total,0.75)))

# Forecast
fc_fh7<-fabletools::forecast(fit_model, new_data = Mal_N_cases_fr7)
fc_fh14<-fabletools::forecast(fit_model, new_data = Mal_N_cases_fr14)
fc_fh17<-fabletools::forecast(fit_model, new_data = Mal_N_cases_fr17)

# Accuracy
fabletools::accuracy(fc_fh7, Mal_N_cases)

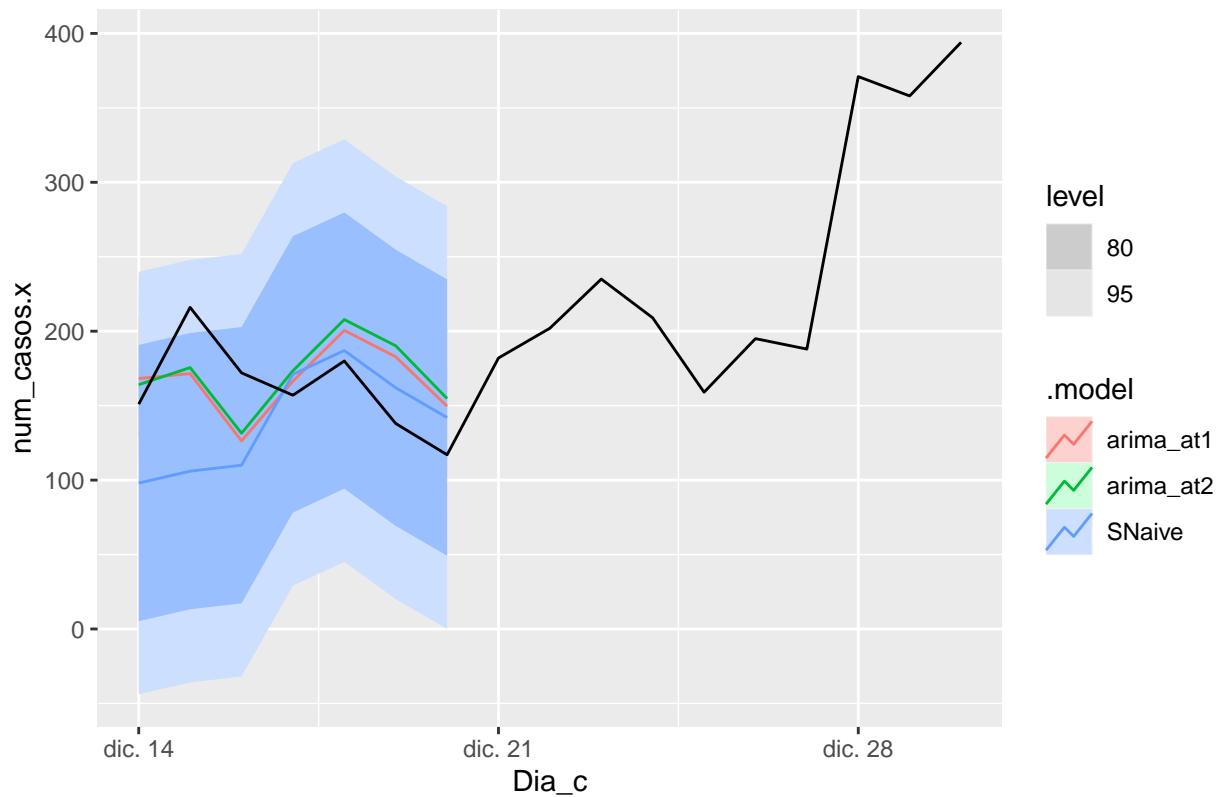
## # A tibble: 3 x 11
##   .model    sub_region_2 .type     ME   RMSE    MAE    MPE    MAPE    MASE   RMSSE   ACF1
##   <chr>      <chr>     <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 arima_at1 Málaga     Test  -4.92  33.7  30.7 -6.00  19.5  0.874  0.620  0.445
## 2 arima_at2 Málaga     Test  -9.48  35.1  32.6 -8.92  21.0  0.929  0.647  0.510
## 3 SNaive     Málaga     Test   22.1  53.7  42.1 10.1   24.8  1.20   0.989  0.557
fabletools::accuracy(fc_fh14, Mal_N_cases)

## # A tibble: 3 x 11
##   .model    sub_region_2 .type     ME   RMSE    MAE    MPE    MAPE    MASE   RMSSE   ACF1
##   <chr>      <chr>     <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 arima_at1 Málaga     Test   9.81  41.5  33.6  1.92  18.9  0.957  0.765  0.301
## 2 arima_at2 Málaga     Test   3.64  40.0  31.8 -1.55  18.4  0.905  0.736  0.316
## 3 SNaive     Málaga     Test  39.2  64.1  53.2 18.5   28.4  1.52   1.18   0.436
fabletools::accuracy(fc_fh17, Mal_N_cases)

## # A tibble: 3 x 11
##   .model    sub_region_2 .type     ME   RMSE    MAE    MPE    MAPE    MASE   RMSSE   ACF1
##   <chr>      <chr>     <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 arima_at1 Málaga     Test  44.3  93.1  63.4 11.4   25.1  1.81   1.71  0.544
## 2 arima_at2 Málaga     Test  37.5  88.8  60.1  8.12   24.2  1.71   1.63  0.543
## 3 SNaive     Málaga     Test  79.9 127.   91.4 28.0   36.1  2.61   2.35  0.595
# Plots
fc_fh7 %>%
  autoplot(Mal_N_cases_tt) +
  labs(title="Málaga - forecast h7")

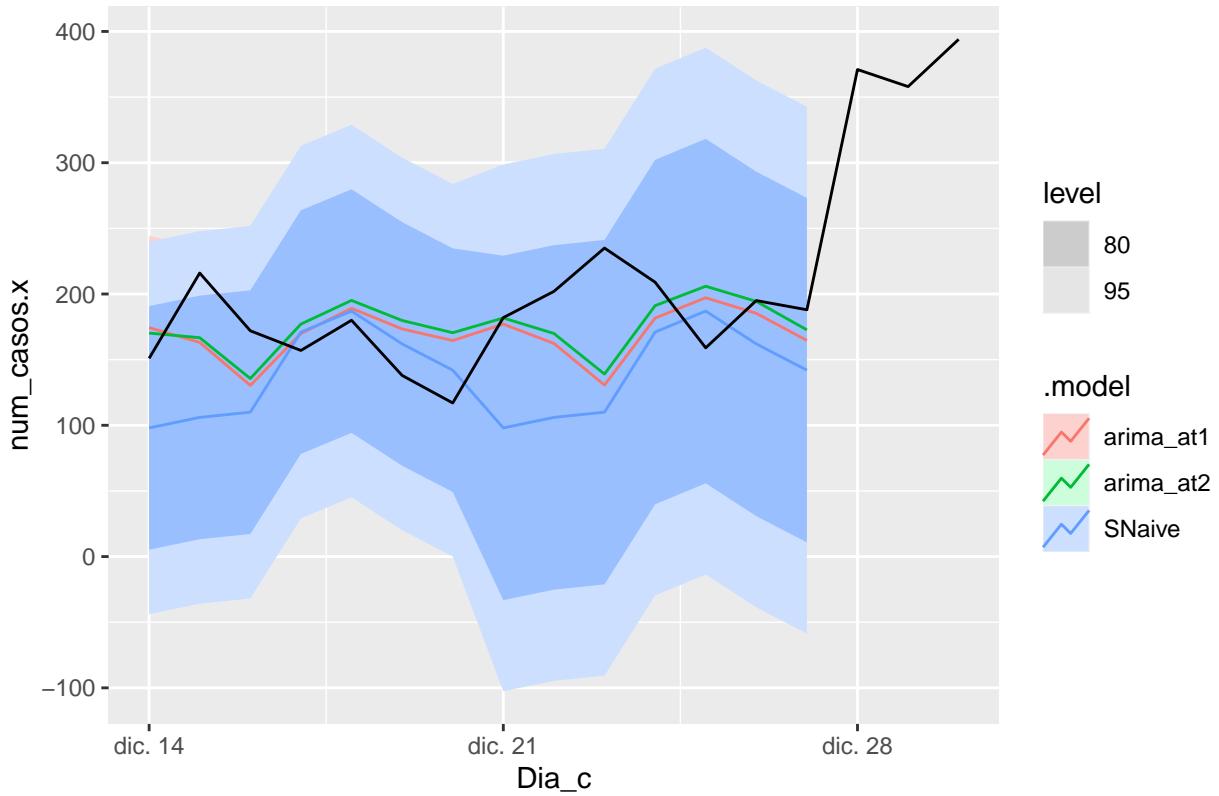
```

Málaga – forecast h7



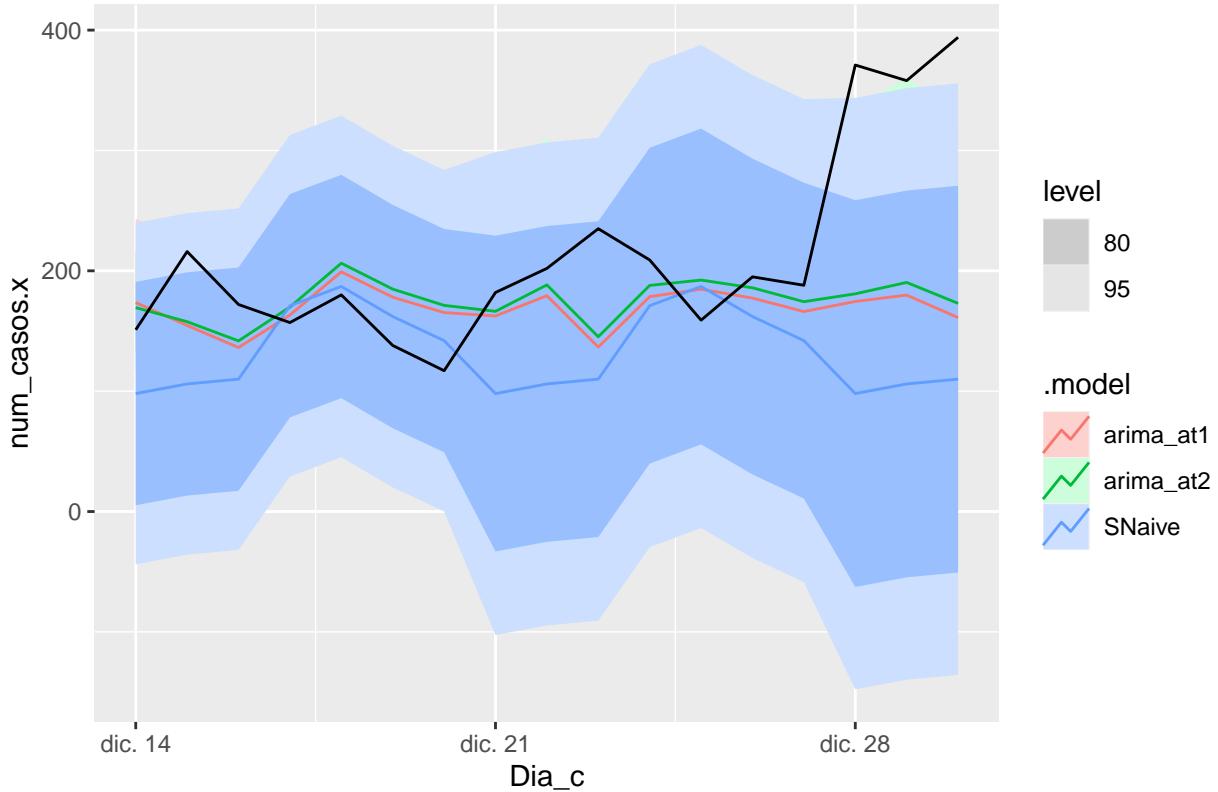
```
fc_fh14 %>%
  autoplot(Mal_N_cases_tt) +
  labs(title="Málaga - forecast h14")
```

Málaga – forecast h14



```
fc_fh17 %>%
  autoplot(Mal_N_cases_tt) +
  labs(title="Málaga - forecast h17")
```

Málaga – forecast h17



```
# Opt A
# We have added "residential_percent_change" variable to models
lambda_cad_b <- Cad_N_cases %>%
  features(residential_percent_change_from_baseline, features = guerero) %>%
  pull(lambda_guerrero)
lambda_cad_h <- Cad_N_cases %>%
  features(Total, features = guerero) %>%
  pull(lambda_guerrero)

fit_model <- Cad_N_cases_tr %>%
  model(
    SNaive = SNAIVE(num_casos.x),
    arima_at1 = ARIMA(box_cox(num_casos.x,lambda_cad) ~
      box_cox(residential_percent_change_from_baseline,lambda_cad_b)+ 
      box_cox(Total,lambda_cad_h)),
    arima_at2 = ARIMA(box_cox(num_casos.x,lambda_bar) ~
      box_cox(residential_percent_change_from_baseline,lambda_cad_b)+ 
      box_cox(Total,lambda_cad_h) ,
      stepwise = FALSE, approx = FALSE))

# Show and report model
fit_model
```

3.3.2.4 Cádiz

```

## # A mable: 1 x 4
## # Key:      sub_region_2 [1]
##   sub_region_2    SNaive                      arima_at1
##   <chr>          <model>                     <model>
## 1 Cádiz         <SNAIVE> <LM w/ ARIMA(0,1,1)(1,0,0)[7] errors>
## # ... with 1 more variable: arima_at2 <model>
fit_model %>% select(arima_at1) %>% report()

## Series: num_casos.x
## Model: LM w/ ARIMA(0,1,1)(1,0,0)[7] errors
## Transformation: box_cox(num_casos.x, lambda_cad)
##
## Coefficients:
##       ma1     sar1
##       -0.5536  0.5053
##   s.e.  0.0738  0.0852
##       box_cox(residential_percent_change_from_baseline, lambda_cad_b)
##                                         -0.0761
##   s.e.                               0.0605
##       box_cox(Total, lambda_cad_h)
##                                         0.0874
##   s.e.                               0.0321
##
## sigma^2 estimated as 1.341: log likelihood=-194.89
## AIC=399.77  AICc=400.28  BIC=413.91
fit_model %>% select(arima_at2) %>% report()

## Series: num_casos.x
## Model: LM w/ ARIMA(0,1,1)(1,0,1)[7] errors
## Transformation: box_cox(num_casos.x, lambda_bar)
##
## Coefficients:
##       ma1     sar1     sma1
##       -0.5192  0.8550 -0.5220
##   s.e.  0.0952  0.1706  0.3379
##       box_cox(residential_percent_change_from_baseline, lambda_cad_b)
##                                         -0.0198
##   s.e.                               0.0210
##       box_cox(Total, lambda_cad_h)
##                                         0.0222
##   s.e.                               0.0145
##
## sigma^2 estimated as 0.1405: log likelihood=-53.89
## AIC=119.78  AICc=120.49  BIC=136.75

# Good model >> Less Sigma / AICc
fit_model %>% pivot_longer(!sub_region_2,
                           names_to = "Model name",
                           values_to = "Orders")

## # A mable: 3 x 3
## # Key:      sub_region_2, Model name [3]
##   sub_region_2 `Model name`           Orders
##   <chr>          <chr>                 <model>

```

```

## 1 Cádiz      SNaive          <SNAIVE>
## 2 Cádiz      arima_at1     <LM w/ ARIMA(0,1,1)(1,0,0)[7] errors>
## 3 Cádiz      arima_at2     <LM w/ ARIMA(0,1,1)(1,0,1)[7] errors>
glance(fit_model) %>% arrange(AICc) %>% select(.model:BIC)

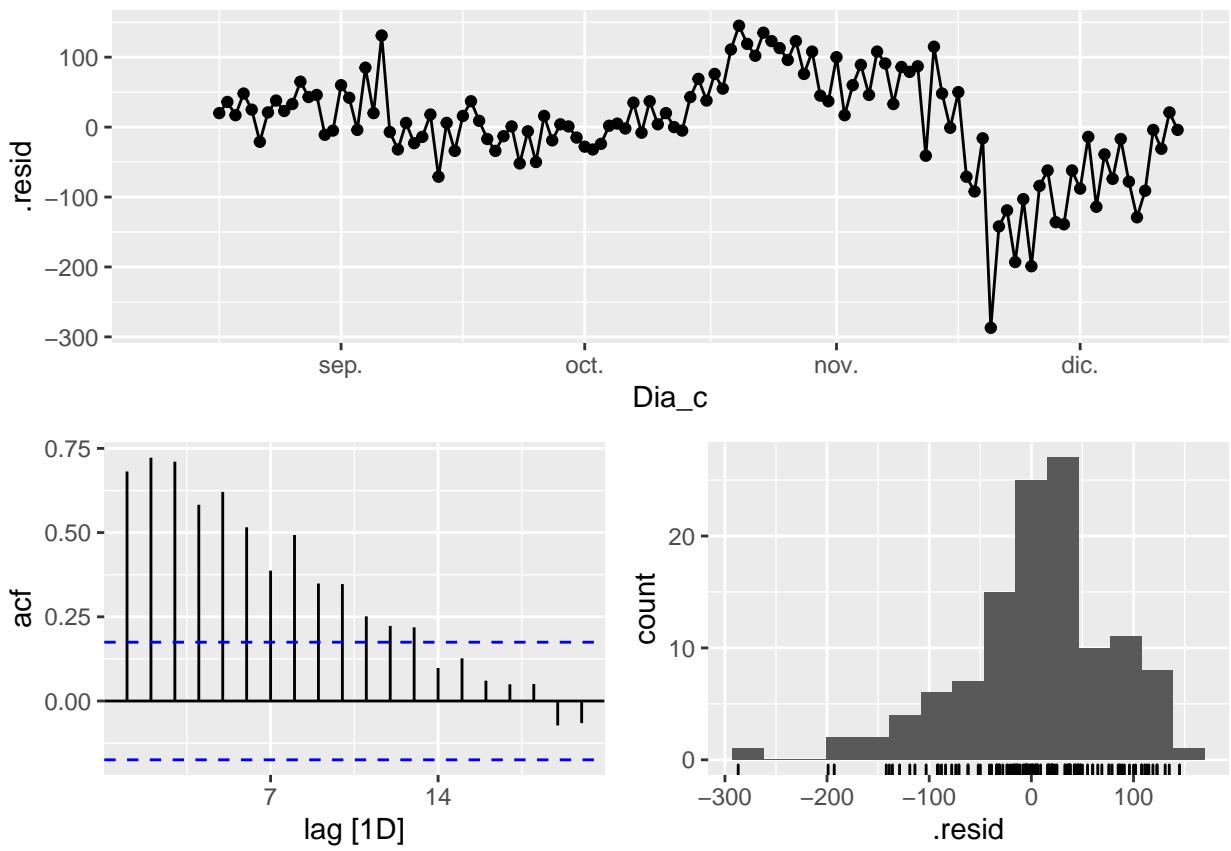
## # A tibble: 3 x 6
##   .model    sigma2 log_lik   AIC   AICc     BIC
##   <chr>     <dbl>   <dbl> <dbl> <dbl>   <dbl>
## 1 arima_at2  0.141   -53.9  120.  120.   137.
## 2 arima_at1  1.34    -195.   400.  400.   414.
## 3 SNaive    5510.     NA     NA     NA     NA
# We use a Ljung-Box test >> large p-value, confirms residuals are similar to white noise.
augment(fit_model) %>%
  features(.innov, lag=7)

## # A tibble: 3 x 4
##   sub_region_2 .model    lb_stat lb_pvalue
##   <chr>        <chr>     <dbl>     <dbl>
## 1 Cádiz       arima_at1  2.82     0.901
## 2 Cádiz       arima_at2  3.98     0.782
## 3 Cádiz       SNaive    328.      0
augment(fit_model) %>%
  features(.innov, lag=14)

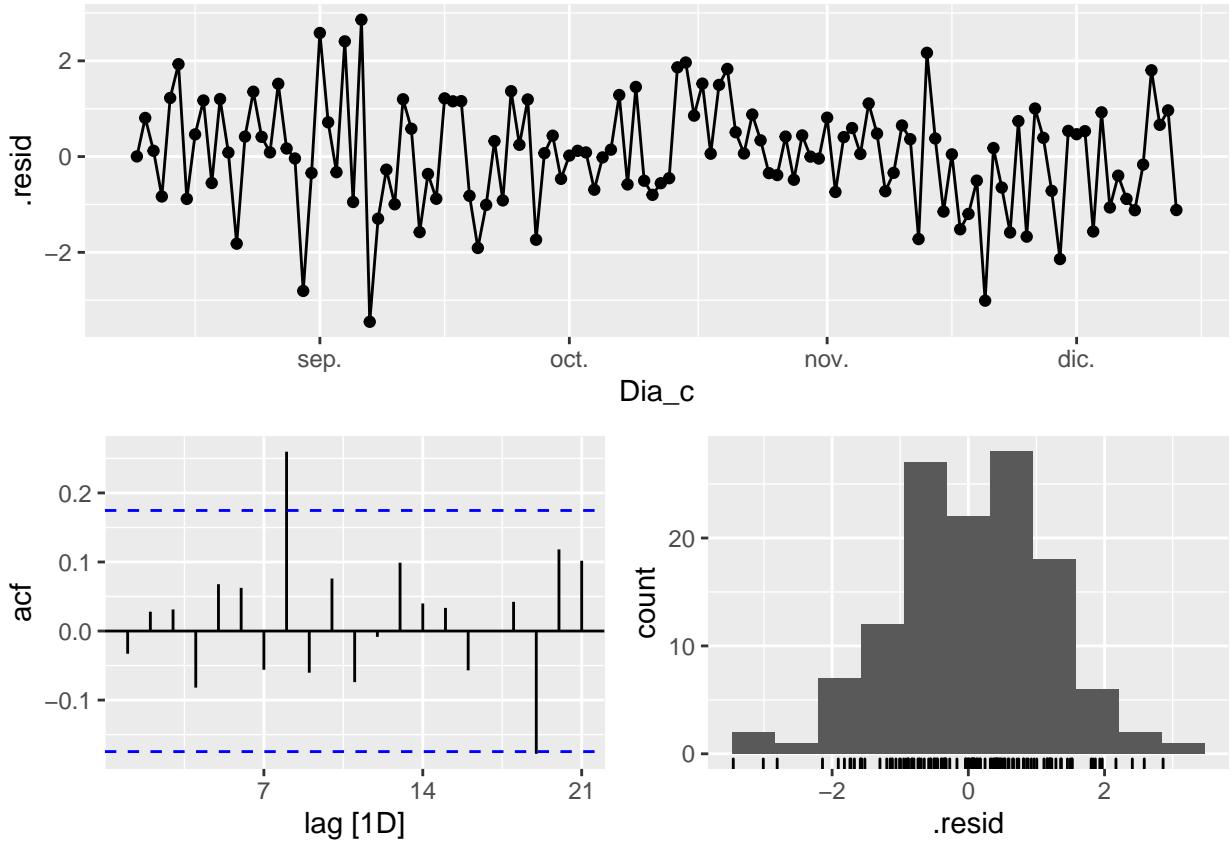
## # A tibble: 3 x 4
##   sub_region_2 .model    lb_stat lb_pvalue
##   <chr>        <chr>     <dbl>     <dbl>
## 1 Cádiz       arima_at1  15.8     0.329
## 2 Cádiz       arima_at2  17.1     0.251
## 3 Cádiz       SNaive    414.      0
augment(fit_model) %>%
  features(.innov, lag=21)

## # A tibble: 3 x 4
##   sub_region_2 .model    lb_stat lb_pvalue
##   <chr>        <chr>     <dbl>     <dbl>
## 1 Cádiz       arima_at1  25.1     0.241
## 2 Cádiz       arima_at2  24.5     0.270
## 3 Cádiz       SNaive    420.      0
fit_model %>% select(SNaive) %>% gg_tsresiduals()

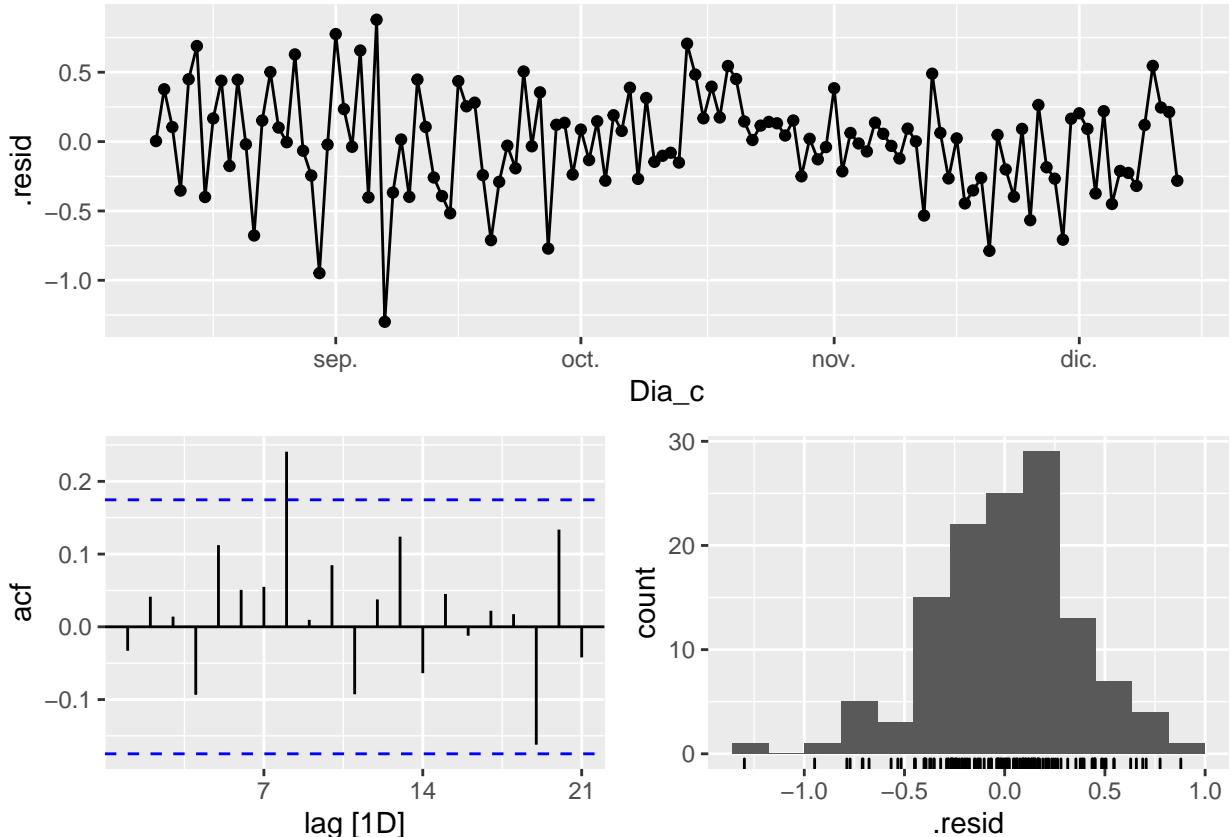
```



```
fit_model %>% select(arima_at1) %>% gg_tsresiduals()
```



```
fit_model %>% select(arima_at2) %>% gg_tsresiduals()
```



```

# Significant spikes out of 30 is still consistent with white noise.
# To be sure, use a Ljung-Box test, which has a large p-value, confirming that the
# residuals are similar to white noise.
# Note that the alternative models also pass this test.

# New data (dynamic regression)
# Here it is needed generate future values for the exogenous variables
# For simplicity we select a rand number included into the 2nd
# and 3rd quantile for the variable

# h7
Cad_N_cases_fr7 <- new_data(Cad_N_cases_tr, 7) %>%
  mutate(residential_percent_change_from_baseline =
    runif(7,quantile(Cad_N_cases_tt$residential_percent_change_from_baseline,
                      0.25),
          quantile(Cad_N_cases_tt$residential_percent_change_from_baseline,
                      0.75)),
    Total = runif(7,quantile(Cad_N_cases_tt$Total,0.25),
                  quantile(Cad_N_cases_tt$Total,0.75)))

# h14
Cad_N_cases_fr14 <- new_data(Cad_N_cases_tr, 14) %>%
  mutate(residential_percent_change_from_baseline =
    runif(14,quantile(Cad_N_cases_tt$residential_percent_change_from_baseline,
                      0.25),
          quantile(Cad_N_cases_tt$residential_percent_change_from_baseline,
                      0.75)))

```

```

          0.75)),
Total = runif(14,quantile(Cad_N_cases_tt$Total,0.25),
               quantile(Cad_N_cases_tt$Total,0.75))

# h17
Cad_N_cases_fr17 <- new_data(Cad_N_cases_tr, 17) %>%
  mutate(residential_percent_change_from_baseline =
    runif(17,quantile(Cad_N_cases_tt$residential_percent_change_from_baseline,
                      0.25),
           quantile(Cad_N_cases_tt$residential_percent_change_from_baseline,
                     0.75)),
  Total = runif(17,quantile(Cad_N_cases_tt$Total,0.25),
                quantile(Cad_N_cases_tt$Total,0.75)))

# Forecast
fc_fh7<-fabletools::forecast(fit_model, new_data = Cad_N_cases_fr7)
fc_fh14<-fabletools::forecast(fit_model, new_data = Cad_N_cases_fr14)
fc_fh17<-fabletools::forecast(fit_model, new_data = Cad_N_cases_fr17)

# Accuracy
fabletools::accuracy(fc_fh7, Cad_N_cases)

## # A tibble: 3 x 11
##   .model sub_region_2 .type     ME   RMSE   MAE     MPE   MAPE   MASE RMSSE      ACF1
##   <chr>   <chr>     <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 arima~ Cádiz     Test    -4.62  13.6  12.1   -3.51   7.96  0.431  0.270 -0.481
## 2 arima~ Cádiz     Test     2.95  10.2   8.56   1.59   5.47  0.304  0.203 -0.630
## 3 SNaive Cádiz    Test     0.143 25.1  22.4    0.430 15.0   0.797  0.499  0.00424
fabletools::accuracy(fc_fh14, Cad_N_cases)

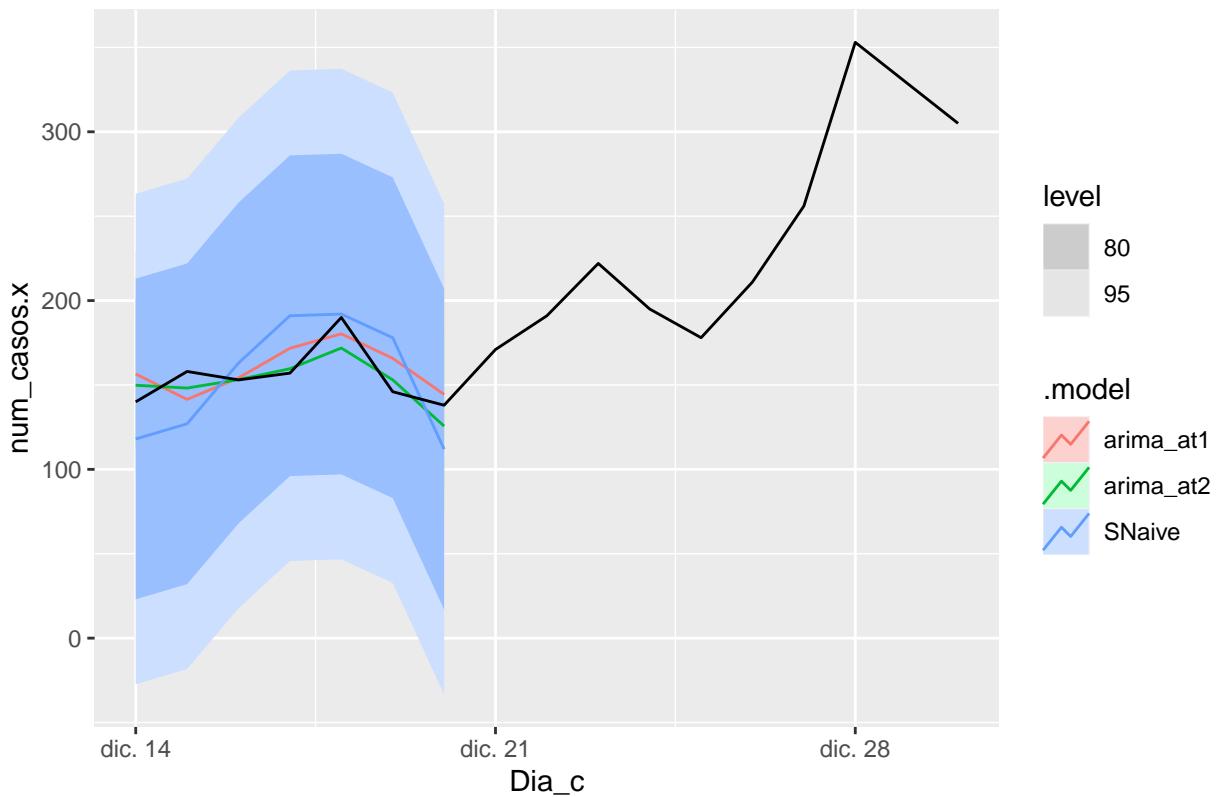
## # A tibble: 3 x 11
##   .model      sub_region_2 .type     ME   RMSE   MAE     MPE   MAPE   MASE RMSSE      ACF1
##   <chr>      <chr>     <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 arima_at1 Cádiz     Test    20.3   42.7  30.5   8.38  15.3   1.08  0.850  0.375
## 2 arima_at2 Cádiz     Test    33.7   54.5  38.1  15.7   18.7   1.36  1.08   0.435
## 3 SNaive Cádiz     Test    24.6   51.3  37.7  11.4   19.9   1.34  1.02   0.284
fabletools::accuracy(fc_fh17, Cad_N_cases)

## # A tibble: 3 x 11
##   .model      sub_region_2 .type     ME   RMSE   MAE     MPE   MAPE   MASE RMSSE      ACF1
##   <chr>      <chr>     <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 arima_at1 Cádiz     Test    46.8   79.9  53.8   16.3  21.2   1.91  1.59   0.747
## 2 arima_at2 Cádiz     Test    63.2   96.4  66.2   23.9  26.0   2.35  1.92   0.777
## 3 SNaive Cádiz     Test    54.3   94.9  65.1   19.7  26.6   2.32  1.89   0.766

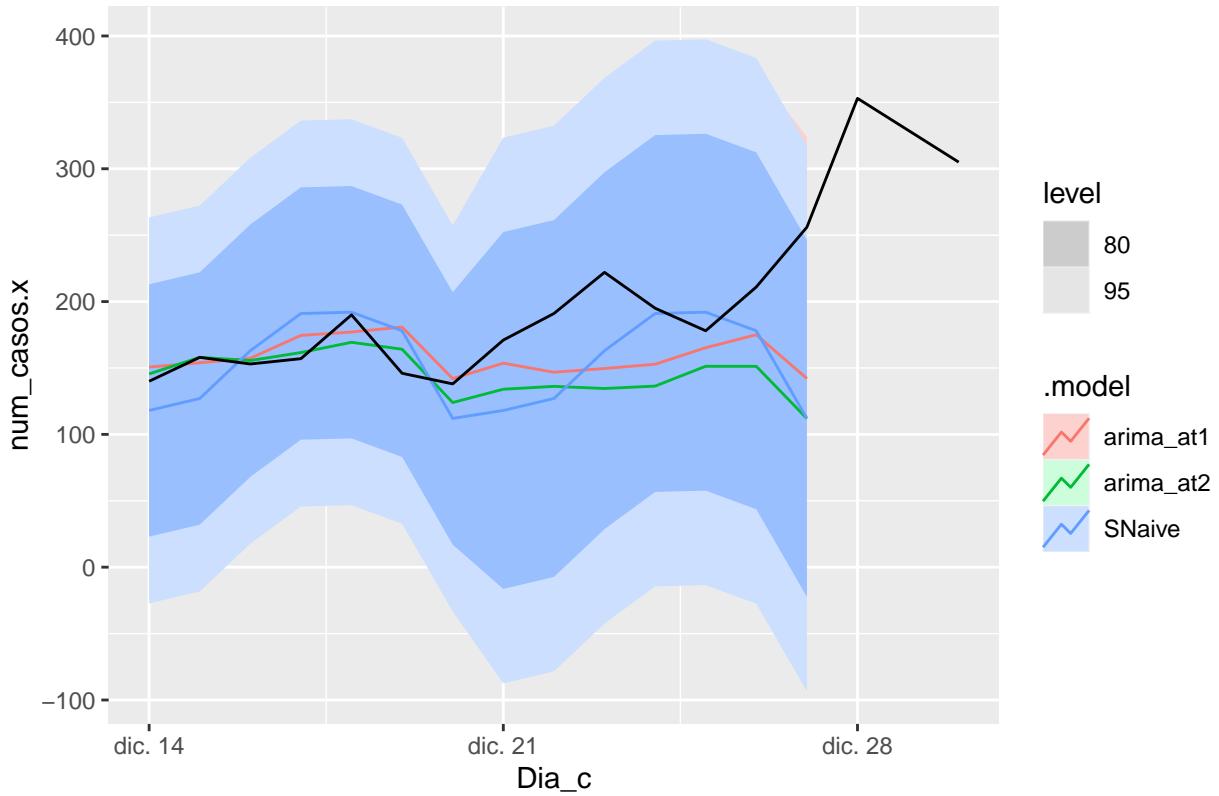
# Plots
fc_fh7 %>%
  autoplot(Cad_N_cases_tt) +
  labs(title="Málaga - forecast h7")

```

Málaga – forecast h7

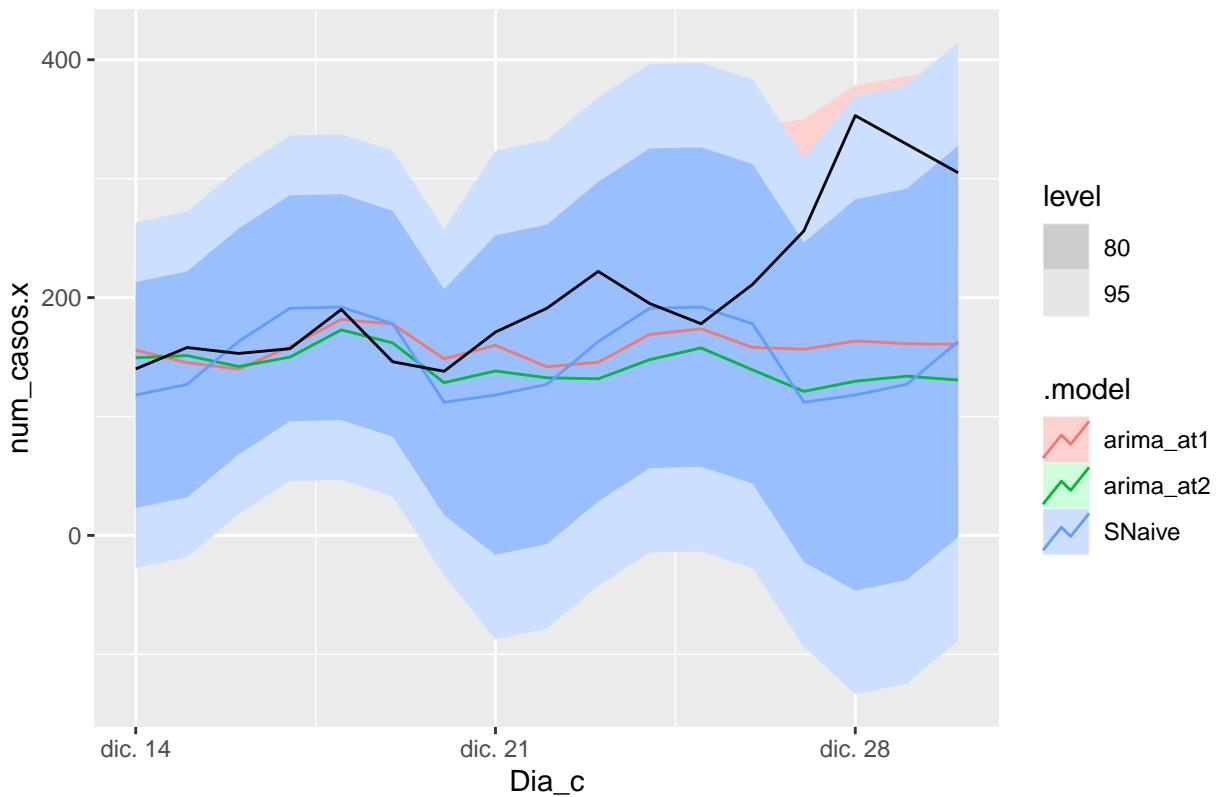


Málaga – forecast h14



```
fc_fh17 %>%
  autoplot(Cad_N_cases_tt) +
  labs(title="Málaga - forecast h17")
```

Málaga – forecast h17



```
# Opt A
# We have added "residential_percent_change" variable to models
lambda_sev_b <- Sev_N_cases %>%
  features(residential_percent_change_from_baseline, features = guerero) %>%
  pull(lambda_guerrero)
lambda_sev_h <- Sev_N_cases %>%
  features(Total, features = guerero) %>%
  pull(lambda_guerrero)

fit_model <- Sev_N_cases_tr %>%
  model(
    SNaive = SNAIVE(num_casos.x),
    arima_at1 = ARIMA(box_cox(num_casos.x,lambda_sev) ~
      box_cox(residential_percent_change_from_baseline,lambda_sev_b) +
      box_cox(Total,lambda_sev_h)),
    arima_at2 = ARIMA(box_cox(num_casos.x,lambda_bar) ~
      box_cox(residential_percent_change_from_baseline,lambda_sev_b) +
      box_cox(Total,lambda_sev_h) ,
      stepwise = FALSE, approx = FALSE))

# Show and report model
fit_model
```

3.3.2.5 Sevilla

```

## # A mable: 1 x 4
## # Key:      sub_region_2 [1]
##   sub_region_2    SNaive                               arima_at1
##   <chr>          <model>                               <model>
## 1 Sevilla        <SNAIVE> <LM w/ ARIMA(0,1,1)(2,0,0)[7] errors>
## # ... with 1 more variable: arima_at2 <model>
fit_model %>% select(arima_at1) %>% report()

## Series: num_casos.x
## Model: LM w/ ARIMA(0,1,1)(2,0,0)[7] errors
## Transformation: box_cox(num_casos.x, lambda_sev)
##
## Coefficients:
##       ma1     sar1     sar2
##       -0.4635  0.4187  0.3214
## s.e.  0.0772  0.0923  0.0912
##       box_cox(residential_percent_change_from_baseline, lambda_sev_b)
##                                         -0.1396
## s.e.                           0.0374
##       box_cox(Total, lambda_sev_h)
##                                         0.7397
## s.e.                           0.2833
##
## sigma^2 estimated as 0.7324: log likelihood=-157.91
## AIC=327.83  AICc=328.54  BIC=344.8
fit_model %>% select(arima_at2) %>% report()

## Series: num_casos.x
## Model: LM w/ ARIMA(2,1,2)(2,0,0)[7] errors
## Transformation: box_cox(num_casos.x, lambda_bar)
##
## Coefficients:
##       ar1     ar2     ma1     ma2     sar1     sar2
##       0.6604 -0.5969 -1.0937  0.7156  0.4488  0.2984
## s.e.  0.1468  0.1415  0.1348  0.1265  0.1028  0.0975
##       box_cox(residential_percent_change_from_baseline, lambda_sev_b)
##                                         -0.0663
## s.e.                           0.0156
##       box_cox(Total, lambda_sev_h)
##                                         0.4086
## s.e.                           0.0996
##
## sigma^2 estimated as 0.1453: log likelihood=-55.18
## AIC=128.36  AICc=129.92  BIC=153.81

# Good model >> Less Sigma / AICc
fit_model %>% pivot_longer(!sub_region_2,
                           names_to = "Model name",
                           values_to = "Orders")

## # A mable: 3 x 3
## # Key:      sub_region_2, Model name [3]
##   sub_region_2 `Model name`                               Orders
##   <chr>          <chr>                               <model>

```

```

## 1 Sevilla      SNaive                      <SNAIVE>
## 2 Sevilla      arima_at1      <LM w/ ARIMA(0,1,1)(2,0,0)[7] errors>
## 3 Sevilla      arima_at2      <LM w/ ARIMA(2,1,2)(2,0,0)[7] errors>
glance(fit_model) %>% arrange(AICc) %>% select(.model:BIC)

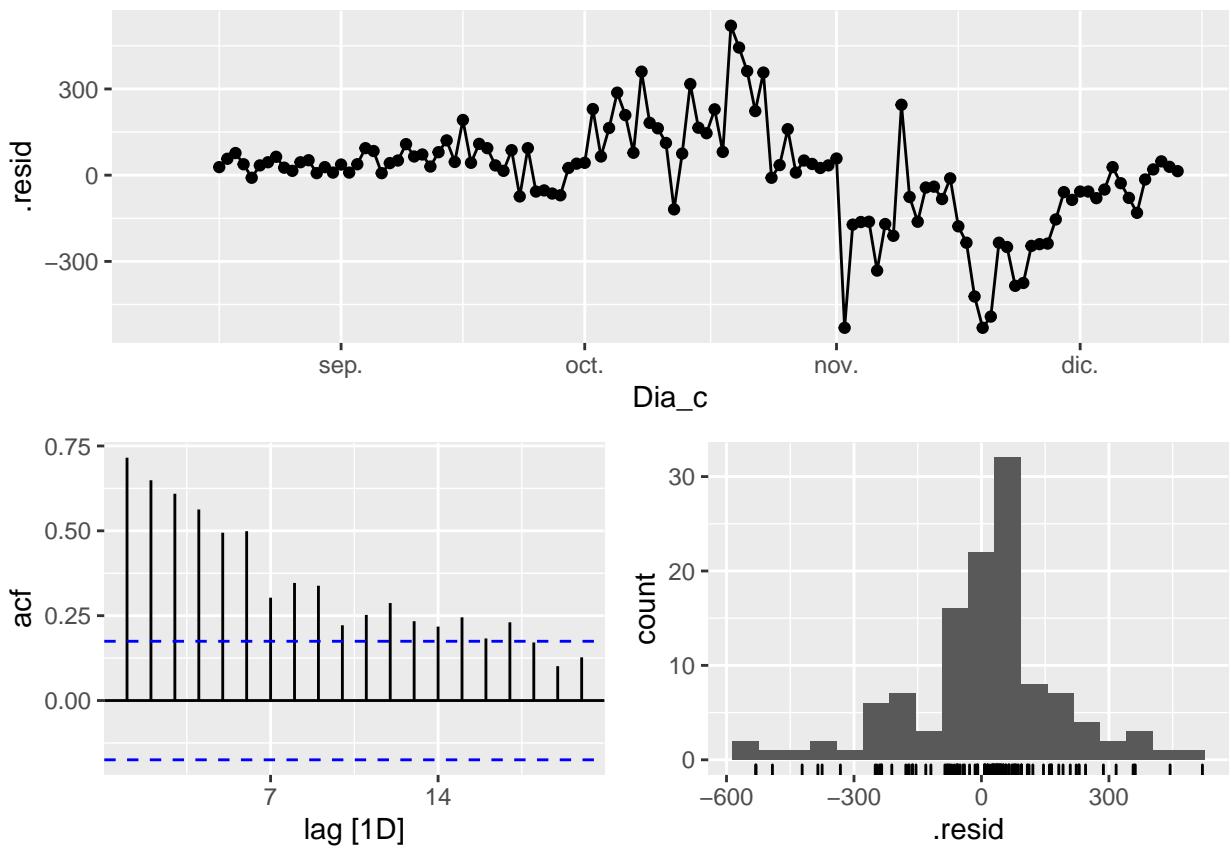
## # A tibble: 3 x 6
##   .model      sigma2 log_lik    AIC   AICc     BIC
##   <chr>        <dbl>    <dbl> <dbl> <dbl>    <dbl>
## 1 arima_at2    0.145   -55.2  128.  130.   154.
## 2 arima_at1    0.732   -158.   328.  329.   345.
## 3 SNaive       31787.      NA     NA     NA     NA
# We use a Ljung-Box test >> large p-value, confirms residuals are similar to white noise.
augment(fit_model) %>%
  features(.innov, lag=7)

## # A tibble: 3 x 4
##   sub_region_2 .model      lb_stat lb_pvalue
##   <chr>        <chr>      <dbl>     <dbl>
## 1 Sevilla      arima_at1   8.62     0.281
## 2 Sevilla      arima_at2   3.48     0.837
## 3 Sevilla      SNaive     274.      0
augment(fit_model) %>%
  features(.innov, lag=14)

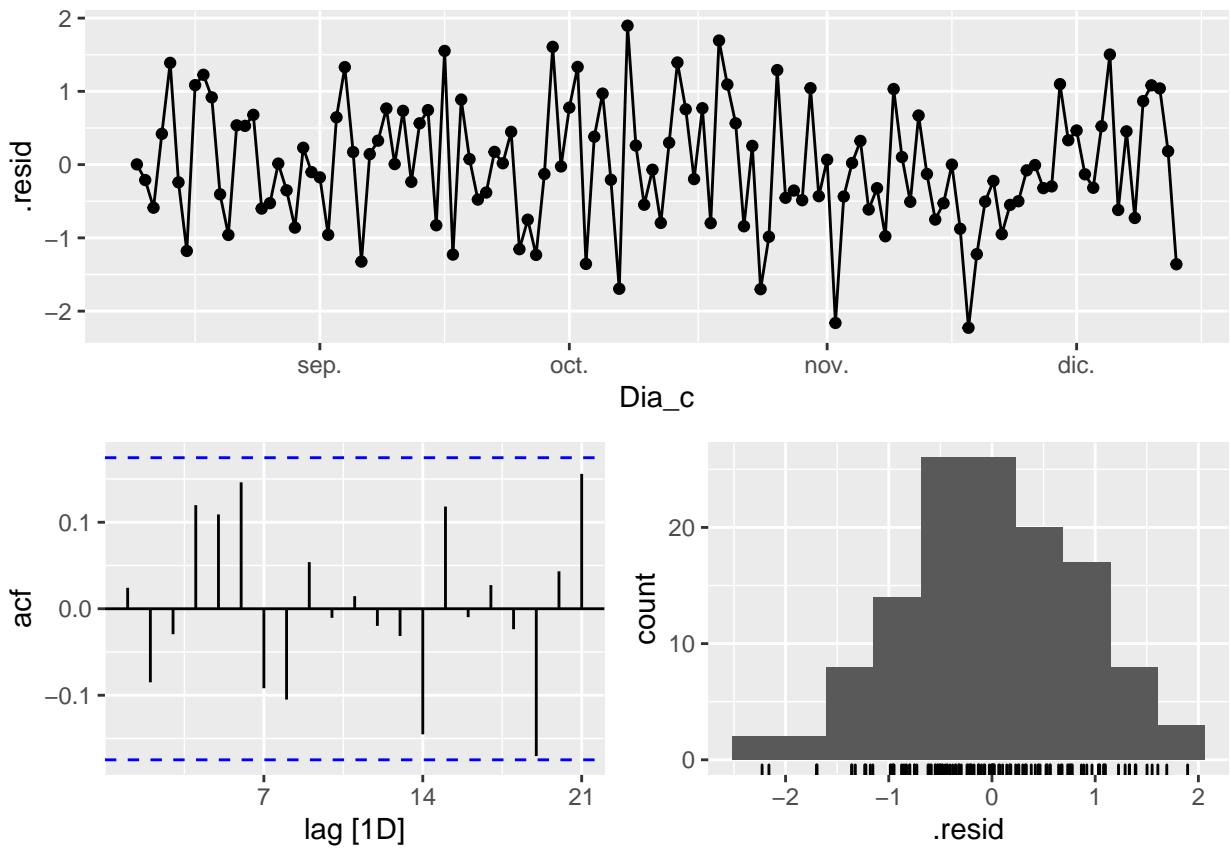
## # A tibble: 3 x 4
##   sub_region_2 .model      lb_stat lb_pvalue
##   <chr>        <chr>      <dbl>     <dbl>
## 1 Sevilla      arima_at1  13.8     0.464
## 2 Sevilla      arima_at2   7.50     0.914
## 3 Sevilla      SNaive     345.      0
augment(fit_model) %>%
  features(.innov, lag=21)

## # A tibble: 3 x 4
##   sub_region_2 .model      lb_stat lb_pvalue
##   <chr>        <chr>      <dbl>     <dbl>
## 1 Sevilla      arima_at1  24.4     0.273
## 2 Sevilla      arima_at2  11.9     0.943
## 3 Sevilla      SNaive     377.      0
fit_model %>% select(SNaive) %>% gg_tsresiduals()

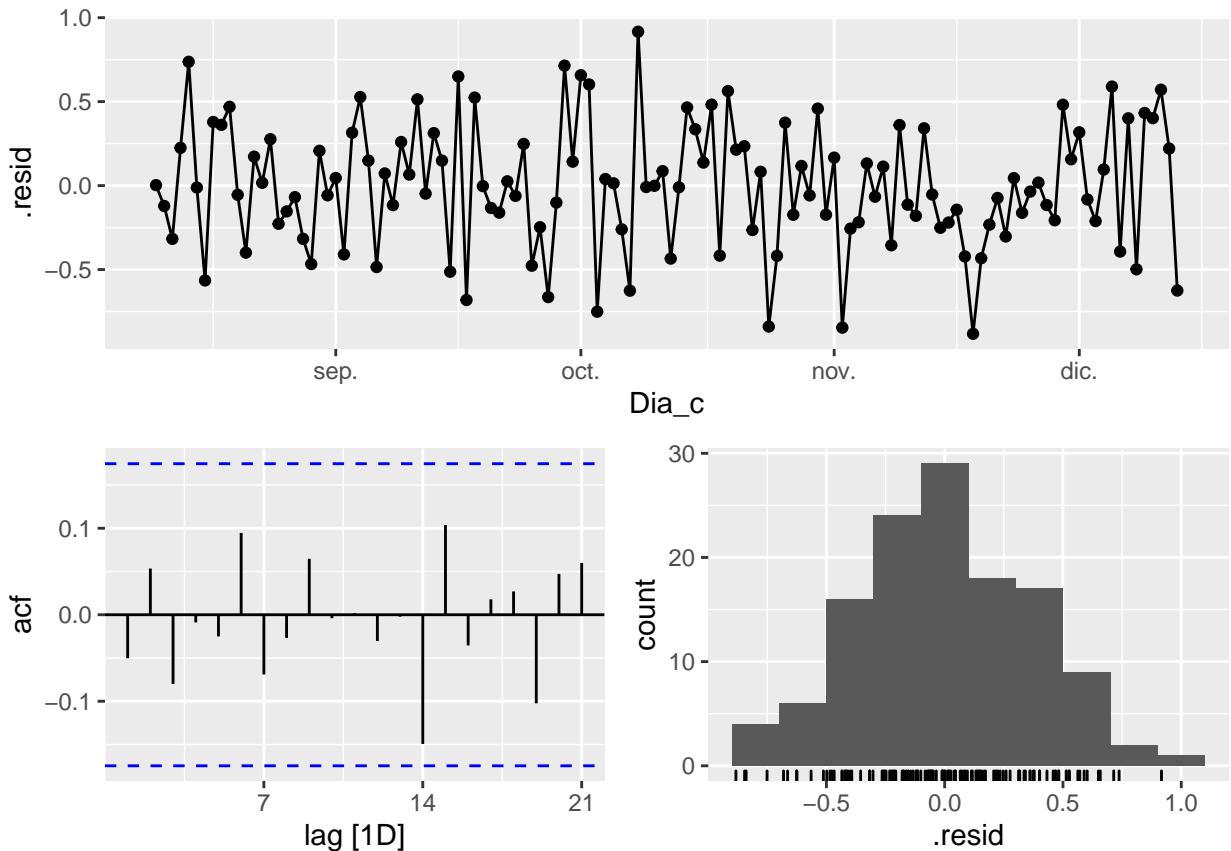
```



```
fit_model %>% select(arima_at1) %>% gg_tsresiduals()
```



```
fit_model %>% select(arima_at2) %>% gg_tsresiduals()
```



```

# Significant spikes out of 30 is still consistent with white noise.
# To be sure, use a Ljung-Box test, which has a large p-value, confirming that the
# residuals are similar to white noise.
# Note that the alternative models also pass this test.

# New data (dynamic regression)
# Here it is needed generate future values for the exogenous variables
# For simplicity we select a rand number included into the 2nd
# and 3rd quantile for the variable

# h7
Sev_N_cases_fr7 <- new_data(Sev_N_cases_tr, 7) %>%
  mutate(residential_percent_change_from_baseline =
    runif(7,quantile(Sev_N_cases_tt$residential_percent_change_from_baseline,
      0.25),
    quantile(Sev_N_cases_tt$residential_percent_change_from_baseline,
      0.75)),
  Total = runif(7,quantile(Sev_N_cases_tt$Total,0.25),
    quantile(Sev_N_cases_tt$Total,0.75)))

# h14
Sev_N_cases_fr14 <- new_data(Sev_N_cases_tr, 14) %>%
  mutate(residential_percent_change_from_baseline =
    runif(14,quantile(Sev_N_cases_tt$residential_percent_change_from_baseline,
      0.25),
    quantile(Sev_N_cases_tt$residential_percent_change_from_baseline,
      0.75)))
  
```

```

        0.75)),
Total = runif(14,quantile(Sev_N_cases_tt$Total,0.25),
               quantile(Sev_N_cases_tt$Total,0.75))

# h17
Sev_N_cases_fr17 <- new_data(Sev_N_cases_tr, 17) %>%
  mutate(residential_percent_change_from_baseline =
    runif(17,quantile(Sev_N_cases_tt$residential_percent_change_from_baseline,
                      0.25),
          quantile(Sev_N_cases_tt$residential_percent_change_from_baseline,
                      0.75)),
  Total = runif(17,quantile(Sev_N_cases_tt$Total,0.25),
                quantile(Sev_N_cases_tt$Total,0.75)))

# Forecast
fc_fh7<-fabletools::forecast(fit_model, new_data = Sev_N_cases_fr7)
fc_fh14<-fabletools::forecast(fit_model, new_data = Sev_N_cases_fr14)
fc_fh17<-fabletools::forecast(fit_model, new_data = Sev_N_cases_fr17)

# Accuracy
fabletools::accuracy(fc_fh7, Sev_N_cases)

## # A tibble: 3 x 11
##   .model   sub_region_2 .type     ME   RMSE   MAE    MPE   MAPE   MASE RMSSE   ACF1
##   <chr>    <chr>      <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 arima_at1 Sevilla    Test   -6.99  30.8  26.1 -11.4  22.7  0.420  0.258  0.548
## 2 arima_at2 Sevilla    Test  -21.9   32.4  26.6 -23.0  25.7  0.428  0.271  0.516
## 3 SNaive    Sevilla    Test    -7    57.4  49.6 -11.2  36.4  0.797  0.479  0.482
fabletools::accuracy(fc_fh14, Sev_N_cases)

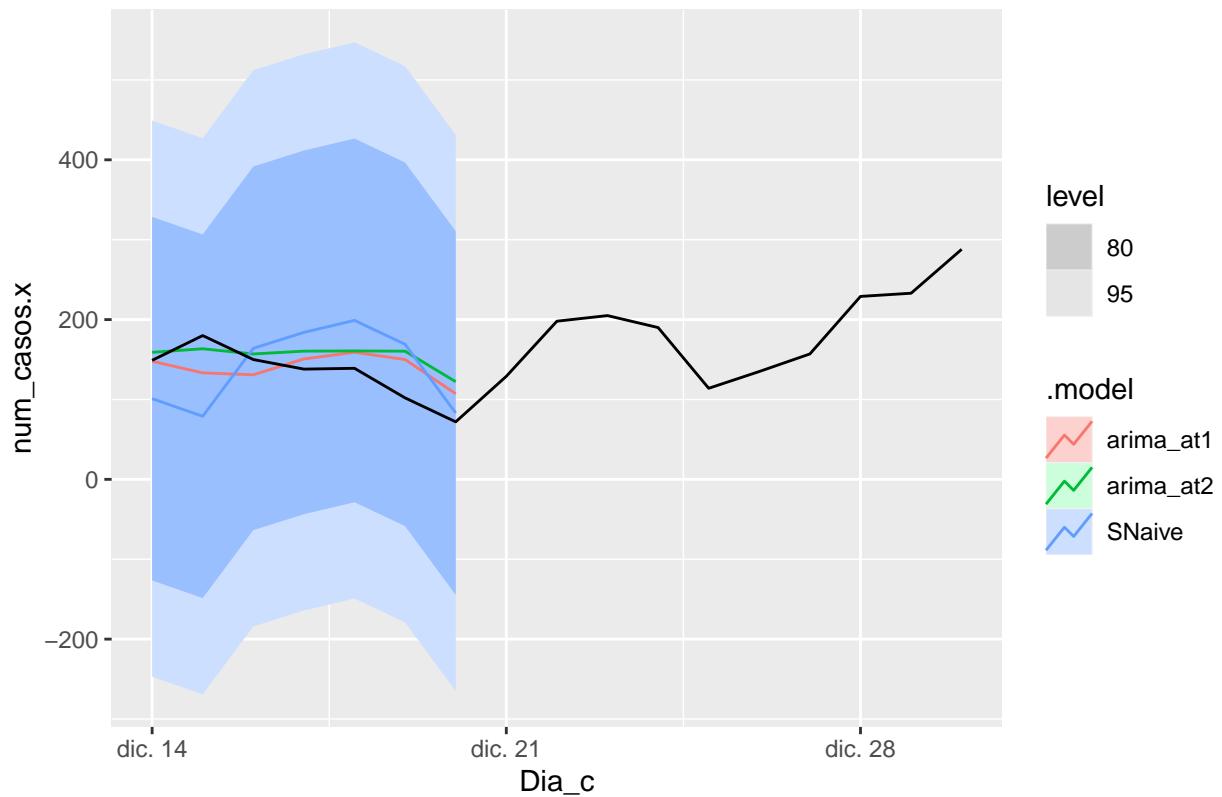
## # A tibble: 3 x 11
##   .model   sub_region_2 .type     ME   RMSE   MAE    MPE   MAPE   MASE RMSSE   ACF1
##   <chr>    <chr>      <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 arima_at1 Sevilla    Test    8.46  47.5  39.5 -1.29  27.3  0.636  0.397  0.418
## 2 arima_at2 Sevilla    Test  -10.3   45.2  39.2 -14.4  29.7  0.630  0.378  0.364
## 3 SNaive    Sevilla    Test   7.14  61.8  52.4 -1.87  36.2  0.843  0.517  0.373
fabletools::accuracy(fc_fh17, Sev_N_cases)

## # A tibble: 3 x 11
##   .model   sub_region_2 .type     ME   RMSE   MAE    MPE   MAPE   MASE RMSSE   ACF1
##   <chr>    <chr>      <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 arima_at1 Sevilla    Test   27.1   63.3  51.5  7.45  29.7  0.829  0.529  0.493
## 2 arima_at2 Sevilla    Test   5.06   50.6  43.6 -5.98  28.4  0.701  0.422  0.421
## 3 SNaive    Sevilla    Test   29.8   80.1  67.1  8.17  39.5  1.08   0.669  0.607

# Plots
fc_fh7 %>%
  autoplot(Sev_N_cases_tt) +
  labs(title="Sevilla - forecast h7")

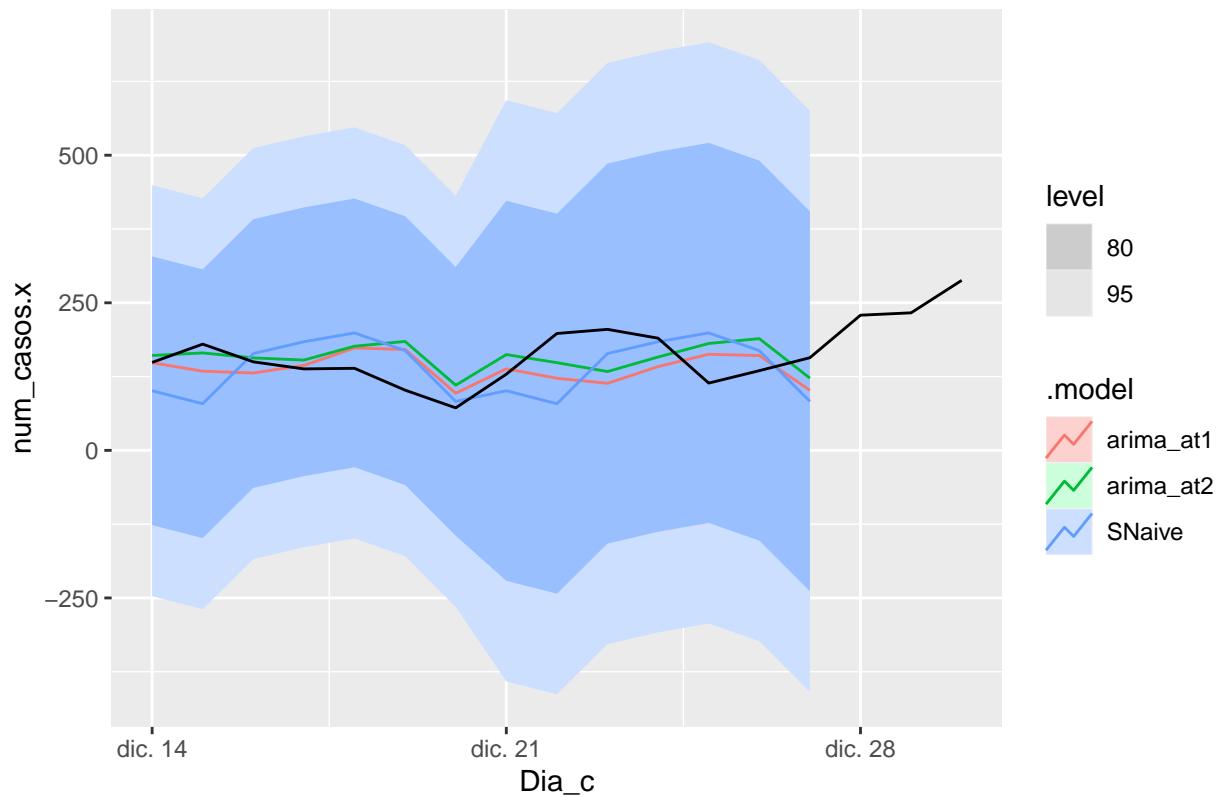
```

Sevilla – forecast h7



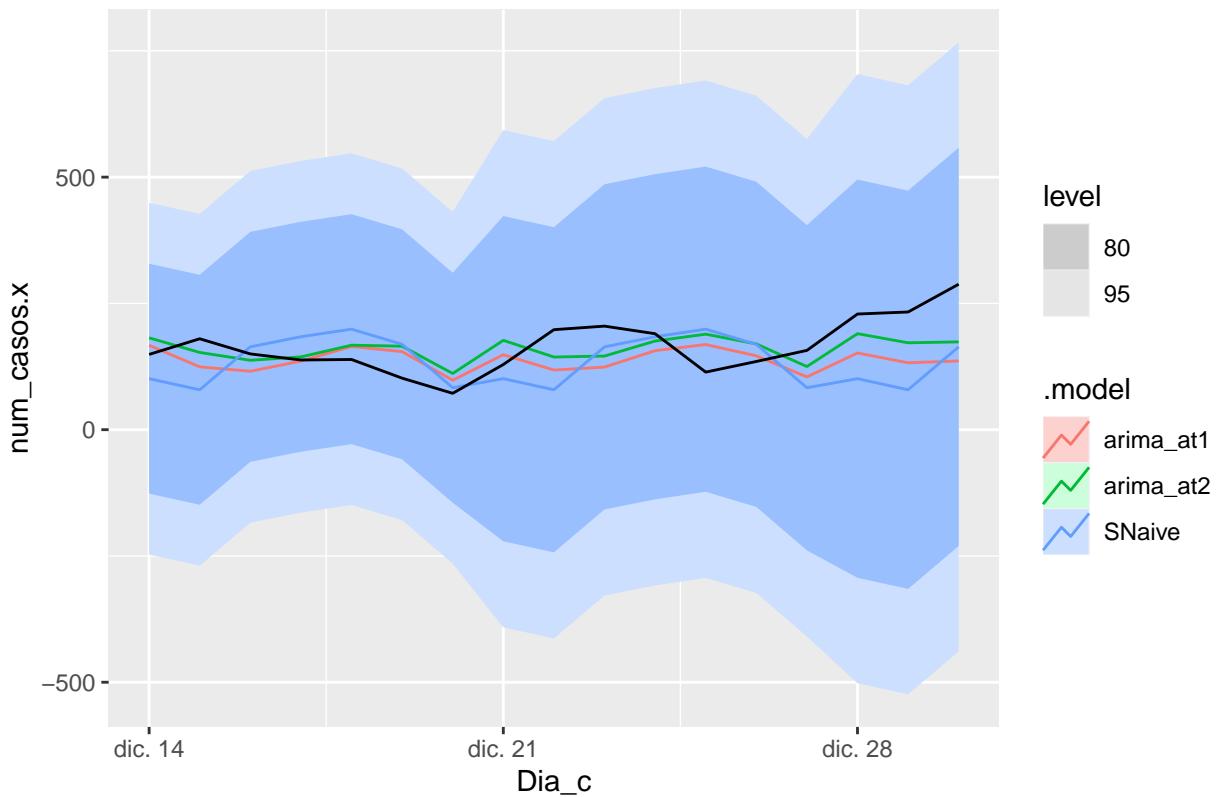
```
fc_fh14 %>%
  autoplot(Sev_N_cases_tt) +
  labs(title="Sevilla - forecast h14")
```

Sevilla – forecast h14



```
fc_fh17 %>%
  autoplot(Sev_N_cases_tt) +
  labs(title="Sevilla - forecast h17")
```

Sevilla – forecast h17



3.3.3 Multivariate (7, 14, 17 days) + All mobility

```
# Model train
# Opt B
# We have added all mobility variables to models
lambda_bar_b <- Bar_N_cases %>%
  features(retail_and_recreation_percent_change_from_baseline, features = guerrero) %>%
  pull(lambda_guerrero)
lambda_bar_c <- Bar_N_cases %>%
  features(grocery_and_pharmacy_percent_change_from_baseline, features = guerrero) %>%
  pull(lambda_guerrero)
lambda_bar_d <- Bar_N_cases %>%
  features(parks_percent_change_from_baseline, features = guerrero) %>%
  pull(lambda_guerrero)
lambda_bar_e <- Bar_N_cases %>%
  features(transit_stations_percent_change_from_baseline, features = guerrero) %>%
  pull(lambda_guerrero)
lambda_bar_f <- Bar_N_cases %>%
  features(workplaces_percent_change_from_baseline, features = guerrero) %>%
  pull(lambda_guerrero)
lambda_bar_g <- Bar_N_cases %>%
  features(residential_percent_change_from_baseline, features = guerrero) %>%
  pull(lambda_guerrero)
lambda_bar_h <- Bar_N_cases %>%
  features>Total, features = guerrero) %>%
```

```

pull(lambda_guerrero)

fit_model <- Bar_N_cases_tr %>%
  model(
    SNaive = SNAIVE(num_casos.x),
    arima_at1 = ARIMA(box_cox(num_casos.x,lambda_bar) ~
      box_cox(retail_and_recreation_percent_change_from_baseline,lambda_bar_b) +
      box_cox(grocery_and_pharmacy_percent_change_from_baseline,lambda_bar_c) +
      box_cox(parks_percent_change_from_baseline,lambda_bar_d) +
      box_cox(transit_stations_percent_change_from_baseline,lambda_bar_e) +
      box_cox(workplaces_percent_change_from_baseline,lambda_bar_f) +
      box_cox(residential_percent_change_from_baseline,lambda_bar_g) +
      box_cox(Total,lambda_bar_h)),
    arima_at2 = ARIMA(box_cox(num_casos.x,lambda_bar) ~
      box_cox(retail_and_recreation_percent_change_from_baseline,lambda_bar_b) +
      box_cox(grocery_and_pharmacy_percent_change_from_baseline,lambda_bar_c) +
      box_cox(parks_percent_change_from_baseline,lambda_bar_d) +
      box_cox(transit_stations_percent_change_from_baseline,lambda_bar_e) +
      box_cox(workplaces_percent_change_from_baseline,lambda_bar_f) +
      box_cox(residential_percent_change_from_baseline,lambda_bar_g) +
      box_cox(Total,lambda_bar_h),
      stepwise = FALSE,
      approx = FALSE))

# Show and report model
fit_model

```

3.3.3.1 Barcelona

```

## # A mable: 1 x 4
## # Key:     sub_region_2 [1]
##   sub_region_2   SNaive                      arima_at1
##   <chr>          <model>                     <model>
## 1 Barcelona    <SNAIVE> <LM w/ ARIMA(1,0,0)(0,0,1)[7] errors>
## # ... with 1 more variable: arima_at2 <model>

fit_model %>% select(arima_at1) %>% report()

## Series: num_casos.x
## Model: LM w/ ARIMA(1,0,0)(0,0,1)[7] errors
## Transformation: box_cox(num_casos.x, lambda_bar)
##
## Coefficients:
##       ar1     sma1
##       0.9128  0.5354
## s.e.  0.0362  0.1173
##       box_cox(retail_and_recreation_percent_change_from_baseline, lambda_bar_b)
##                                         0e+00
##                                         3e-04
## s.e.
##       box_cox(grocery_and_pharmacy_percent_change_from_baseline, lambda_bar_c)
##                                         0.0130
##                                         0.0058
## s.e.
##       box_cox(parks_percent_change_from_baseline, lambda_bar_d)
##                                         -0.0148
##                                         0.0026
## s.e.

```

```

##      box_cox(transit_stations_percent_change_from_baseline, lambda_bar_e)
##                                         0.0024
## s.e.                                         0.0005
##      box_cox(workplaces_percent_change_from_baseline, lambda_bar_f)
##                                         -4e-04
## s.e.                                         2e-04
##      box_cox(residential_percent_change_from_baseline, lambda_bar_g)
##                                         0.1046
## s.e.                                         0.0310
##      box_cox(Total, lambda_bar_h)  intercept
##                               0.0980    10.3128
## s.e.                               0.0704    0.5106
##
## sigma^2 estimated as 0.1027:  log likelihood=-32.5
## AIC=86.99  AICc=89.31  BIC=118.19
fit_model %>% select(arima_at2) %>% report()

## Series: num_casos.x
## Model: LM w/ ARIMA(0,0,5)(1,0,0)[7] errors
## Transformation: box_cox(num_casos.x, lambda_bar)
##
## Coefficients:
##      ma1     ma2     ma3     ma4     ma5     sar1
##      0.8360  0.7053  0.5056  0.4640  0.3485  0.8671
## s.e.  0.1111  0.1333  0.0929  0.0816  0.1028    NaN
##      box_cox(retail_and_recreation_percent_change_from_baseline, lambda_bar_b)
##                                         -3e-04
## s.e.                                         NaN
##      box_cox(grocery_and_pharmacy_percent_change_from_baseline, lambda_bar_c)
##                                         0.0190
## s.e.                                         0.0053
##      box_cox(parks_percent_change_from_baseline, lambda_bar_d)
##                                         -0.0048
## s.e.                                         0.0021
##      box_cox(transit_stations_percent_change_from_baseline, lambda_bar_e)
##                                         6e-04
## s.e.                                         4e-04
##      box_cox(workplaces_percent_change_from_baseline, lambda_bar_f)
##                                         1e-04
## s.e.                                         0e+00
##      box_cox(residential_percent_change_from_baseline, lambda_bar_g)
##                                         0.0202
## s.e.                                         0.0318
##      box_cox(Total, lambda_bar_h)  intercept
##                               0.1926    9.3503
## s.e.                               NaN      NaN
##
## sigma^2 estimated as 0.08968:  log likelihood=-25.35
## AIC=80.71  AICc=85.07  BIC=123.25

# Good model >> Less Sigma / AICc
fit_model %>% pivot_longer(!sub_region_2,
                           names_to = "Model name",
                           values_to = "Orders")

```

```

## # A mable: 3 x 3
## # Key:      sub_region_2, Model name [3]
##   sub_region_2 `Model name`          Orders
##   <chr>           <chr>            <model>
## 1 Barcelona     SNaive            <SNAIVE>
## 2 Barcelona     arima_at1       <LM w/ ARIMA(1,0,0)(0,0,1)[7] errors>
## 3 Barcelona     arima_at2       <LM w/ ARIMA(0,0,5)(1,0,0)[7] errors>
glance(fit_model) %>% arrange(AICc) %>% select(.model:BIC)

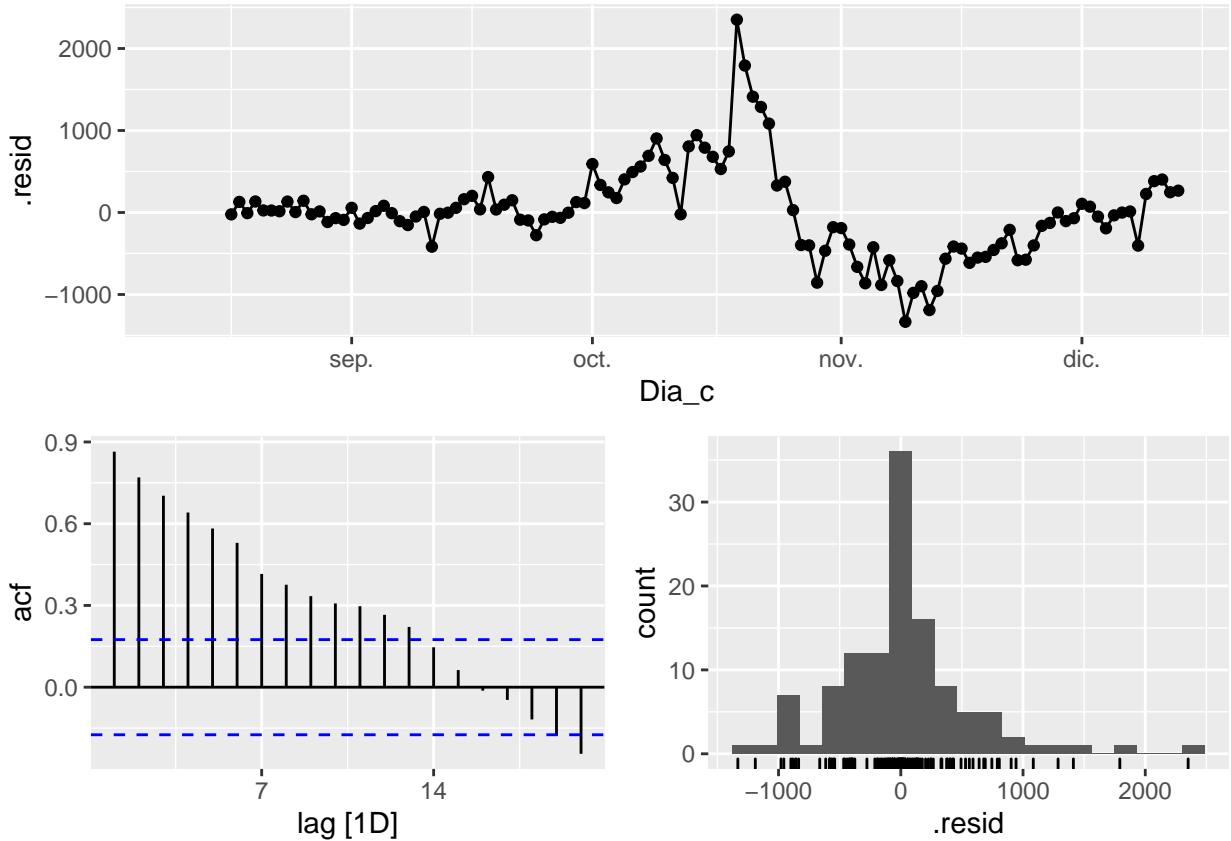
## # A tibble: 3 x 6
##   .model      sigma2 log_lik    AIC   AICc     BIC
##   <chr>      <dbl>   <dbl> <dbl> <dbl> <dbl>
## 1 arima_at2  0.0897 -25.4  80.7  85.1  123.
## 2 arima_at1  0.103   -32.5  87.0  89.3  118.
## 3 SNaive     299977.     NA     NA     NA     NA
# We use a Ljung-Box test >> large p-value, confirms residuals are similar to white noise.
augment(fit_model) %>%
  features(.innov, lag=7)

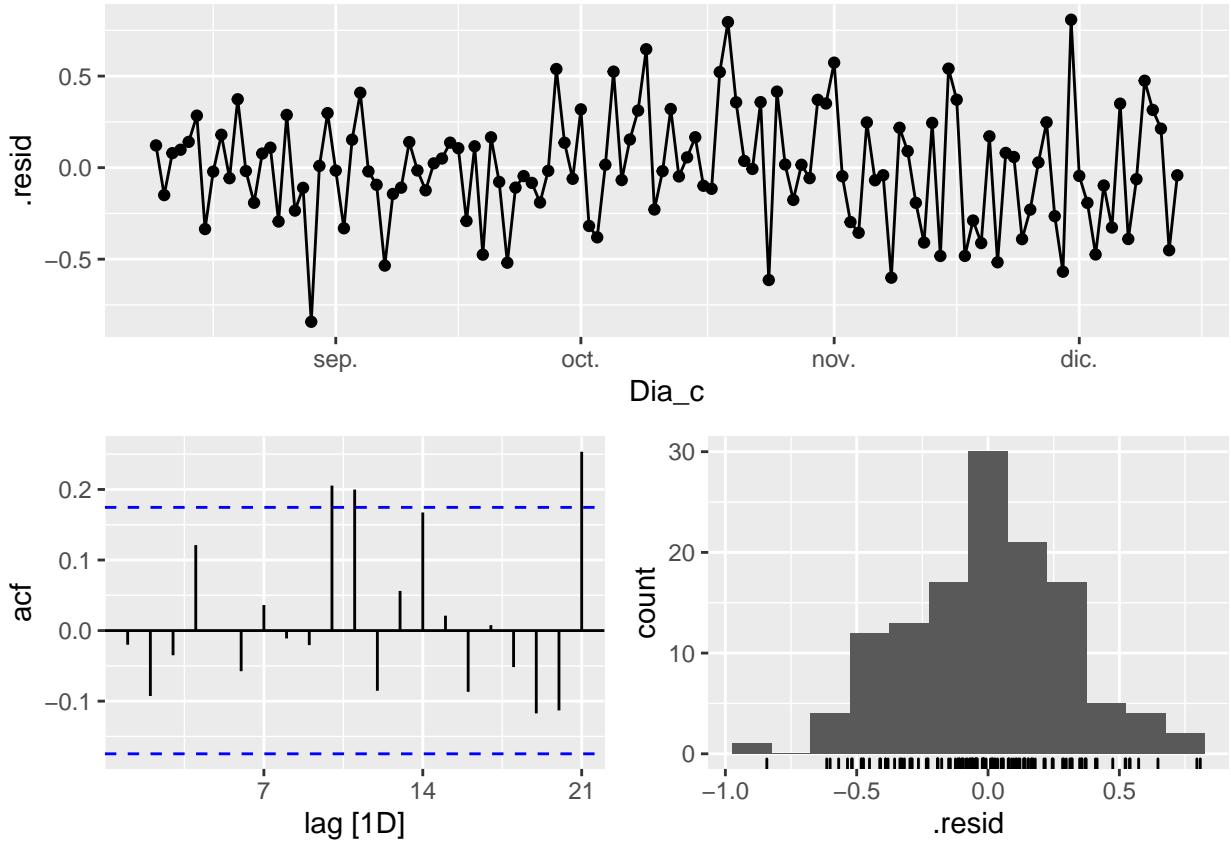
## # A tibble: 3 x 4
##   sub_region_2 .model    lb_stat lb_pvalue
##   <chr>        <chr>    <dbl>     <dbl>
## 1 Barcelona    arima_at1  3.89  0.792
## 2 Barcelona    arima_at2  30.4   0.0000814
## 3 Barcelona    SNaive    378.   0
augment(fit_model) %>%
  features(.innov, lag=14)

## # A tibble: 3 x 4
##   sub_region_2 .model    lb_stat lb_pvalue
##   <chr>        <chr>    <dbl>     <dbl>
## 1 Barcelona    arima_at1  20.9  0.103
## 2 Barcelona    arima_at2  52.6  0.00000227
## 3 Barcelona    SNaive    454.   0
augment(fit_model) %>%
  features(.innov, lag=21)

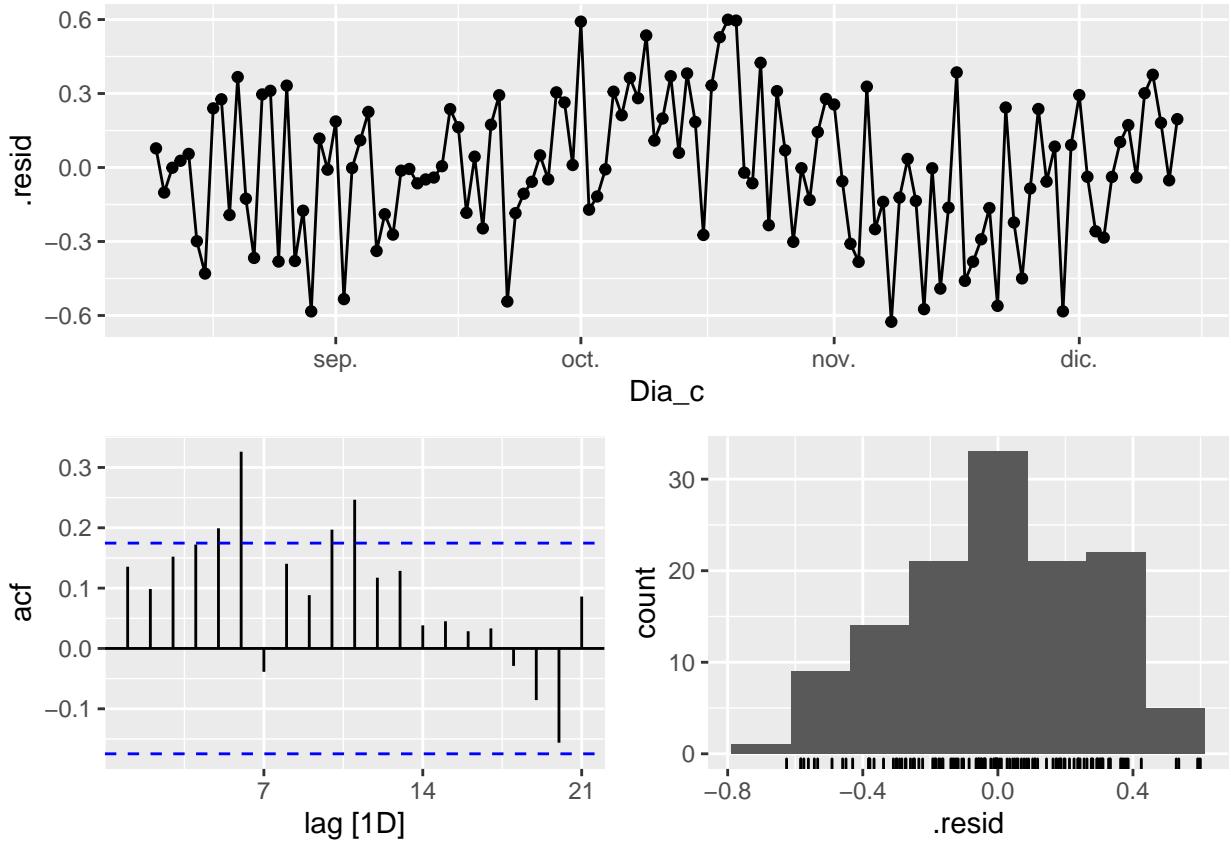
## # A tibble: 3 x 4
##   sub_region_2 .model    lb_stat lb_pvalue
##   <chr>        <chr>    <dbl>     <dbl>
## 1 Barcelona    arima_at1  36.4  0.0196
## 2 Barcelona    arima_at2  59.2  0.0000168
## 3 Barcelona    SNaive    482.   0
fit_model %>% select(SNaive) %>% gg_tsresiduals()

```





```
fit_model %>% select(arima_at2) %>% gg_tsresiduals()
```



```

# Significant spikes out of 30 is still consistent with white noise.
# To be sure, use a Ljung-Box test, which has a large p-value, confirming that the
# residuals are similar to white noise.
# Note that the alternative models also pass this test.

# New data (dynamic regression)
# Here it is needed generate future values for the exogenous variables
# For simplicity we select a rand number included into the 2nd and
# 3rd quantile for the variable

# h7
Bar_N_cases_fr7 <- new_data(Bar_N_cases_tr, 7) %>%
  mutate(retail_and_recreation_percent_change_from_baseline =
    runif(7,quantile(Bar_N_cases_tt$retail_and_recreation_percent_change_from_baseline,
      0.25),
    quantile(Bar_N_cases_tt$retail_and_recreation_percent_change_from_baseline,
      0.75)),
  grocery_and_pharmacy_percent_change_from_baseline =
    runif(7,quantile(Bar_N_cases_tt$grocery_and_pharmacy_percent_change_from_baseline,
      0.25),
    quantile(Bar_N_cases_tt$grocery_and_pharmacy_percent_change_from_baseline,
      0.75)),
  parks_percent_change_from_baseline =
    runif(7,quantile(Bar_N_cases_tr$parks_percent_change_from_baseline,
      0.25),
    quantile(Bar_N_cases_tt$parks_percent_change_from_baseline,
      0.75)))
  
```

```

        0.75)),
transit_stations_percent_change_from_baseline =
  runif(7,quantile(Bar_N_cases_tt$transit_stations_percent_change_from_baseline,
  0.25),
  quantile(Bar_N_cases_tt$transit_stations_percent_change_from_baseline,
  0.75)),
workplaces_percent_change_from_baseline =
  runif(7,quantile(Bar_N_cases_tt$workplaces_percent_change_from_baseline,
  0.25),
  quantile(Bar_N_cases_tt$workplaces_percent_change_from_baseline,
  0.75)),
residential_percent_change_from_baseline =
  runif(7,quantile(Bar_N_cases_tt$residential_percent_change_from_baseline,
  0.25),
  quantile(Bar_N_cases_tt$residential_percent_change_from_baseline,
  0.75)),
Total = runif(7,quantile(Bar_N_cases_tt$Total,0.25),
  quantile(Bar_N_cases_tt$Total,0.75)))

# h14
Bar_N_cases_fr14 <- new_data(Bar_N_cases_tr, 14) %>%
  mutate(retail_and_recreation_percent_change_from_baseline =
    runif(14,quantile(Bar_N_cases_tt$retail_and_recreation_percent_change_from_baseline,
    0.25),
    quantile(Bar_N_cases_tt$retail_and_recreation_percent_change_from_baseline,
    0.75)),
  grocery_and_pharmacy_percent_change_from_baseline =
    runif(14,quantile(Bar_N_cases_tt$grocery_and_pharmacy_percent_change_from_baseline,
    0.25),
    quantile(Bar_N_cases_tt$grocery_and_pharmacy_percent_change_from_baseline,
    0.75)),
  parks_percent_change_from_baseline =
    runif(14,quantile(Bar_N_cases_tt$parks_percent_change_from_baseline,
    0.25),
    quantile(Bar_N_cases_tt$parks_percent_change_from_baseline,
    0.75)),
  transit_stations_percent_change_from_baseline =
    runif(14,quantile(Bar_N_cases_tt$transit_stations_percent_change_from_baseline,
    0.25),
    quantile(Bar_N_cases_tt$transit_stations_percent_change_from_baseline,
    0.75)),
  workplaces_percent_change_from_baseline =
    runif(14,quantile(Bar_N_cases_tt$workplaces_percent_change_from_baseline,
    0.25),
    quantile(Bar_N_cases_tt$workplaces_percent_change_from_baseline,
    0.75)),
  residential_percent_change_from_baseline =
    runif(14,quantile(Bar_N_cases_tt$residential_percent_change_from_baseline,
    0.25),
    quantile(Bar_N_cases_tt$residential_percent_change_from_baseline,
    0.75)),
Total = runif(14,quantile(Bar_N_cases_tt$Total,0.25),
  quantile(Bar_N_cases_tt$Total,0.75)))

```

```

# h17
Bar_N_cases_fr17 <- new_data(Bar_N_cases_tr, 17) %>%
  mutate(retail_and_recreation_percent_change_from_baseline =
    runif(17,quantile(Bar_N_cases_tt$retail_and_recreation_percent_change_from_baseline,
                      0.25),
          quantile(Bar_N_cases_tt$retail_and_recreation_percent_change_from_baseline,
                    0.75)),
    grocery_and_pharmacy_percent_change_from_baseline =
    runif(17,quantile(Bar_N_cases_tt$grocery_and_pharmacy_percent_change_from_baseline,
                      0.25),
          quantile(Bar_N_cases_tt$grocery_and_pharmacy_percent_change_from_baseline,
                    0.75)),
    parks_percent_change_from_baseline =
    runif(17,quantile(Bar_N_cases_tt$parks_percent_change_from_baseline,
                      0.25),
          quantile(Bar_N_cases_tt$parks_percent_change_from_baseline,
                    0.75)),
    transit_stations_percent_change_from_baseline =
    runif(17,quantile(Bar_N_cases_tt$transit_stations_percent_change_from_baseline,
                      0.25),
          quantile(Bar_N_cases_tt$transit_stations_percent_change_from_baseline,
                    0.75)),
    workplaces_percent_change_from_baseline =
    runif(17,quantile(Bar_N_cases_tt$workplaces_percent_change_from_baseline,
                      0.25),
          quantile(Bar_N_cases_tt$workplaces_percent_change_from_baseline,
                    0.75)),
    residential_percent_change_from_baseline =
    runif(17,quantile(Bar_N_cases_tt$residential_percent_change_from_baseline,
                      0.25),
          quantile(Bar_N_cases_tt$residential_percent_change_from_baseline,
                    0.75)),
    Total = runif(17,quantile(Bar_N_cases_tt$Total,0.25),
                  quantile(Bar_N_cases_tt$Total,0.75)))

# Forecast
fc_fh7<-fabletools::forecast(fit_model, new_data = Bar_N_cases_fr7)
fc_fh14<-fabletools::forecast(fit_model, new_data = Bar_N_cases_fr14)
fc_fh17<-fabletools::forecast(fit_model, new_data = Bar_N_cases_fr17)

# Accuracy
fabletools::accuracy(fc_fh7, Bar_N_cases)

## # A tibble: 3 x 11
##   .model    sub_region_2 .type     ME   RMSE    MAE    MPE   MAPE   MASE RMSSE    ACF1
##   <chr>      <chr>     <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 arima_at1 Barcelona   Test  84.5  296.  271.  3.71  21.0  1.19  0.771  0.387
## 2 arima_at2 Barcelona   Test  184.   199.  184.  14.0  14.0  0.810  0.518 -0.101
## 3 SNaive     Barcelona   Test  377.  442.  377.  28.5  28.5  1.66  1.15   0.113
fabletools::accuracy(fc_fh14, Bar_N_cases)

## # A tibble: 3 x 11
##   .model    sub_region_2 .type     ME   RMSE    MAE    MPE   MAPE   MASE RMSSE    ACF1
##   <chr>      <chr>     <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>

```

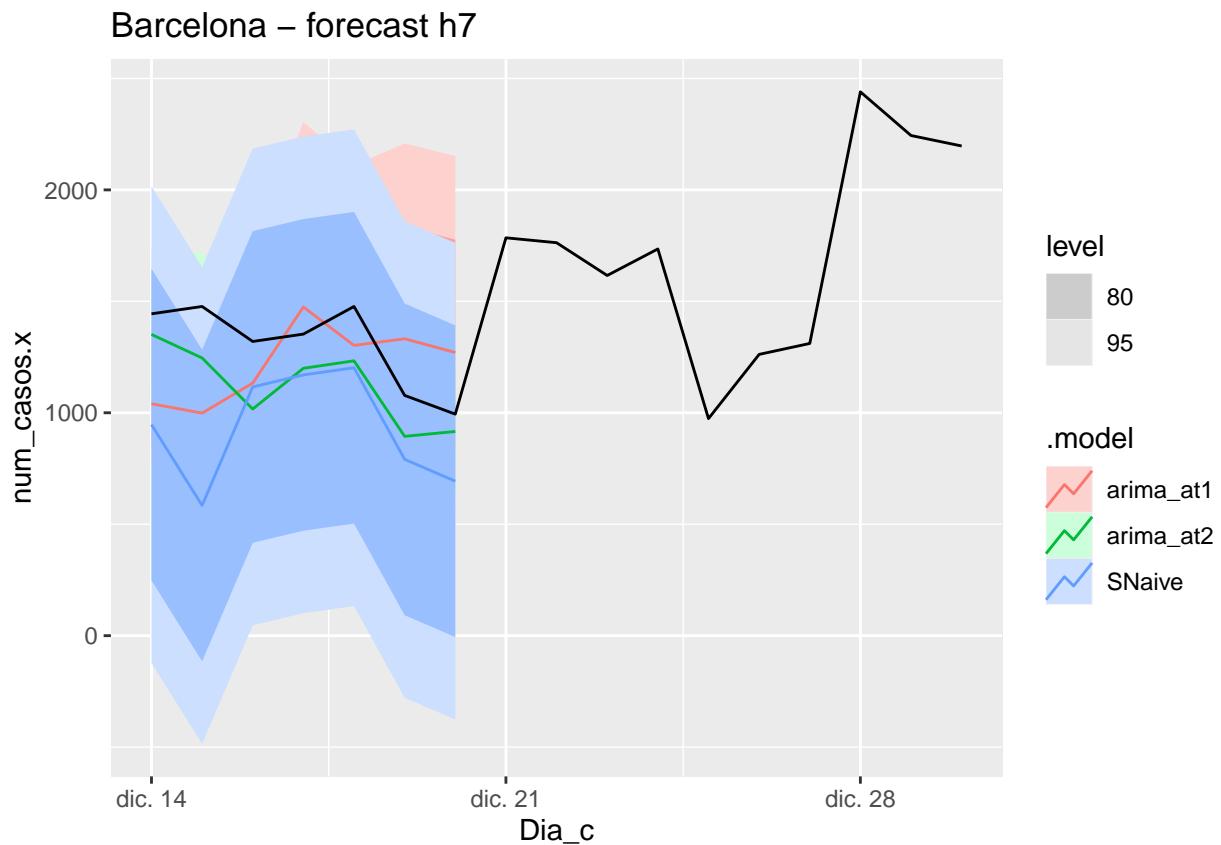
```

##   <chr>     <chr>      <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 arima_a~ Barcelona Test  25.4 373.  282. -2.75 22.6 1.24 0.972 9.18e-2
## 2 arima_a~ Barcelona Test  251. 367.  325. 15.6  23.2 1.43 0.956 -2.20e-4
## 3 SNaive    Barcelona Test  470. 579.  503. 31.3  34.6 2.22 1.51   1.60e-1
fabletools::accuracy(fc_fh17, Bar_N_cases)

## # A tibble: 3 x 11
##   .model    sub_region_2 .type     ME   RMSE   MAE   MPE   MAPE   MASE   RMSSE   ACF1
##   <chr>      <chr>      <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 arima_at1 Barcelona Test  39.8 370. 278. -2.20 17.9 1.23 0.965 0.181
## 2 arima_at2 Barcelona Test  371. 516. 404. 20.2  23.6 1.78 1.34  0.523
## 3 SNaive    Barcelona Test  636. 798. 663. 36.6  39.3 2.92 2.08  0.495

# Plots
fc_fh7 %>%
  autoplot(Bar_N_cases_tt) +
  labs(title="Barcelona – forecast h7")

```

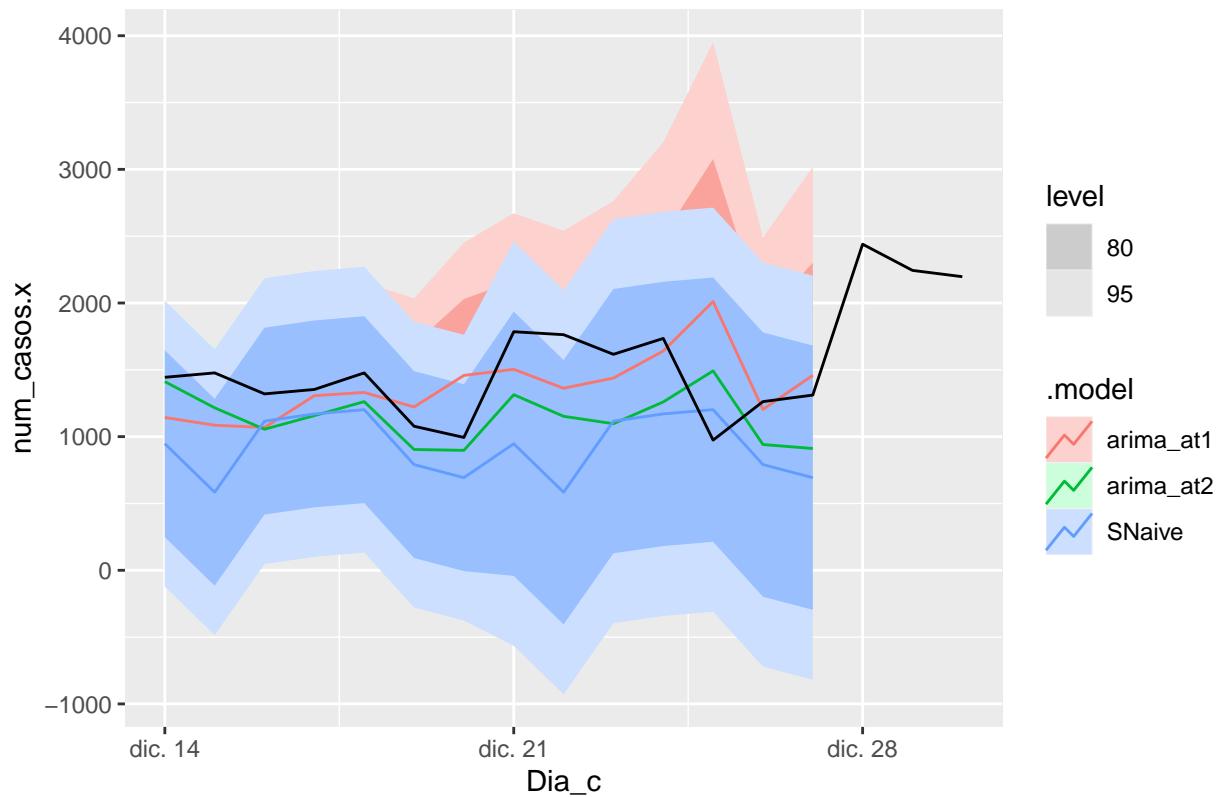


```

fc_fh14 %>%
  autoplot(Bar_N_cases_tt) +
  labs(title="Barcelona – forecast h14")

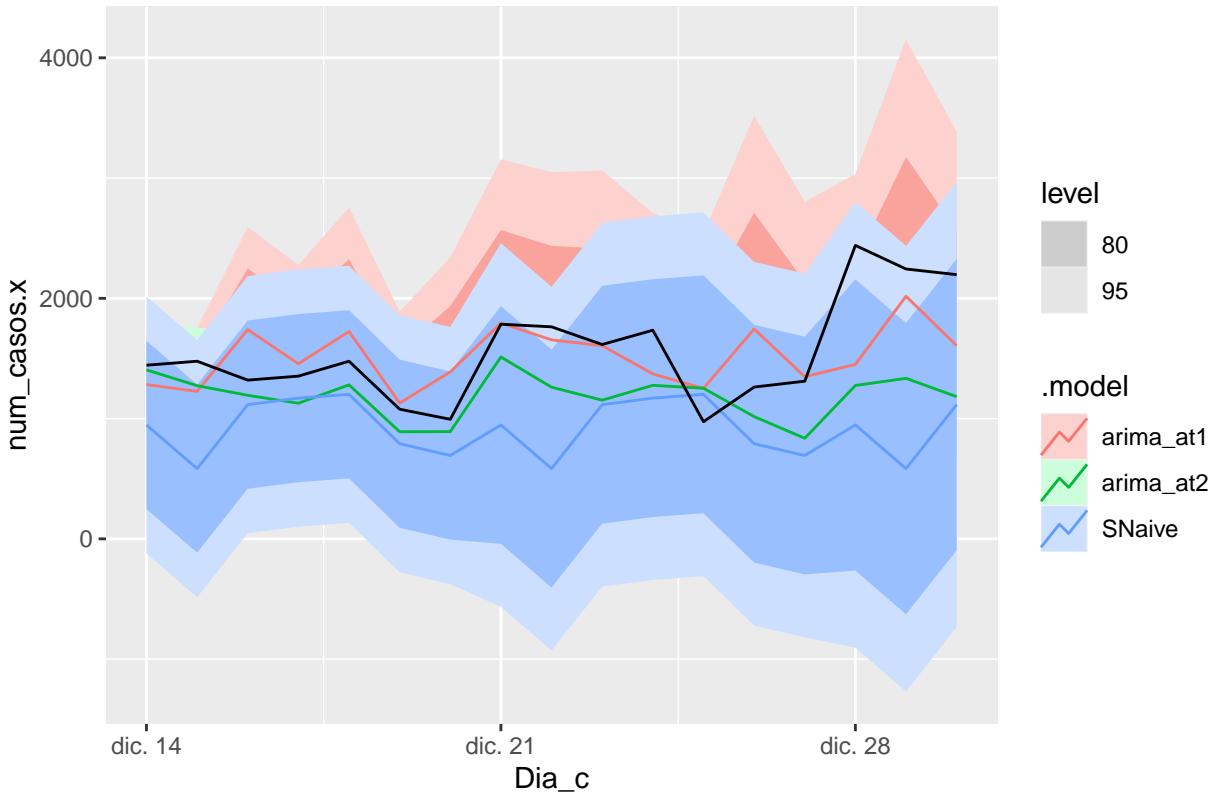
```

Barcelona – forecast h14



```
fc_fh17 %>%
  autoplot(Bar_N_cases_tt) +
  labs(title="Barcelona - forecast h17")
```

Barcelona – forecast h17



For arima_at2 - As stated by (Hyndman and Athanasopoulos 2021)... “There is clear heteroscedasticity in the residuals... The model also has some significant autocorrelation in the residuals, and the histogram of the residuals shows long tails. All of these issues with the residuals may affect the coverage of the prediction intervals, but the point forecasts should still be ok.”

```
# Model train
# Opt B
# We have added all mobility variables to models
lambda_mal_b <- Mal_N_cases %>%
  features(retail_and_recreation_percent_change_from_baseline, features = guererro) %>%
  pull(lambda_guererro)
lambda_mal_c <- Mal_N_cases %>%
  features(grocery_and_pharmacy_percent_change_from_baseline, features = guererro) %>%
  pull(lambda_guererro)
lambda_mal_d <- Mal_N_cases %>%
  features(parks_percent_change_from_baseline, features = guererro) %>%
  pull(lambda_guererro)
lambda_mal_e <- Mal_N_cases %>%
  features(transit_stations_percent_change_from_baseline, features = guererro) %>%
  pull(lambda_guererro)
lambda_mal_f <- Mal_N_cases %>%
  features(workplaces_percent_change_from_baseline, features = guererro) %>%
  pull(lambda_guererro)
lambda_mal_g <- Mal_N_cases %>%
  features(residential_percent_change_from_baseline, features = guererro) %>%
```

```

pull(lambda_guerrero)
lambda_mal_h <- Mal_N_cases %>%
  features(Total, features = guerrero) %>%
  pull(lambda_guerrero)

fit_model <- Mal_N_cases_tr %>%
  model(
    SNaive = SNAIVE(num_casos.x),
    arima_at1 = ARIMA(box_cox(num_casos.x,lambda_mal) ~
      box_cox(retail_and_recreation_percent_change_from_baseline,lambda_mal_b) +
      box_cox(grocery_and_pharmacy_percent_change_from_baseline,lambda_mal_c) +
      box_cox(parks_percent_change_from_baseline,lambda_mal_d) +
      box_cox(transit_stations_percent_change_from_baseline,lambda_mal_e) +
      box_cox(workplaces_percent_change_from_baseline,lambda_mal_f) +
      box_cox(residential_percent_change_from_baseline,lambda_mal_g) +
      box_cox(Total,lambda_mal_h)),
    arima_at2 = ARIMA(box_cox(num_casos.x,lambda_mal) ~
      box_cox(retail_and_recreation_percent_change_from_baseline,lambda_mal_b) +
      box_cox(grocery_and_pharmacy_percent_change_from_baseline,lambda_mal_c) +
      box_cox(parks_percent_change_from_baseline,lambda_mal_d) +
      box_cox(transit_stations_percent_change_from_baseline,lambda_mal_e) +
      box_cox(workplaces_percent_change_from_baseline,lambda_mal_f) +
      box_cox(residential_percent_change_from_baseline,lambda_mal_g) +
      box_cox(Total,lambda_mal_h)),
    stepwise = FALSE,
    approx = FALSE))

# Show and report model
fit_model

```

3.3.3.2 Málaga

```

## # A mable: 1 x 4
## # Key:     sub_region_2 [1]
##   sub_region_2   SNaive                      arima_at1
##   <chr>          <model>                     <model>
## 1 Málaga        <SNAIVE> <LM w/ ARIMA(1,0,1)(2,0,0)[7] errors>
## # ... with 1 more variable: arima_at2 <model>

fit_model %>% select(arima_at1) %>% report()

## Series: num_casos.x
## Model: LM w/ ARIMA(1,0,1)(2,0,0)[7] errors
## Transformation: box_cox(num_casos.x, lambda_mal)
##
## Coefficients:
##       ar1      ma1      sar1      sar2
##       0.9570 -0.5242  0.3591  0.1835
##   s.e.  0.0269  0.0850  0.1011  0.0964
##       box_cox(retail_and_recreation_percent_change_from_baseline, lambda_mal_b)
##                                         0.0030
##   s.e.
##       box_cox(grocery_and_pharmacy_percent_change_from_baseline, lambda_mal_c)
##                                         -0.0028
##   s.e.
##                                         0.0094
##   s.e.

```

```

##      box_cox(parks_percent_change_from_baseline, lambda_mal_d)
##                                         -0.0149
## s.e.                                         0.0100
##      box_cox(transit_stations_percent_change_from_baseline, lambda_mal_e)
##                                         -0.0008
## s.e.                                         0.0013
##      box_cox(workplaces_percent_change_from_baseline, lambda_mal_f)
##                                         0.0012
## s.e.                                         0.0005
##      box_cox(residential_percent_change_from_baseline, lambda_mal_g)
##                                         0.1675
## s.e.                                         0.1155
##      box_cox(Total, lambda_mal_h)  intercept
##                               0.1605    10.5888
## s.e.                               0.0564    2.5493
##
## sigma^2 estimated as 0.7138:  log likelihood=-153.38
## AIC=332.77  AICC=336.02  BIC=369.64
fit_model %>% select(arima_at2) %>% report()

## Series: num_casos.x
## Model: LM w/ ARIMA(2,0,3)(0,0,1)[7] errors
## Transformation: box_cox(num_casos.x, lambda_mal)
##
## Coefficients:
##      ar1      ar2      ma1      ma2      ma3      sma1
##      1.9343   -0.9464  -1.5603   0.4203   0.2276   0.1774
## s.e.  0.0537   0.0522   0.1128   0.2096   0.1333   0.0909
##      box_cox(retail_and_recreation_percent_change_from_baseline, lambda_mal_b)
##                                         0.0019
## s.e.                                         0.0011
##      box_cox(grocery_and_pharmacy_percent_change_from_baseline, lambda_mal_c)
##                                         0.0061
## s.e.                                         0.0073
##      box_cox(parks_percent_change_from_baseline, lambda_mal_d)
##                                         -0.0189
## s.e.                                         0.0074
##      box_cox(transit_stations_percent_change_from_baseline, lambda_mal_e)
##                                         0.0009
## s.e.                                         0.0011
##      box_cox(workplaces_percent_change_from_baseline, lambda_mal_f)
##                                         6e-04
## s.e.                                         5e-04
##      box_cox(residential_percent_change_from_baseline, lambda_mal_g)
##                                         0.2520
## s.e.                                         0.0872
##      box_cox(Total, lambda_mal_h)  intercept
##                               0.1287    11.7192
## s.e.                               0.0379    1.0469
##
## sigma^2 estimated as 0.6625:  log likelihood=-147.66
## AIC=325.31  AICC=329.68  BIC=367.86

```

```

# Good model >> Less Sigma / AICc
fit_model %>% pivot_longer(!sub_region_2,
                           names_to = "Model name",
                           values_to = "Orders")

## # A mable: 3 x 3
## # Key:   sub_region_2, Model name [3]
##   sub_region_2 `Model name`          Orders
##   <chr>           <chr>           <model>
## 1 Málaga         SNaive          <SNAIVE>
## 2 Málaga         arima_at1     <LM w/ ARIMA(1,0,1)(2,0,0)[7] errors>
## 3 Málaga         arima_at2     <LM w/ ARIMA(2,0,3)(0,0,1)[7] errors>
glance(fit_model) %>% arrange(AICc) %>% select(.model:BIC)

## # A tibble: 3 x 6
##   .model      sigma2 logLik  AIC  AICc    BIC
##   <chr>      <dbl>   <dbl> <dbl> <dbl> <dbl>
## 1 arima_at2  0.663   -148.  325.  330.  368.
## 2 arima_at1  0.714   -153.  333.  336.  370.
## 3 SNaive     5268.     NA     NA     NA     NA

# We use a Ljung-Box test >> large p-value, confirms residuals are similar to white noise.
augment(fit_model) %>%
  features(.innov, ljung_box, lag=7)

## # A tibble: 3 x 4
##   sub_region_2 .model    lb_stat lb_pvalue
##   <chr>        <chr>      <dbl>     <dbl>
## 1 Málaga      arima_at1  11.3     0.126
## 2 Málaga      arima_at2  1.48     0.983
## 3 Málaga      SNaive     178.      0

augment(fit_model) %>%
  features(.innov, ljung_box, lag=14)

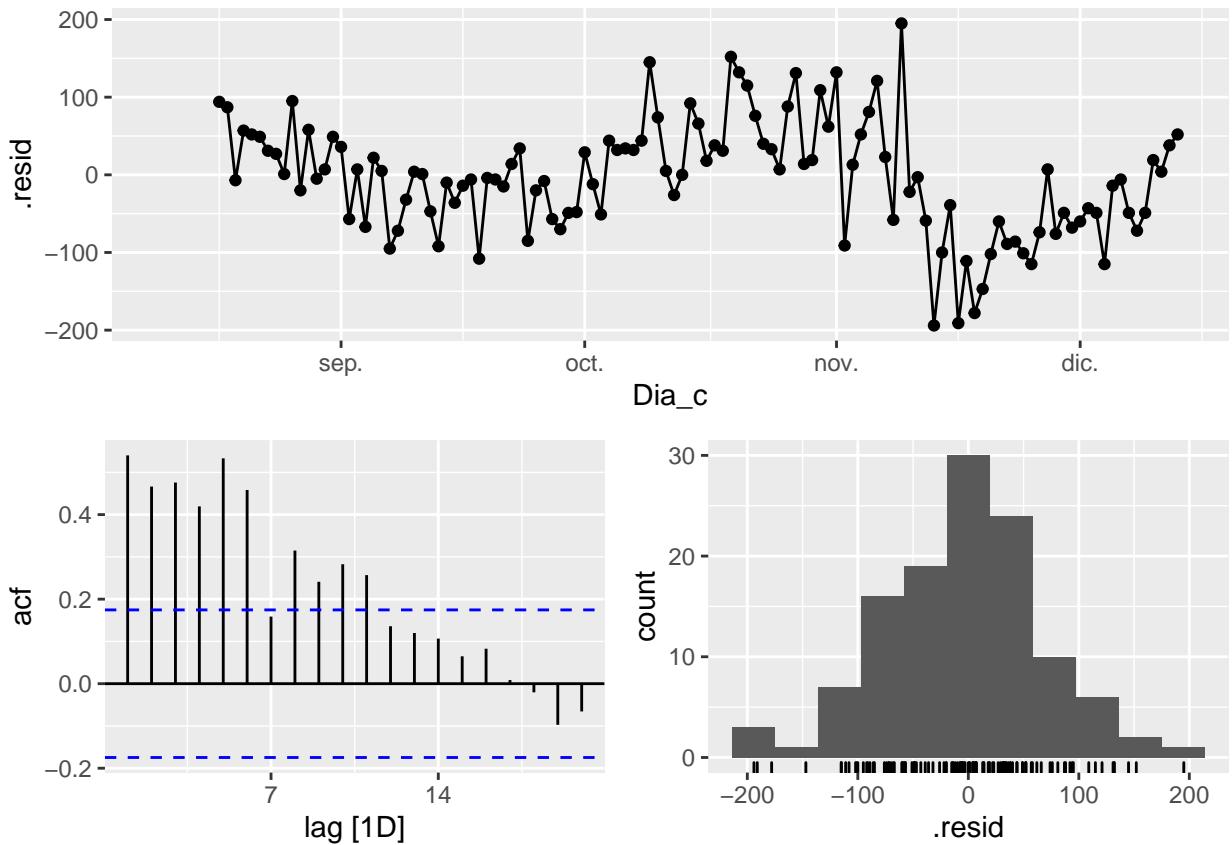
## # A tibble: 3 x 4
##   sub_region_2 .model    lb_stat lb_pvalue
##   <chr>        <chr>      <dbl>     <dbl>
## 1 Málaga      arima_at1  18.4     0.190
## 2 Málaga      arima_at2  15.4     0.352
## 3 Málaga      SNaive     224.      0

augment(fit_model) %>%
  features(.innov, ljung_box, lag=21)

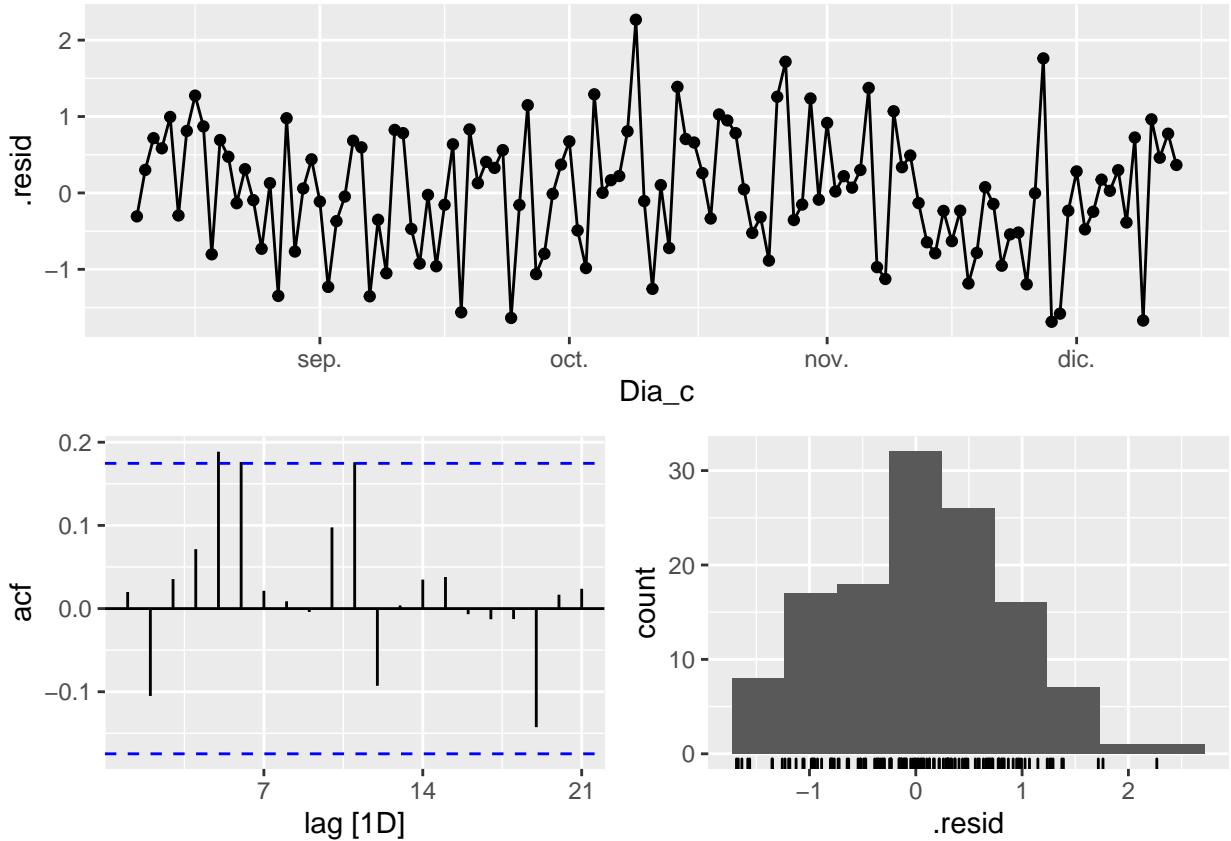
## # A tibble: 3 x 4
##   sub_region_2 .model    lb_stat lb_pvalue
##   <chr>        <chr>      <dbl>     <dbl>
## 1 Málaga      arima_at1  21.8     0.409
## 2 Málaga      arima_at2  20.2     0.510
## 3 Málaga      SNaive     229.      0

```

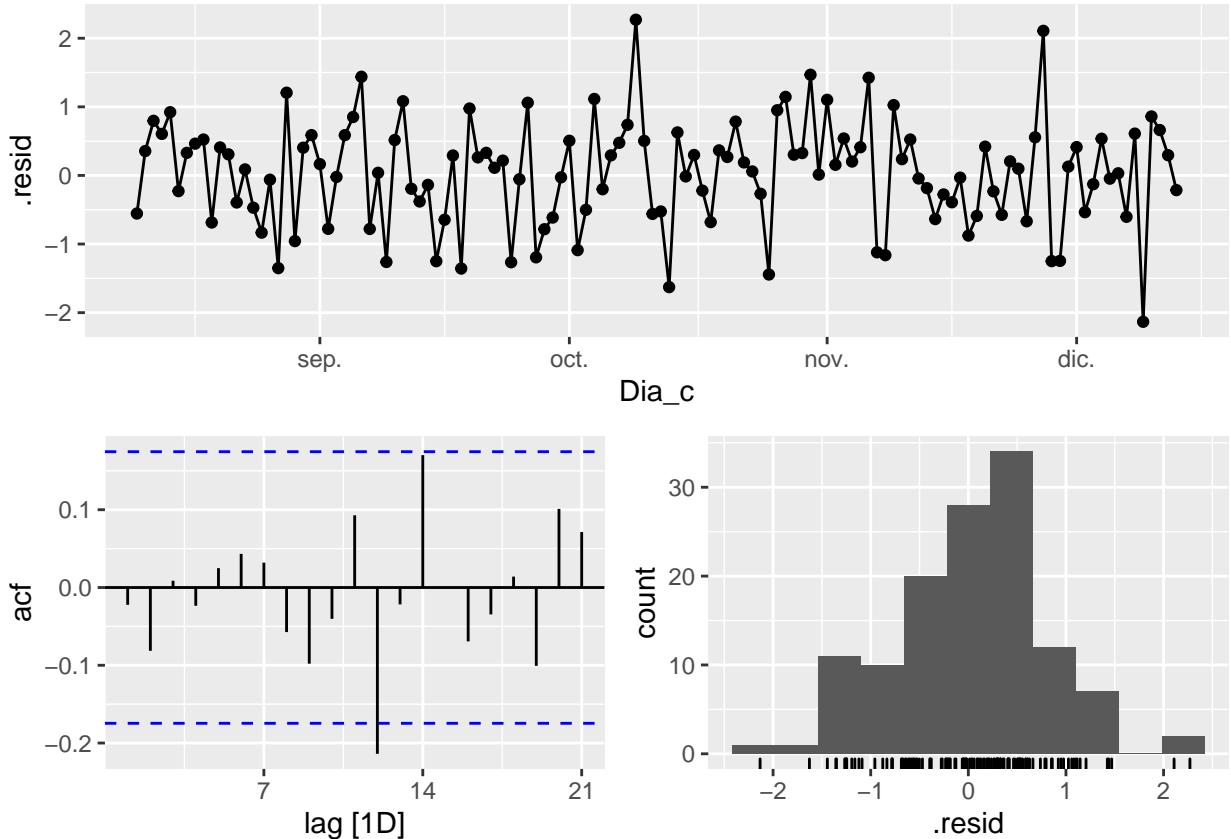
```
fit_model %>% select(SNaive) %>% gg_tsresiduals()
```



```
fit_model %>% select(arima_at1) %>% gg_tsresiduals()
```



```
fit_model %>% select(arima_at2) %>% gg_tsresiduals()
```



```

# Significant spikes out of 30 is still consistent with white noise.
# To be sure, use a Ljung-Box test, which has a large p-value, confirming that the
# residuals are similar to white noise.
# Note that the alternative models also pass this test.

# New data (dynamic regression)
# Here it is needed generate future values for the exogenous variables
# For simplicity we select a rand number included into the 2nd and
# 3rd quantile for the variable

# h7
Mal_N_cases_fr7 <- new_data(Mal_N_cases_tr, 7) %>%
  mutate(retail_and_recreation_percent_change_from_baseline =
    runif(7,quantile(Mal_N_cases_tt$retail_and_recreation_percent_change_from_baseline,
                      0.25),
          quantile(Mal_N_cases_tt$retail_and_recreation_percent_change_from_baseline,
                    0.75)),
    grocery_and_pharmacy_percent_change_from_baseline =
    runif(7,quantile(Mal_N_cases_tt$grocery_and_pharmacy_percent_change_from_baseline,
                      0.25),
          quantile(Mal_N_cases_tt$grocery_and_pharmacy_percent_change_from_baseline,
                    0.75)),
    parks_percent_change_from_baseline =
    runif(7,quantile(Mal_N_cases_tr$parks_percent_change_from_baseline,
                      0.25),
          quantile(Mal_N_cases_tt$parks_percent_change_from_baseline,
                    0.75)),
    ...
  )

```

```

        0.75)),
transit_stations_percent_change_from_baseline =
  runif(7,quantile(Mal_N_cases_tt$transit_stations_percent_change_from_baseline,
  0.25),
  quantile(Mal_N_cases_tt$transit_stations_percent_change_from_baseline,
  0.75)),
workplaces_percent_change_from_baseline =
  runif(7,quantile(Mal_N_cases_tt$workplaces_percent_change_from_baseline,
  0.25),
  quantile(Mal_N_cases_tt$workplaces_percent_change_from_baseline,
  0.75)),
residential_percent_change_from_baseline =
  runif(7,quantile(Mal_N_cases_tt$residential_percent_change_from_baseline,
  0.25),
  quantile(Mal_N_cases_tt$residential_percent_change_from_baseline,
  0.75)),
Total = runif(7,quantile(Mal_N_cases_tt$Total,0.25),
  quantile(Mal_N_cases_tt$Total,0.75)))

# h14
Mal_N_cases_fri14 <- new_data(Mal_N_cases_tr, 14) %>%
  mutate(retail_and_recreation_percent_change_from_baseline =
    runif(14,quantile(Mal_N_cases_tt$retail_and_recreation_percent_change_from_baseline,
    0.25),
    quantile(Mal_N_cases_tt$retail_and_recreation_percent_change_from_baseline,
    0.75)),
  grocery_and_pharmacy_percent_change_from_baseline =
    runif(14,quantile(Mal_N_cases_tt$grocery_and_pharmacy_percent_change_from_baseline,
    0.25),
    quantile(Mal_N_cases_tt$grocery_and_pharmacy_percent_change_from_baseline,
    0.75)),
  parks_percent_change_from_baseline =
    runif(14,quantile(Mal_N_cases_tt$parks_percent_change_from_baseline,
    0.25),
    quantile(Mal_N_cases_tt$parks_percent_change_from_baseline,
    0.75)),
  transit_stations_percent_change_from_baseline =
    runif(14,quantile(Mal_N_cases_tt$transit_stations_percent_change_from_baseline,
    0.25),
    quantile(Mal_N_cases_tt$transit_stations_percent_change_from_baseline,
    0.75)),
  workplaces_percent_change_from_baseline =
    runif(14,quantile(Mal_N_cases_tt$workplaces_percent_change_from_baseline,
    0.25),
    quantile(Mal_N_cases_tt$workplaces_percent_change_from_baseline,
    0.75)),
  residential_percent_change_from_baseline =
    runif(14,quantile(Mal_N_cases_tt$residential_percent_change_from_baseline,
    0.25),
    quantile(Mal_N_cases_tt$residential_percent_change_from_baseline,
    0.75)),
Total = runif(14,quantile(Mal_N_cases_tt$Total,0.25),
  quantile(Mal_N_cases_tt$Total,0.75)))

```

```

# h17
Mal_N_cases_fr17 <- new_data(Mal_N_cases_tr, 17) %>%
  mutate(retail_and_recreation_percent_change_from_baseline =
    runif(17,quantile(Mal_N_cases_tt$retail_and_recreation_percent_change_from_baseline,
                      0.25),
          quantile(Mal_N_cases_tt$retail_and_recreation_percent_change_from_baseline,
                    0.75)),
    grocery_and_pharmacy_percent_change_from_baseline =
    runif(17,quantile(Mal_N_cases_tt$grocery_and_pharmacy_percent_change_from_baseline,
                      0.25),
          quantile(Mal_N_cases_tt$grocery_and_pharmacy_percent_change_from_baseline,
                    0.75)),
    parks_percent_change_from_baseline =
    runif(17,quantile(Mal_N_cases_tt$parks_percent_change_from_baseline,
                      0.25),
          quantile(Mal_N_cases_tt$parks_percent_change_from_baseline,
                    0.75)),
    transit_stations_percent_change_from_baseline =
    runif(17,quantile(Mal_N_cases_tt$transit_stations_percent_change_from_baseline,
                      0.25),
          quantile(Mal_N_cases_tt$transit_stations_percent_change_from_baseline,
                    0.75)),
    workplaces_percent_change_from_baseline =
    runif(17,quantile(Mal_N_cases_tt$workplaces_percent_change_from_baseline,
                      0.25),
          quantile(Mal_N_cases_tt$workplaces_percent_change_from_baseline,
                    0.75)),
    residential_percent_change_from_baseline =
    runif(17,quantile(Mal_N_cases_tt$residential_percent_change_from_baseline,
                      0.25),
          quantile(Mal_N_cases_tt$residential_percent_change_from_baseline,
                    0.75)),
    Total = runif(17,quantile(Mal_N_cases_tt$Total,0.25),
                  quantile(Mal_N_cases_tt$Total,0.75)))

# Forecast
fc_fh7<-fabletools::forecast(fit_model, new_data = Mal_N_cases_fr7)
fc_fh14<-fabletools::forecast(fit_model, new_data = Mal_N_cases_fr14)
fc_fh17<-fabletools::forecast(fit_model, new_data = Mal_N_cases_fr17)

# Accuracy
fabletools::accuracy(fc_fh7, Mal_N_cases)

## # A tibble: 3 x 11
##   .model    sub_region_2 .type     ME   RMSE    MAE    MPE   MAPE   MASE RMSSE   ACF1
##   <chr>      <chr>     <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 arima_at1 Málaga     Test   -25.4  39.8  33.5 -19.5  23.6  0.954  0.732  0.450
## 2 arima_at2 Málaga     Test   -48.0  64.9  52.9 -35.5  37.8  1.51   1.20   0.424
## 3 SNaive     Málaga     Test   22.1   53.7  42.1  10.1   24.8  1.20   0.989  0.557
fabletools::accuracy(fc_fh14, Mal_N_cases)

## # A tibble: 3 x 11
##   .model    sub_region_2 .type     ME   RMSE    MAE    MPE   MAPE   MASE RMSSE   ACF1
##   <chr>      <chr>     <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>

```

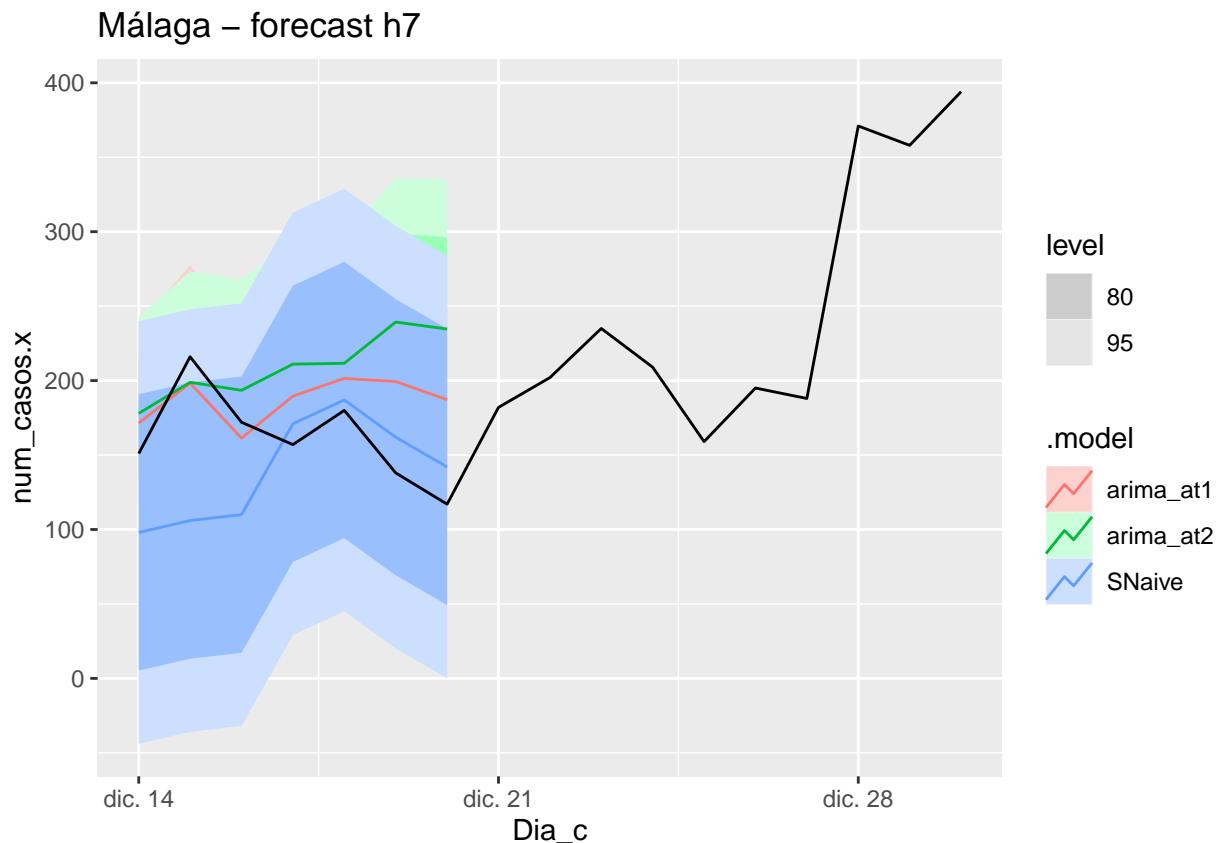
```

##   <chr>      <chr>      <chr>  <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 arima_at1 Málaga Test    -4.68  31.3  25.7 -5.42  15.4  0.734  0.576  0.103
## 2 arima_at2 Málaga Test   -58.8   70.8  62.6 -35.6  37.3  1.78   1.30   0.529
## 3 SNaive     Málaga Test    39.2   64.1  53.2  18.5  28.4  1.52   1.18   0.436
fabletools:::accuracy(fc_fh17, Mal_N_cases)

## # A tibble: 3 x 11
##   .model    sub_region_2 .type     ME   RMSE    MAE    MPE    MAPE    MASE   RMSSE   ACF1
##   <chr>      <chr>      <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 arima_at1 Málaga Test    26.1  79.4  54.6  2.98  23.1  1.56  1.46  0.539
## 2 arima_at2 Málaga Test   -39.9  72.3  62.0 -27.9  34.5  1.77  1.33  0.504
## 3 SNaive     Málaga Test    79.9 127.   91.4  28.0  36.1  2.61  2.35  0.595

# Plots
fc_fh7 %>%
  autoplot(Mal_N_cases_tt) +
  labs(title="Málaga - forecast h7")

```

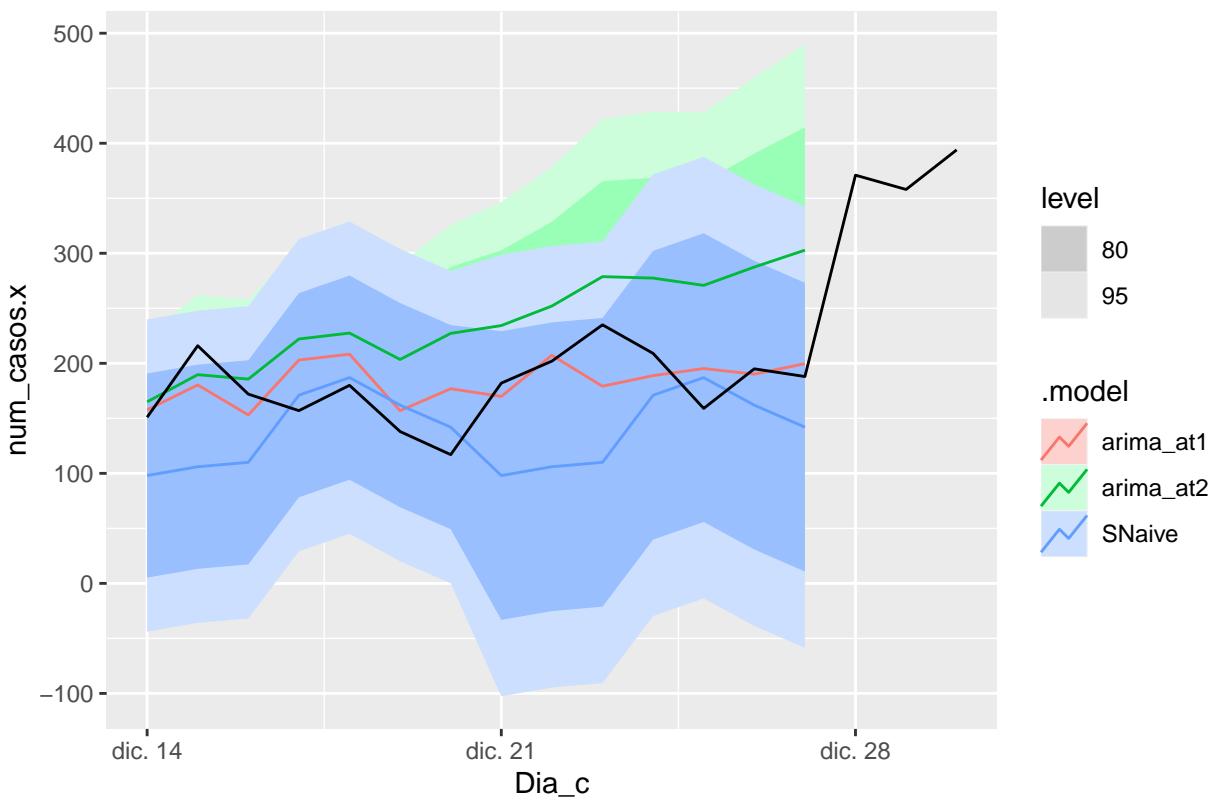


```

fc_fh14 %>%
  autoplot(Mal_N_cases_tt) +
  labs(title="Málaga - forecast h14")

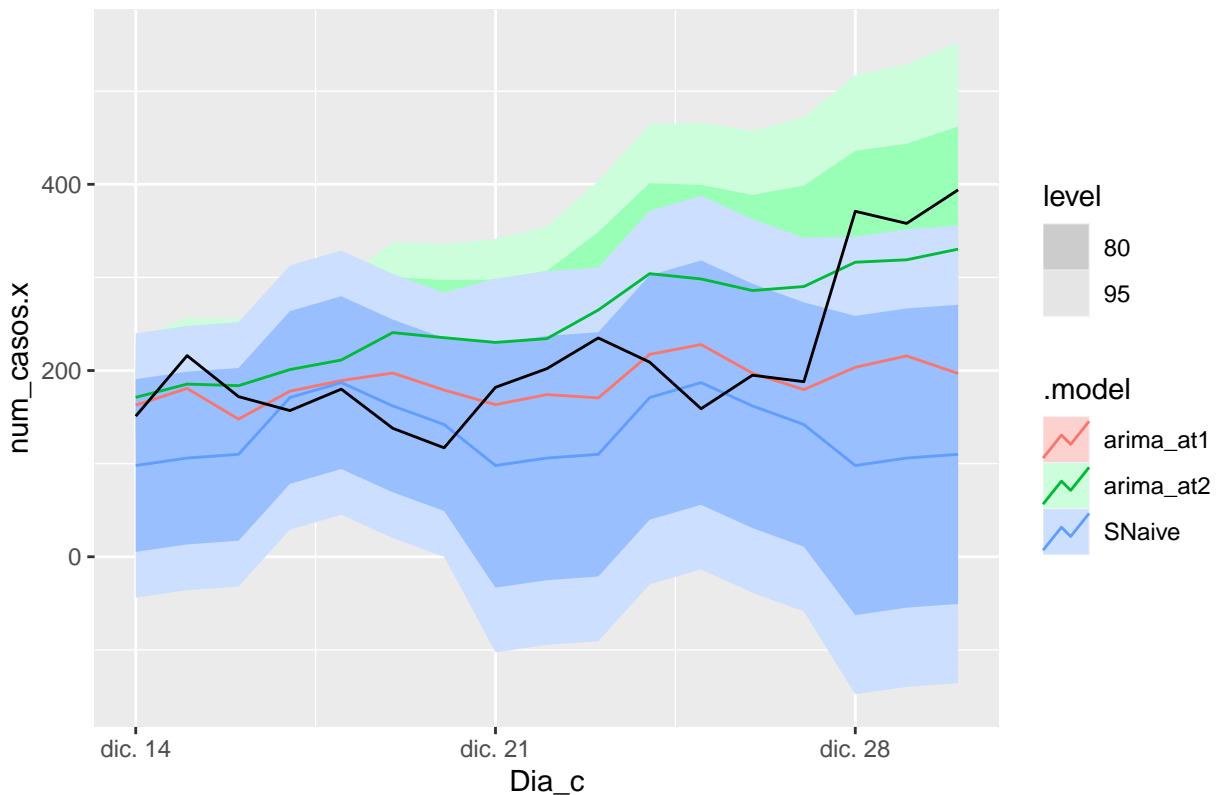
```

Málaga – forecast h14



```
fc_fh17 %>%
  autoplot(Mal_N_cases_tt) +
  labs(title="Málaga - forecast h17")
```

Málaga – forecast h17



Bibliography

- Baayen, R Harald. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics Using r*. Cambridge University Press.
- Hothorn, Torsten, and Brian S Everitt. 2014. *A Handbook of Statistical Analyses Using r*. CRC press.
- Hyndman, Rob J., and George Athanasopoulos. 2021. “Forecasting: Principles and Practice, 3rd.” OTexts: Melbourne, Australia. OTexts.com.
- Liviano Solas, Daniel, and Maria Pujol Jover. nd. *Analisis de Datos y Estadistica Descriptiva Con r y r-Commander*. UOC.
- Tector, Paul. 2011. *R Cookbook: Proven Recipes for Data Analysis, Statistics, and Graphics*. O'Reilly Media, Inc.
- Vegas Lozano, Esteban. nd. *Preprocesamiento de Los Datos*. UOC.