2- Modelación\_Datos

Angie Lorena Pradilla

28/12/2020

# Modelación de datos

En primera instancia cargamos la base de datos para realizar el modelaje y las librerías necesarias

library(readr)  
library(timeSeries)

## Loading required package: timeDate

library(ggplot2)  
library(forecast)

## Registered S3 method overwritten by 'quantmod':  
## method from  
## as.zoo.data.frame zoo

library(tseries)  
library(readr)  
library(dplyr)

##   
## Attaching package: 'dplyr'

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

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

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

options(scipen=999)  
data\_modelo <- read\_delim("C:/Users/User/Desktop/20201124-test-convocatoria/02. modelacion\_datos/data\_modelo.csv", ";", escape\_double = FALSE, trim\_ws = TRUE)

## Warning: Missing column names filled in: 'X6' [6]

##   
## -- Column specification --------------------------------------------------------  
## cols(  
## tipologia = col\_character(),  
## Fecha = col\_character(),  
## Pas = col\_number(),  
## KmT = col\_number(),  
## Gal = col\_number(),  
## X6 = col\_logical()  
## )

datos <- data\_modelo

Convertimos las fechas en formato “Date”

## [1] "Date"

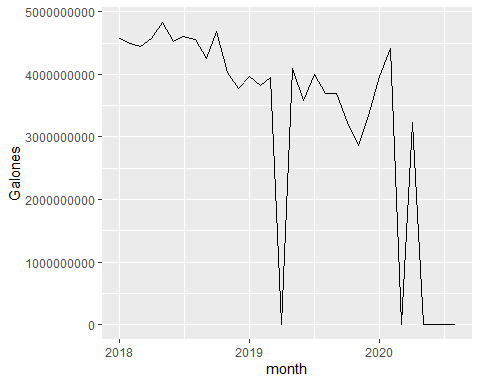
# TIPOLIGÍA T1

Filtramos por la primera tipología para realizar la predicción:

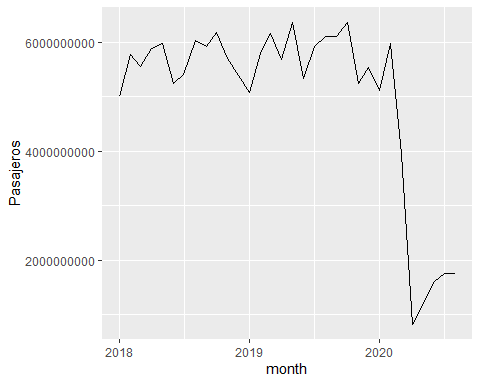
datos1 <- datos %>%  
 filter(tipologia == "T1")  
  
borrar <- c("X6")  
datos1 <- datos1[ , !(names(datos1) %in% borrar)]  
datos1$Fecha <- format(as.Date(datos1$Fecha ,format="%m/%d/%Y"),"%Y-%m-%d")  
datos1$Fecha<- as.Date(datos1$Fecha,format = "%Y-%m-%d")

A continuación se presenta el gráfico del comportamiento de las variables para esta primera tipología:

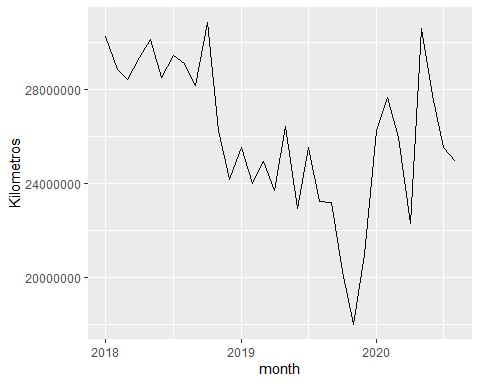
ggplot(datos1, aes(Fecha, Gal)) + geom\_line() + scale\_x\_date('month') + ylab("Galones") +  
 xlab("")



ggplot(datos1, aes(Fecha, Pas)) + geom\_line() + scale\_x\_date('month') + ylab("Pasajeros") +  
 xlab("")



ggplot(datos1, aes(Fecha, KmT)) + geom\_line() + scale\_x\_date('month') + ylab("Kilometros") +  
 xlab("")



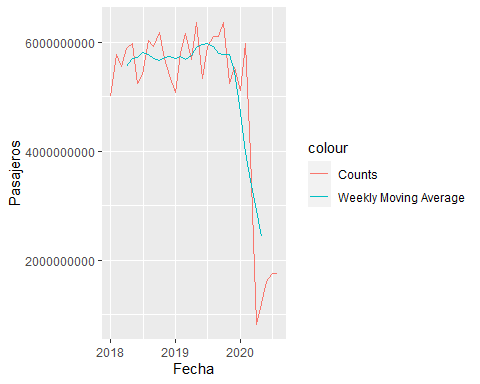
De lo anterior se puede evidenciar que de las tres variables no se tiene una tendencia clara en su comportamiento pero si un pico muy fuerte finalizando el periodo de estudio. Debido a que los datos tienen una estructura de serie temporal se procederá a tomar una media móvil, se escogerá una media movil semanal pues no hay suficientes datos y de esta manera no habrá tantos valores perdidos.

datos1$Pas7 = ma(datos1$Pas, order=7)  
  
datos1$Gal7 = ma(datos1$Gal, order=7)  
  
datos1$KmT7 = ma(datos1$KmT, order=7)

Una vez hallada la media móvil, se evidencia en el gráfico la serie suavizada vs la serie original

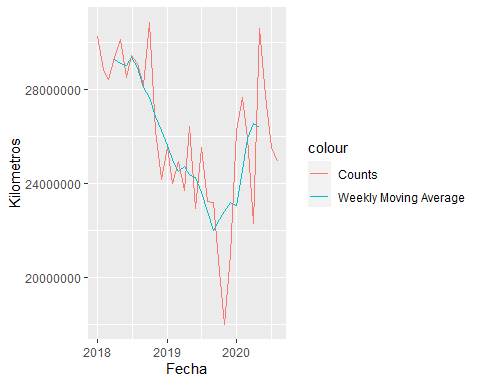
ggplot() +  
 geom\_line(data = datos1, aes(x = Fecha, y = Pas, colour = "Counts")) +  
 geom\_line(data = datos1, aes(x = Fecha, y = Pas7, colour = "Weekly Moving Average")) +  
 ylab('Pasajeros')

## Warning: Removed 6 row(s) containing missing values (geom\_path).



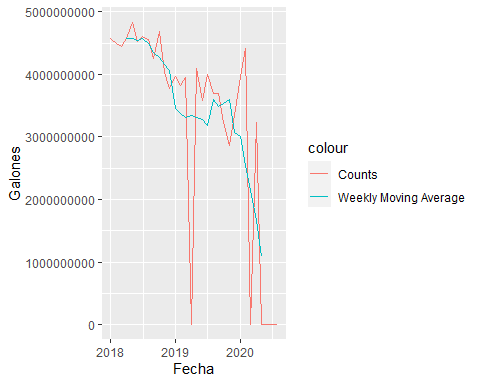
ggplot() +  
 geom\_line(data = datos1, aes(x = Fecha, y = KmT, colour = "Counts")) +  
 geom\_line(data = datos1, aes(x = Fecha, y = KmT7, colour = "Weekly Moving Average")) +  
 ylab('Kilometros')

## Warning: Removed 6 row(s) containing missing values (geom\_path).



ggplot() +  
 geom\_line(data = datos1, aes(x = Fecha, y = Gal, colour = "Counts")) +  
 geom\_line(data = datos1, aes(x = Fecha, y = Gal7, colour = "Weekly Moving Average")) +  
 ylab('Galones')

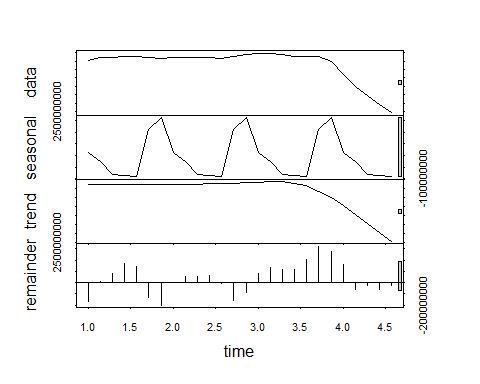
## Warning: Removed 6 row(s) containing missing values (geom\_path).



## Variable “Pas”

Comenzamos con la revisión de la variable “Pas” y hallamos el componente estacional de la serie mendiante suavizado

count\_ma = ts(na.omit(datos1$Pas7), frequency=7)  
decomp = stl(count\_ma, s.window="periodic")  
deseasonal\_Pas <- seasadj(decomp)  
plot(decomp)



De acuerdo al gráfico se puede concluir que la serie no presenta estacionariedad, tampoco una tendencia clara pero si un pico descendente significativo al final del periodo.

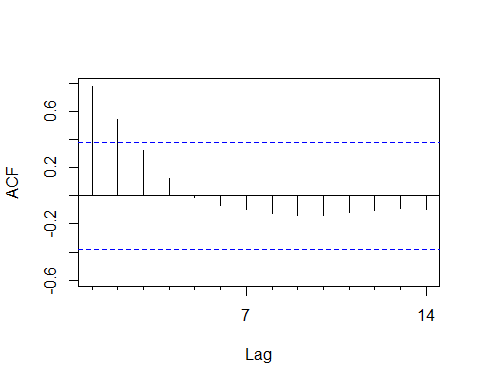
La instalación de un modelo ARIMA requiere que la serie sea estacionaria e dice que una serie es estacionaria cuando su media, varianza y autocovarianza son invariantes en el tiempo, para confirmar la estacionalidad de la serie, corremos la prueba de dickey - Fuller aumentada:

adf.test(count\_ma, alternative = "stationary")

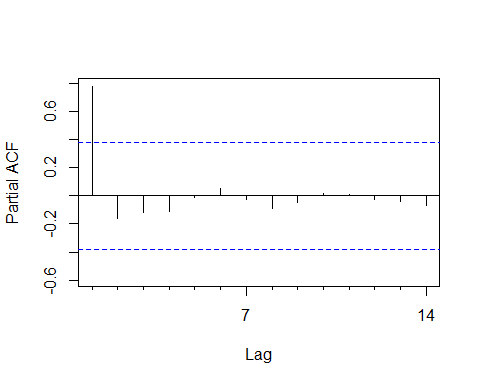
## Warning in adf.test(count\_ma, alternative = "stationary"): p-value greater than  
## printed p-value

##   
## Augmented Dickey-Fuller Test  
##   
## data: count\_ma  
## Dickey-Fuller = 0.47837, Lag order = 2, p-value = 0.99  
## alternative hypothesis: stationary

Acf(count\_ma, main='')

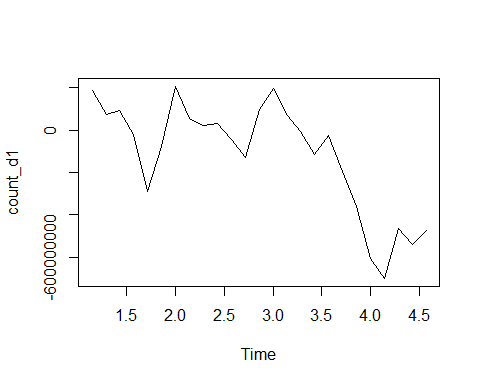


Pacf(count\_ma, main='')



Debido al que el p - valor es mayor al 5% de significancia, no se rechaza la hipótesis nula, lo que quiere decir que no es estacionaria. Como la serie no es estacionaria es necesario diferenciar, al diferenciar el modelo que se ajusta es un modelo ARIMA.

count\_d1 = diff(deseasonal\_Pas, differences = 1)  
plot(count\_d1)

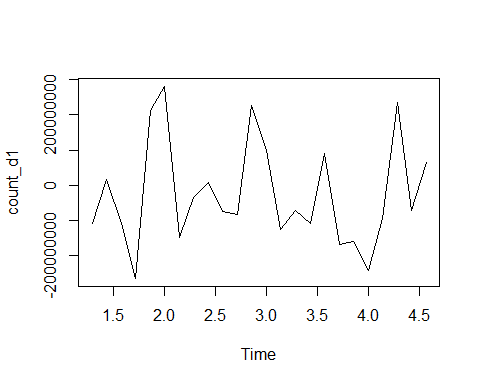


adf.test(count\_d1, alternative = "stationary")

##   
## Augmented Dickey-Fuller Test  
##   
## data: count\_d1  
## Dickey-Fuller = -1.6241, Lag order = 2, p-value = 0.7156  
## alternative hypothesis: stationary

Con la primera diferenciación la serie aun sigue sin tener estacionariedad, se vuelve a realizar una segunda diferenciación:

count\_d1 = diff(deseasonal\_Pas, differences = 2)  
plot(count\_d1)



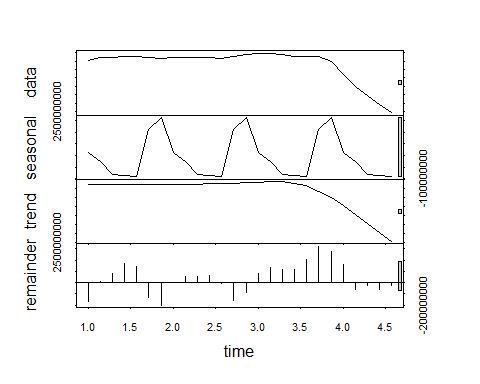
adf.test(count\_d1, alternative = "stationary")

##   
## Augmented Dickey-Fuller Test  
##   
## data: count\_d1  
## Dickey-Fuller = -2.6578, Lag order = 2, p-value = 0.3218  
## alternative hypothesis: stationary

Aun diferenciando 2 veces la serie sigue sin tener estacionariedad, se debe tener en cuenta que las veces máximas para diferenciar una serie son 2 veces debido a la pérdida de datos que se genera. De acuerdo a lo anterior, no se puede ajustar un modelo arima para esta variable.

## Variable Gal

count\_ma1 = ts(na.omit(datos1$Gal7), frequency=7)  
decomp1 = stl(count\_ma1, s.window="periodic")  
deseasonal\_Gal <- seasadj(decomp1)  
plot(decomp)



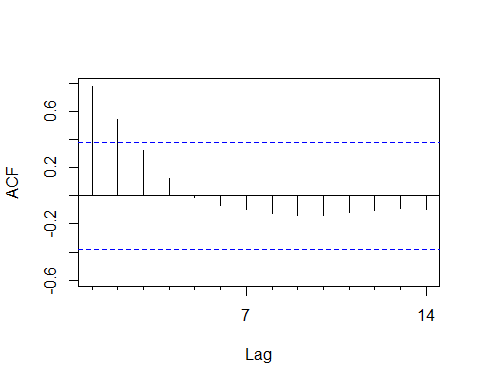
De acuerdo al gráfico se puede concluir que la serie no presenta estacionariedad, y una tendencia suave hacia la baja y pero si un pico descendente significativo al final del periodo.

La instalación de un modelo ARIMA requiere que la serie sea estacionaria e dice que una serie es estacionaria cuando su media, varianza y autocovarianza son invariantes en el tiempo, para confirmar la estacionalidad de la serie, corremos la prueba de dickey - Fuller aumentada:

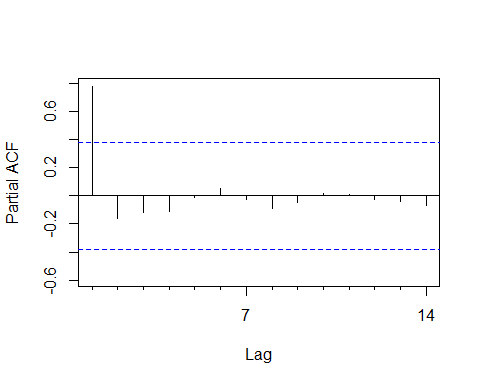
adf.test(count\_ma1, alternative = "stationary")

##   
## Augmented Dickey-Fuller Test  
##   
## data: count\_ma1  
## Dickey-Fuller = -1.254, Lag order = 2, p-value = 0.8566  
## alternative hypothesis: stationary

Acf(count\_ma, main='')

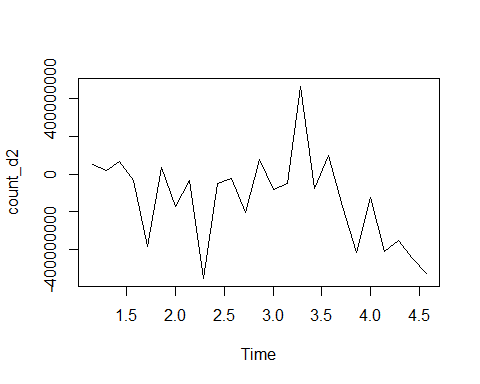


Pacf(count\_ma, main='')



Debido al que el p - valor es mayor al 5% de significancia, no se rechaza la hipótesis nula, lo que quiere decir que no es estacionaria. Como la serie no es estacionaria es necesario diferenciar, al diferenciar el modelo que se ajusta es un modelo ARIMA.

count\_d2 = diff(deseasonal\_Gal, differences = 1)  
plot(count\_d2)

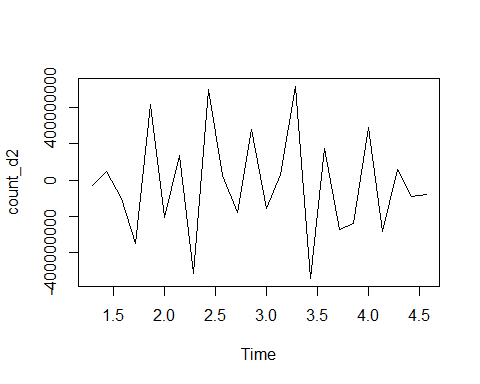


adf.test(count\_d2, alternative = "stationary")

##   
## Augmented Dickey-Fuller Test  
##   
## data: count\_d2  
## Dickey-Fuller = -1.1609, Lag order = 2, p-value = 0.892  
## alternative hypothesis: stationary

Con la primera diferenciación la serie aun sigue sin tener estacionariedad, se vuelve a realizar una segunda diferenciación:

count\_d2 = diff(deseasonal\_Gal, differences = 2)  
plot(count\_d2)



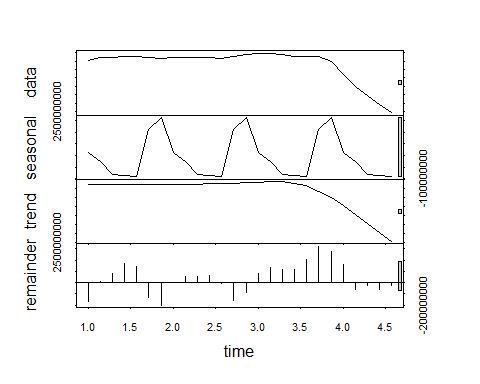
adf.test(count\_d2, alternative = "stationary")

##   
## Augmented Dickey-Fuller Test  
##   
## data: count\_d2  
## Dickey-Fuller = -2.9818, Lag order = 2, p-value = 0.1984  
## alternative hypothesis: stationary

Aun diferenciando 2 veces la serie sigue sin tener estacionariedad, se debe tener en cuenta que las veces máximas para diferenciar una serie son 2 veces debido a la pérdida de datos que se genera. De acuerdo a lo anterior, no se puede ajustar un modelo arima para esta variable.

## Variable KmT

count\_ma2 = ts(na.omit(datos1$KmT7), frequency=7)  
decomp2 = stl(count\_ma2, s.window="periodic")  
deseasonal\_KmT <- seasadj(decomp2)  
plot(decomp)



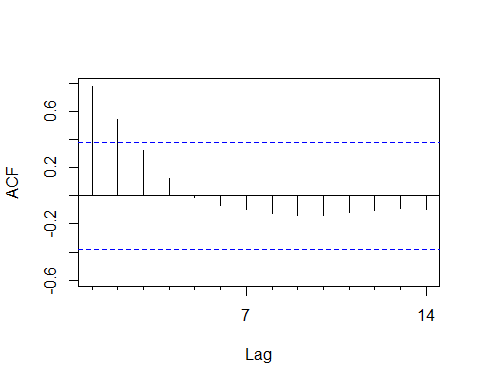
De acuerdo al gráfico se puede concluir que la serie no presenta estacionariedad, y una tendencia suave hacia la baja y pero si un pico descendente significativo al final del periodo.

La instalación de un modelo ARIMA requiere que la serie sea estacionaria e dice que una serie es estacionaria cuando su media, varianza y autocovarianza son invariantes en el tiempo, para confirmar la estacionalidad de la serie, corremos la prueba de dickey - Fuller aumentada:

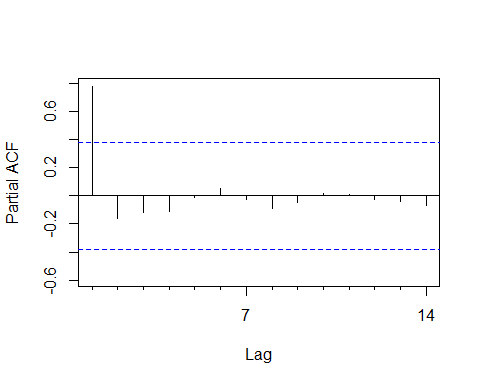
adf.test(count\_ma2, alternative = "stationary")

##   
## Augmented Dickey-Fuller Test  
##   
## data: count\_ma2  
## Dickey-Fuller = -0.57714, Lag order = 2, p-value = 0.9686  
## alternative hypothesis: stationary

Acf(count\_ma, main='')

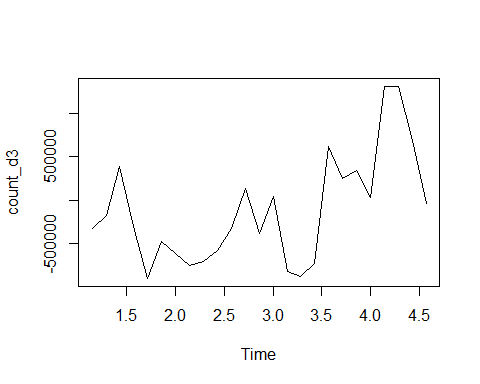


Pacf(count\_ma, main='')



Debido al que el p - valor es mayor al 5% de significancia, no se rechaza la hipótesis nula, lo que quiere decir que no es estacionaria. Como la serie no es estacionaria es necesario diferenciar, al diferenciar el modelo que se ajusta es un modelo ARIMA.

count\_d3 = diff(deseasonal\_KmT, differences = 1)  
plot(count\_d3)

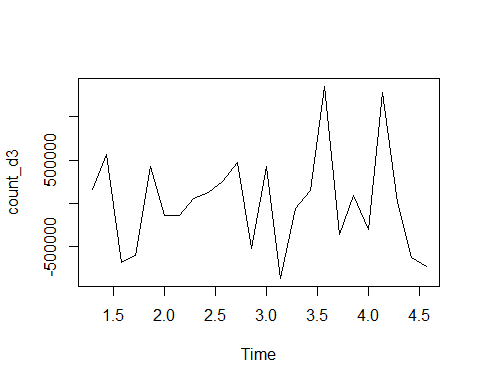


adf.test(count\_d3, alternative = "stationary")

##   
## Augmented Dickey-Fuller Test  
##   
## data: count\_d3  
## Dickey-Fuller = -3.3361, Lag order = 2, p-value = 0.08665  
## alternative hypothesis: stationary

Con la primera diferenciación la serie aun sigue sin tener estacionariedad al 5% pero tiene estacionariedad al 10% de significancia, se vuelve a realizar una segunda diferenciación:

count\_d3 = diff(deseasonal\_KmT, differences = 2)  
plot(count\_d3)

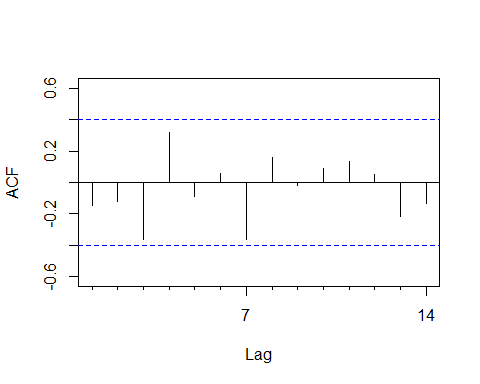


adf.test(count\_d3, alternative = "stationary")

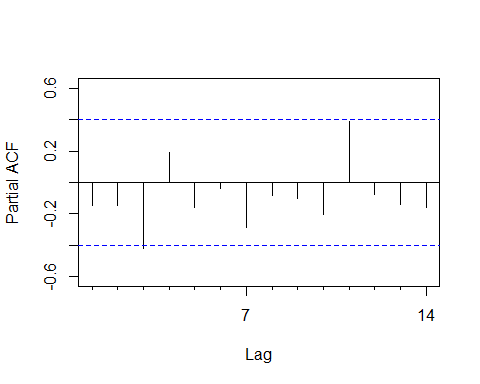
## Warning in adf.test(count\_d3, alternative = "stationary"): p-value smaller than  
## printed p-value

##   
## Augmented Dickey-Fuller Test  
##   
## data: count\_d3  
## Dickey-Fuller = -4.7444, Lag order = 2, p-value = 0.01  
## alternative hypothesis: stationary

Acf(count\_d3, main='')



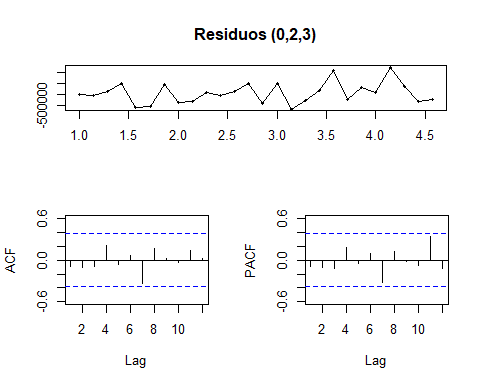
Pacf(count\_d3, main='')

 Se observa que a la segunda diferenciación se rechaza la hipótesis nula confirmando que la serie es estacionaria, de esta manera, según la función de autocorrelación y la autocorrelación parcial, el orden p del modelo es 0 y el orden q es 3.

arima1<- arima(deseasonal\_KmT,order=c(0,2,3), method = "ML")  
arima1$aic

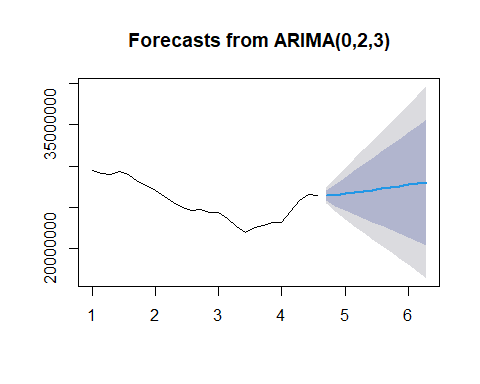
## [1] 706.2143

tsdisplay(residuals(arima1), lag.max=12, main='Residuos (0,2,3)')



De acuerdo al modelo anterior, a continuación se presenta la predicción para la variable Kilometráje en la tipología 1.

prediccion <- forecast(arima1, h=12)  
plot(prediccion)



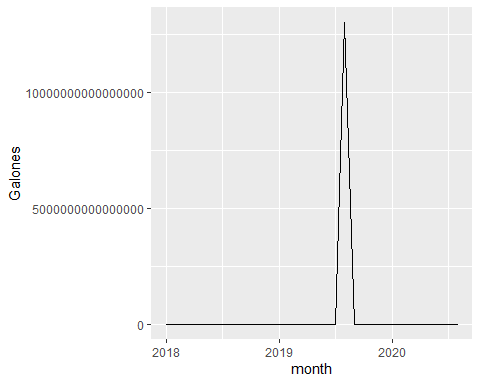
# TIPOLIGÍA T2

Filtramos por la primera tipología para realizar la predicción:

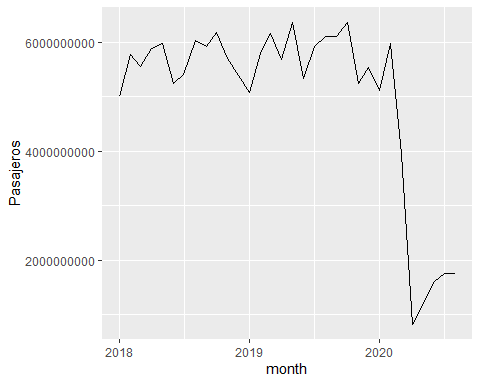
datos2 <- datos %>%  
 filter(tipologia == "T2")  
  
borrar <- c("X6")  
datos2 <- datos2[ , !(names(datos2) %in% borrar)]  
datos2$Fecha <- format(as.Date(datos2$Fecha ,format="%m/%d/%Y"),"%Y-%m-%d")  
datos2$Fecha<- as.Date(datos2$Fecha,format = "%Y-%m-%d")

A continuación se presenta el gráfico del comportamiento de las variables para esta primera tipología:

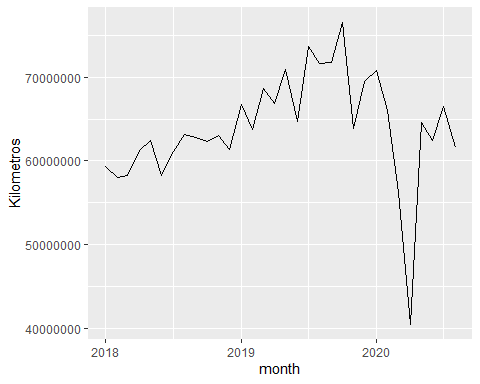
ggplot(datos2, aes(Fecha, Gal)) + geom\_line() + scale\_x\_date('month') + ylab("Galones") +  
 xlab("")



ggplot(datos2, aes(Fecha, Pas)) + geom\_line() + scale\_x\_date('month') + ylab("Pasajeros") +  
 xlab("")



ggplot(datos2, aes(Fecha, KmT)) + geom\_line() + scale\_x\_date('month') + ylab("Kilometros") +  
 xlab("")



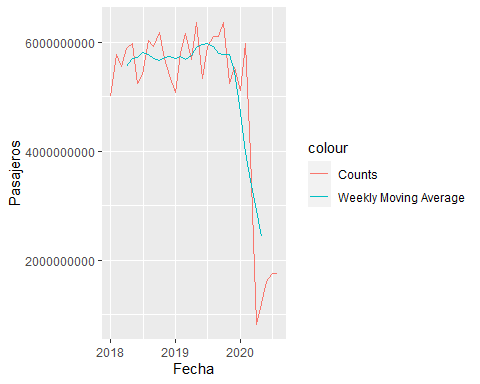
De lo anterior se puede evidenciar que de las tres variables no se tiene una tendencia clara en su comportamiento pero si un pico muy fuerte finalizando el periodo de estudio. Debido a que los datos tienen una estructura de serie temporal se procederá a tomar una media móvil, se escogerá una media movil semanal pues no hay suficientes datos y de esta manera no habrá tantos valores perdidos.

datos2$Pas7 = ma(datos2$Pas, order=7)  
  
datos2$Gal7 = ma(datos2$Gal, order=7)  
  
datos2$KmT7 = ma(datos2$KmT, order=7)

Una vez hallada la media móvil, se evidencia en el gráfico la serie suavizada vs la serie original

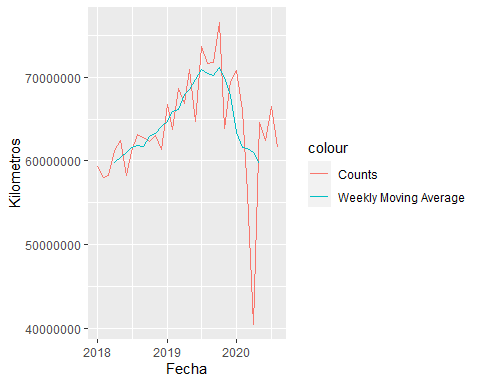
ggplot() +  
 geom\_line(data = datos2, aes(x = Fecha, y = Pas, colour = "Counts")) +  
 geom\_line(data = datos2, aes(x = Fecha, y = Pas7, colour = "Weekly Moving Average")) +  
 ylab('Pasajeros')

## Warning: Removed 6 row(s) containing missing values (geom\_path).



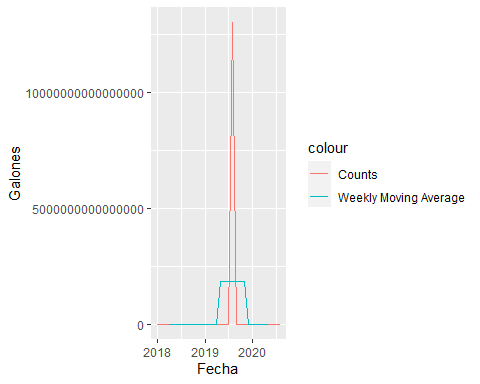
ggplot() +  
 geom\_line(data = datos2, aes(x = Fecha, y = KmT, colour = "Counts")) +  
 geom\_line(data = datos2, aes(x = Fecha, y = KmT7, colour = "Weekly Moving Average")) +  
 ylab('Kilometros')

## Warning: Removed 6 row(s) containing missing values (geom\_path).



ggplot() +  
 geom\_line(data = datos2, aes(x = Fecha, y = Gal, colour = "Counts")) +  
 geom\_line(data = datos2, aes(x = Fecha, y = Gal7, colour = "Weekly Moving Average")) +  
 ylab('Galones')

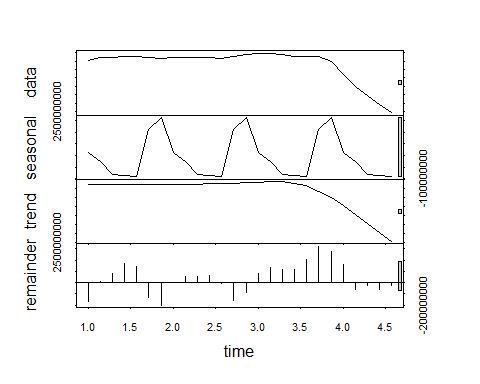
## Warning: Removed 6 row(s) containing missing values (geom\_path).



## Variable “Pas”

Comenzamos con la revisión de la variable “Pas” y hallamos el componente estacional de la serie mendiante suavizado

count\_ma = ts(na.omit(datos2$Pas7), frequency=7)  
decomp = stl(count\_ma, s.window="periodic")  
deseasonal\_Pas <- seasadj(decomp)  
plot(decomp)



De acuerdo al gráfico se puede concluir que la serie no presenta estacionariedad, tampoco una tendencia clara pero si un pico descendente significativo al final del periodo.

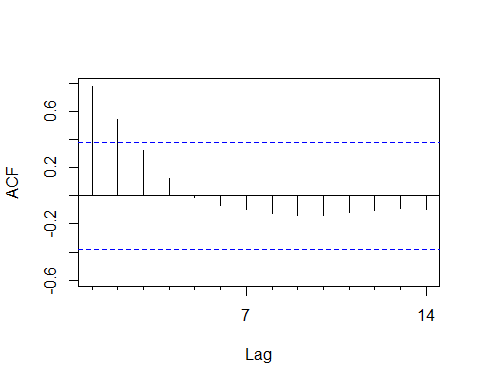
La instalación de un modelo ARIMA requiere que la serie sea estacionaria e dice que una serie es estacionaria cuando su media, varianza y autocovarianza son invariantes en el tiempo, para confirmar la estacionalidad de la serie, corremos la prueba de dickey - Fuller aumentada:

adf.test(count\_ma, alternative = "stationary")

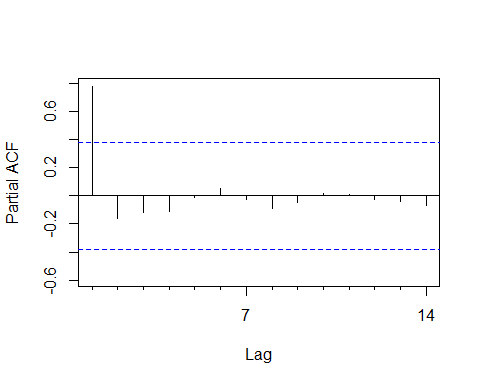
## Warning in adf.test(count\_ma, alternative = "stationary"): p-value greater than  
## printed p-value

##   
## Augmented Dickey-Fuller Test  
##   
## data: count\_ma  
## Dickey-Fuller = 0.47837, Lag order = 2, p-value = 0.99  
## alternative hypothesis: stationary

Acf(count\_ma, main='')

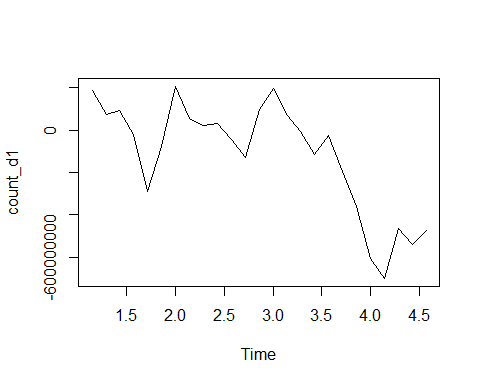


Pacf(count\_ma, main='')



Debido al que el p - valor es mayor al 5% de significancia, no se rechaza la hipótesis nula, lo que quiere decir que no es estacionaria. Como la serie no es estacionaria es necesario diferenciar, al diferenciar el modelo que se ajusta es un modelo ARIMA.

count\_d1 = diff(deseasonal\_Pas, differences = 1)  
plot(count\_d1)

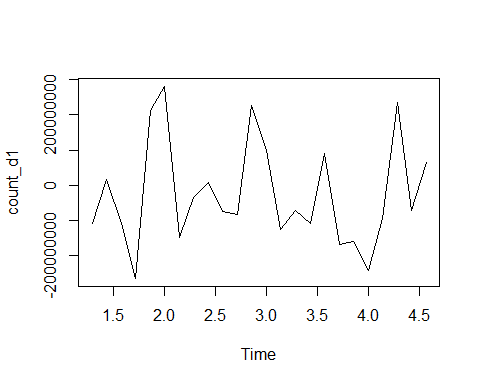


adf.test(count\_d1, alternative = "stationary")

##   
## Augmented Dickey-Fuller Test  
##   
## data: count\_d1  
## Dickey-Fuller = -1.6241, Lag order = 2, p-value = 0.7156  
## alternative hypothesis: stationary

Con la primera diferenciación la serie aun sigue sin tener estacionariedad, se vuelve a realizar una segunda diferenciación:

count\_d1 = diff(deseasonal\_Pas, differences = 2)  
plot(count\_d1)



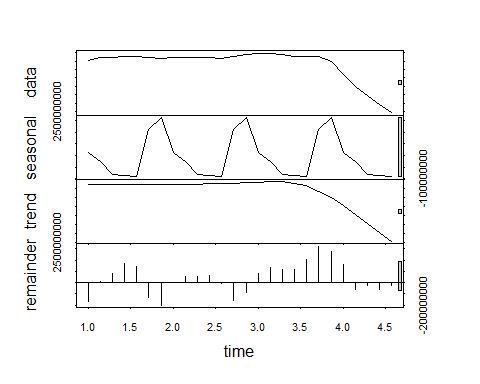
adf.test(count\_d1, alternative = "stationary")

##   
## Augmented Dickey-Fuller Test  
##   
## data: count\_d1  
## Dickey-Fuller = -2.6578, Lag order = 2, p-value = 0.3218  
## alternative hypothesis: stationary

Aun diferenciando 2 veces la serie sigue sin tener estacionariedad, se debe tener en cuenta que las veces máximas para diferenciar una serie son 2 veces debido a la pérdida de datos que se genera. De acuerdo a lo anterior, no se puede ajustar un modelo arima para esta variable.

## Variable Gal

count\_ma1 = ts(na.omit(datos2$Gal7), frequency=7)  
decomp1 = stl(count\_ma1, s.window="periodic")  
deseasonal\_Gal <- seasadj(decomp1)  
plot(decomp)



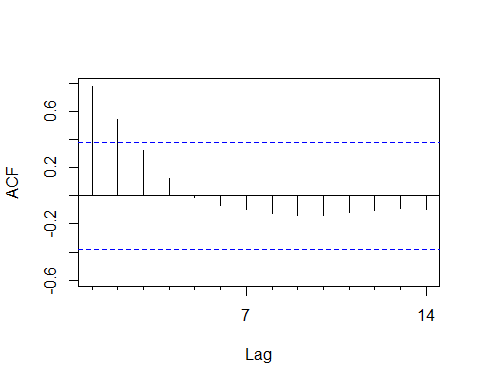
De acuerdo al gráfico se puede concluir que la serie no presenta estacionariedad, y una tendencia suave hacia la baja y pero si un pico descendente significativo al final del periodo.

La instalación de un modelo ARIMA requiere que la serie sea estacionaria e dice que una serie es estacionaria cuando su media, varianza y autocovarianza son invariantes en el tiempo, para confirmar la estacionalidad de la serie, corremos la prueba de dickey - Fuller aumentada:

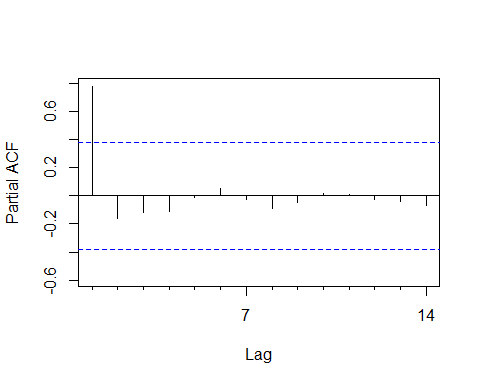
adf.test(count\_ma1, alternative = "stationary")

##   
## Augmented Dickey-Fuller Test  
##   
## data: count\_ma1  
## Dickey-Fuller = -1.4649, Lag order = 2, p-value = 0.7762  
## alternative hypothesis: stationary

Acf(count\_ma, main='')

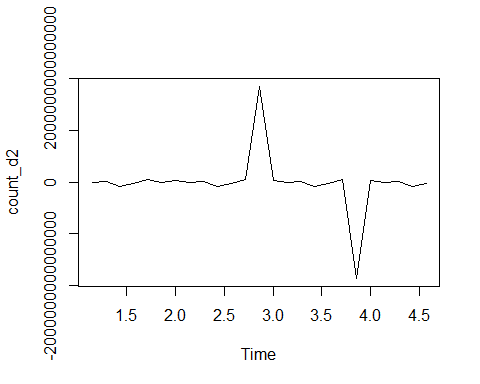


Pacf(count\_ma, main='')



Debido al que el p - valor es mayor al 5% de significancia, no se rechaza la hipótesis nula, lo que quiere decir que no es estacionaria. Como la serie no es estacionaria es necesario diferenciar, al diferenciar el modelo que se ajusta es un modelo ARIMA.

count\_d2 = diff(deseasonal\_Gal, differences = 1)  
plot(count\_d2)

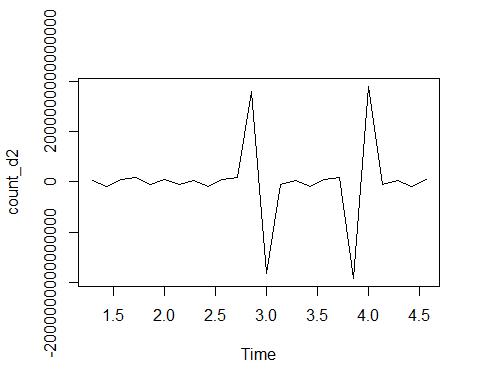


adf.test(count\_d2, alternative = "stationary")

##   
## Augmented Dickey-Fuller Test  
##   
## data: count\_d2  
## Dickey-Fuller = -2.5493, Lag order = 2, p-value = 0.3631  
## alternative hypothesis: stationary

Con la primera diferenciación la serie aun sigue sin tener estacionariedad, se vuelve a realizar una segunda diferenciación:

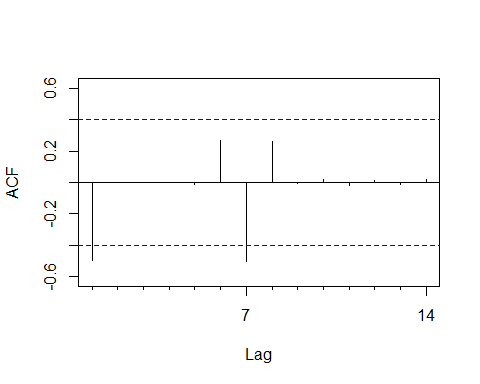
count\_d2 = diff(deseasonal\_Gal, differences = 2)  
plot(count\_d2)



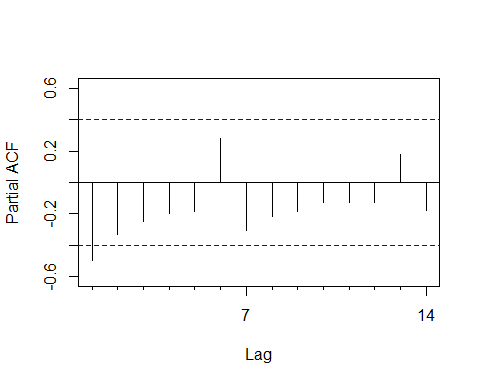
adf.test(count\_d2, alternative = "stationary")

##   
## Augmented Dickey-Fuller Test  
##   
## data: count\_d2  
## Dickey-Fuller = -3.9989, Lag order = 2, p-value = 0.02329  
## alternative hypothesis: stationary

Acf(count\_d2, main='')



Pacf(count\_d2, main='')

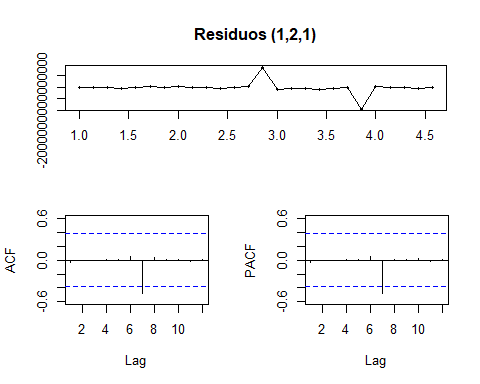


Se observa que a la segunda diferenciación se rechaza la hipótesis nula confirmando que la serie es estacionaria, de esta manera, según la función de autocorrelación y la autocorrelación parcial, el orden p del modelo es 1 y el orden q es 1.

arima2<- arima(deseasonal\_Gal,order=c(1,2,1), method = "ML")  
arima2$aic

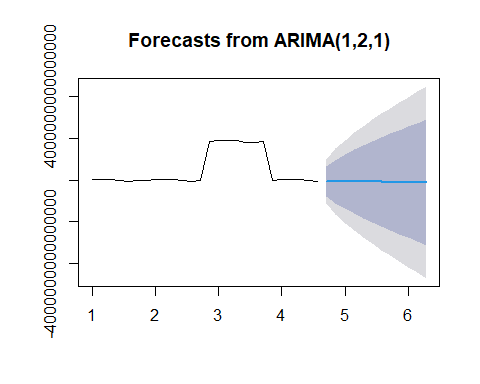
## [1] 1705.408

tsdisplay(residuals(arima2), lag.max=12, main='Residuos (1,2,1)')



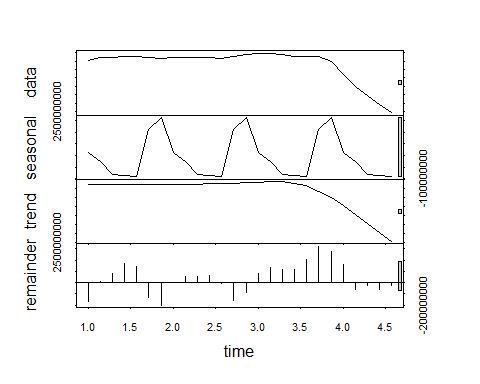
De acuerdo al modelo anterior, a continuación se presenta la predicción para la variable Kilometráje en la tipología 1.

prediccion1 <- forecast(arima2, h=12)  
plot(prediccion1)



## Variable KmT

count\_ma2 = ts(na.omit(datos2$KmT7), frequency=7)  
decomp2 = stl(count\_ma2, s.window="periodic")  
deseasonal\_KmT <- seasadj(decomp2)  
plot(decomp)



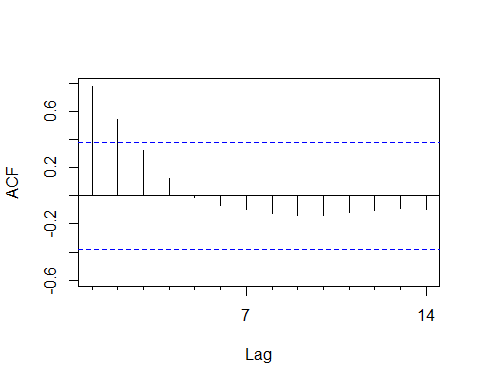
De acuerdo al gráfico se puede concluir que la serie no presenta estacionariedad, y una tendencia suave hacia la baja y pero si un pico descendente significativo al final del periodo.

La instalación de un modelo ARIMA requiere que la serie sea estacionaria e dice que una serie es estacionaria cuando su media, varianza y autocovarianza son invariantes en el tiempo, para confirmar la estacionalidad de la serie, corremos la prueba de dickey - Fuller aumentada:

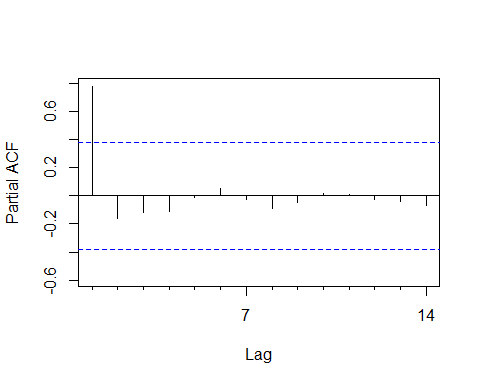
adf.test(count\_ma2, alternative = "stationary")

##   
## Augmented Dickey-Fuller Test  
##   
## data: count\_ma2  
## Dickey-Fuller = -0.39246, Lag order = 2, p-value = 0.9796  
## alternative hypothesis: stationary

Acf(count\_ma, main='')

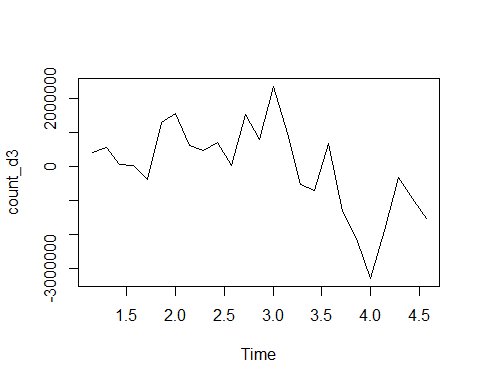


Pacf(count\_ma, main='')



Debido al que el p - valor es mayor al 5% de significancia, no se rechaza la hipótesis nula, lo que quiere decir que no es estacionaria. Como la serie no es estacionaria es necesario diferenciar, al diferenciar el modelo que se ajusta es un modelo ARIMA.

count\_d3 = diff(deseasonal\_KmT, differences = 1)  
plot(count\_d3)

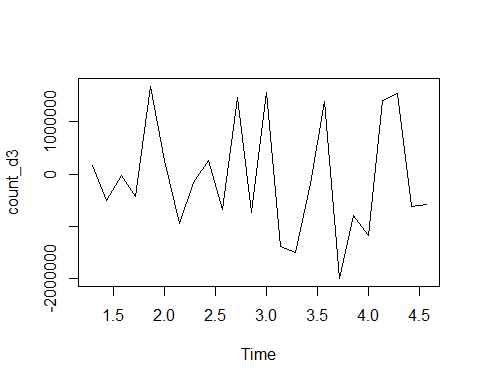


adf.test(count\_d3, alternative = "stationary")

##   
## Augmented Dickey-Fuller Test  
##   
## data: count\_d3  
## Dickey-Fuller = -2.3553, Lag order = 2, p-value = 0.437  
## alternative hypothesis: stationary

Con la primera diferenciación la serie aun sigue sin tener estacionariedad, se vuelve a realizar una segunda diferenciación:

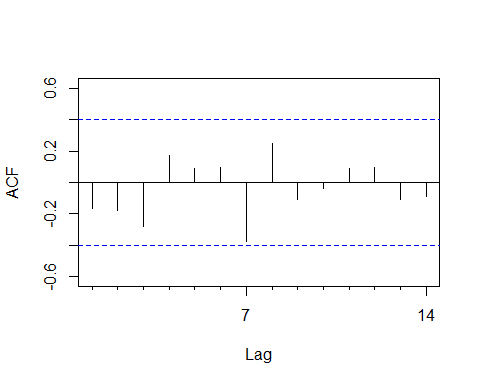
count\_d3 = diff(deseasonal\_KmT, differences = 2)  
plot(count\_d3)



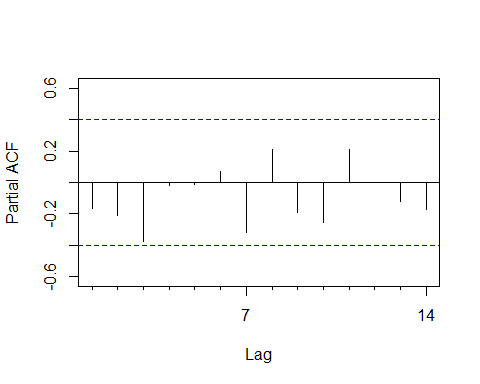
adf.test(count\_d3, alternative = "stationary")

##   
## Augmented Dickey-Fuller Test  
##   
## data: count\_d3  
## Dickey-Fuller = -4.2524, Lag order = 2, p-value = 0.01445  
## alternative hypothesis: stationary

Acf(count\_d3, main='')



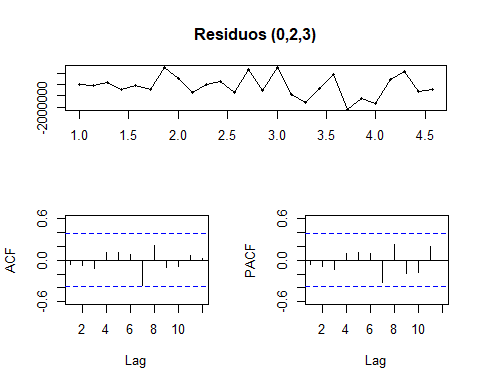
Pacf(count\_d3, main='')

 Se observa que a la segunda diferenciación se rechaza la hipótesis nula confirmando que la serie es estacionaria, de esta manera, según la función de autocorrelación y la autocorrelación parcial, el orden p del modelo es 0 y el orden q es 3.

arima3<- arima(deseasonal\_KmT,order=c(0,2,3), method = "ML")  
arima3$aic

## [1] 738.6955

tsdisplay(residuals(arima3), lag.max=12, main='Residuos (0,2,3)')



De acuerdo al modelo anterior, a continuación se presenta la predicción para la variable Kilometráje en la tipología 2.

prediccion <- forecast(arima3, h=12)  
plot(prediccion)

