Estudio de demanda de bicicletas

Juan Pablo Delzo

1 de marzo de 2018

Una empresa de alquiler de bicicletas en una area de Estados Unidos tiene registrado desde 2011 a 2012 la demanda de usuarios por hora de acuerdo del tipo de cliente (casual o registered), condiciones climaticas, día festivo, estación del año. Por lo que me ha despertado el interes de saber que información ofrecen estos datos.

Este artículo tiene como objetivo en presentar un ejemplo técnico en el análisis inicial en la demanda de bicicletas. Para ello, se han realizados los siguientes estudios:

- Visualización temporal
- Análisis anual
- Análisis promedio diario
- Análisis temporal y predicción de la demanda

En donde se esta mostrando **paso a paso** la programación ejecutada en ***R*** con el fin de difundir un caso en "*el arte de programar en R*".

Descarga de datos:

```
setwd("~/Analisis de Datos/Practicas en R/Bicing sharing")
datos <- read.csv("hour.csv", header=TRUE, sep=",", quote= "\"'" , string
sAsFactors= FALSE, dec=".")
```

Edición de los datos descargados:

Para el análisis inicial, con el objetivo de visualizar los datos, la información metereológica no brinda una directa correlación, por lo que estos datos más bien deben de ser considerado en el análisis dentro del Machine Learning. Por lo que la información metereológica estará excluida en los presentes análisis.

```
datos <- within(datos, rm(instant, yr,mnth,holiday,weekday,weathersit,tem
p,hum,windspeed,atemp,cnt))
colnames(datos)[1] <- c("date")
datos$hr <- as.numeric(datos$hr)
datos$casual <- as.numeric(datos$casual)
datos$season <- as.numeric(datos$season)
datos$workingday <- as.numeric(datos$workingday)
datos$registered <- as.numeric(datos$registered)
```

Descarga de los paquetes auxiliares:

```
library(plyr)
```

```
## Warning: package 'plyr' was built under R version 3.3.1
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.3.1
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

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

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

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.3.3
```

```
library(extrafont)
```

```
## Warning: package 'extrafont' was built under R version 3.3.3
```

```
## Registering fonts with R
```

```
library(TTR)
```

```
## Warning: package 'TTR' was built under R version 3.3.1
```

```
library(forecast)
```

```
## Warning: package 'forecast' was built under R version 3.3.1
```

```
## Loading required package: zoo
```

```
## Warning: package 'zoo' was built under R version 3.3.1
```

```
##
## Attaching package: 'zoo'
```

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

```
## Loading required package: timeDate

## This is forecast 7.2
```

Visualización de los datos de entrada:

```
head(datos)

##         date season hr workingday casual registered
## 1 2011-01-01      1  0          0      3         13
## 2 2011-01-01      1  1          0      8         32
## 3 2011-01-01      1  2          0      5         27
## 4 2011-01-01      1  3          0      3         10
## 5 2011-01-01      1  4          0      0          1
## 6 2011-01-01      1  5          0      0          1
```

Cuadro 1: Primeros 6 valores de entrada de los datos editados en el estudio de demanda de bicicletas

1.Visualización temporal:

En primer lugar, se está realizando una comparación en la evolucion de demanda horaria de las usuarios casual, registered y la demanda total.

Codificación inicial:

```
datetime <- paste(datos$date,paste(if_else(datos$hr<10,paste("0",datos$hr
, sep=""),as.character(datos$hr)),":00:00",sep="") )
datos <- cbind(datetime,datos)
datos$datetime <- strptime (datos$datetime, "%Y-%m-%d %H:%M:%S")
```

Ploteo:

```
p <- ggplot(datos, aes(datetime))
p <- p + geom_line(aes(y= casual + registered,colour= "total"))
p <- p + geom_line(aes(y= registered, colour= "registered"))
p <- p + geom_line(aes(y= casual, colour= "casual"))
p <- p + scale_colour_manual("", values= c("casual"= "deepskyblue3","regi
stered"= " darkorange","total"= "firebrick4"))
p <- p + ggtitle("Demanda horaria")
p <- p + scale_x_datetime(date_labels = "%b %Y")
p <- p + theme_bw() + ylab("Users")
p <- p + theme( axis.line = element_line(size=1, colour= "black"),plot.ti
tle = element_text(hjust = 0.5), panel.border = element_blank())
p <- p + theme(plot.title = element_text(family= "Comic Sans MS"), text=
element_text(family = "Comic Sans MS"))
p <- p + theme(axis.text.x = element_text(colour="black",size= 10), axis.
text.y = element_text(colour="black",size=10), legend.key = element_rect(
fill="white",colour="white"))
p <- p + theme(legend.position = "bottom",legend.direction = "horizontal"
,legend.text = element_text(colour = "black", size= 12))
p <- p + theme(panel.grid.major = element_line(colour= "gray80"), panel.g
```

```
rid.minor = element_line(colour= "gray80"))
p

## Warning: Removed 2 rows containing missing values (geom_path).

## Warning: Removed 2 rows containing missing values (geom_path).

## Warning: Removed 2 rows containing missing values (geom_path).
```

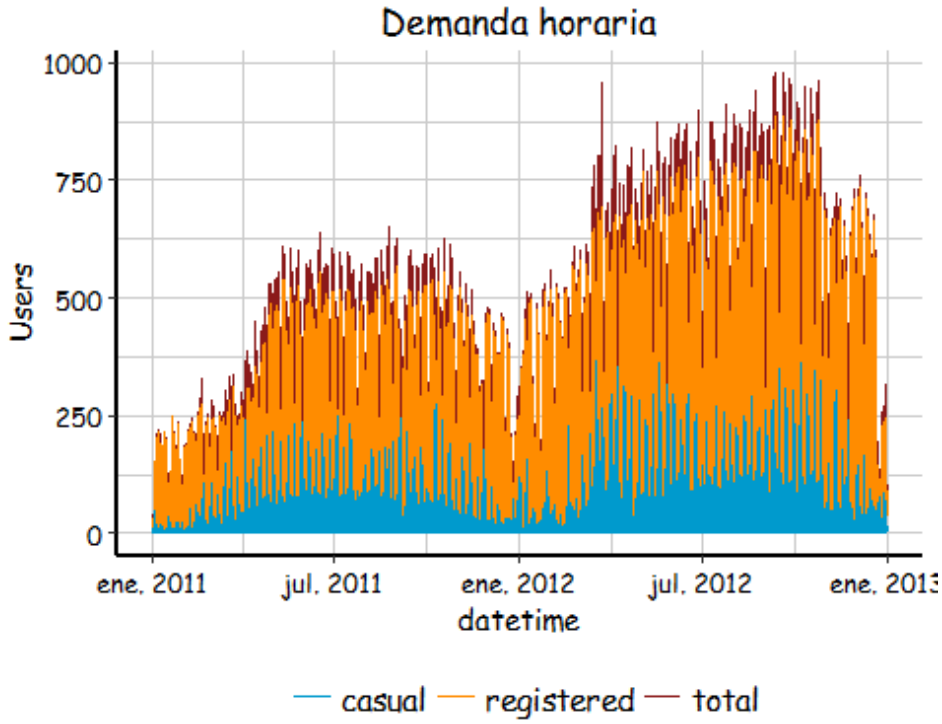

Fig 1: Demanda horaria de las bicicletas entre los años 2011 y 2012

- Se observa una mayor demanda en el año 2012
- Existe una demanda pico entre los meses de marzo y octubre en cada año

Para profundizar la primera impresión de los datos, se necesitará cuantificar la demanda por año.

2.Análisis anual:

Codificación previa:

```
datos$año <- format(datos$datetime,"%Y")
Tusers2011 <- colSums(filter(datos[,-(1:2)], año == 2011)[,4:5])
Tusers2012 <- colSums(filter(datos[,-(1:2)], año == 2012)[,4:5])
Tanual <- data.frame(año = c(rep(2011,2),rep(2012,2) ),Count= c(Tusers201
1,Tusers2012), User= c(rep(c("casual","registered"),2)) )
Panual <- data.frame(año = Tanual$año, label= Tanual$Count)
```

```
Panual$label <- c(sort(Panual$label[1:2],decreasing = TRUE),sort(Panual$l
abel[3:4],decreasing = TRUE))
Panual <- ddply(Panual,.(año),transform,pos= cumsum(label)- 0.5*label)
Panual$label <- round(Panual$label*100/c(rep(sum(Tanual$Count[1:2]),2), r
ep(sum(Tanual$Count[3:4]),2)),0)
Panual$label <- paste(Panual$label,"%",sep="")
totales <- data.frame(año= c(2011,2012), label= c(sum(Tusers2011), sum(Tu
sers2012)), pos= c(sum(Tusers2011),sum(Tusers2012))+100000)
Panual <- rbind(Panual,totales)
```

Ploteo:

```
colores <- c("deepskyblue3", "darkorange")
bar <- ggplot() + geom_bar(data=  Tanual, aes(x= año, y= Count, fill= Use
r),stat= "Identity" )
bar <- bar + geom_text(data= Panual, aes(x= año,y= pos, label= label),siz
e= 5,family= "Comic Sans MS")
bar <- bar + theme_bw() + scale_x_continuous(breaks = 2011:2012) + ggtitl
e("Usuarios")
bar <- bar + scale_fill_manual(values= colores)
bar <- bar + theme(legend.position = "bottom",legend.direction = "horizon
tal",legend.title = element_blank())
bar <- bar + theme(plot.title = element_text(hjust = 0.5), axis.line = el
ement_line(size= 1, colour= "black"))
bar <- bar + theme(panel.border = element_blank(), panel.grid.minor = ele
ment_blank())
bar <- bar + theme(plot.title = element_text(family = "Comic Sans MS"), t
ext = element_text(family= "Comic Sans MS"))
bar <- bar + theme(axis.text.x = element_text(colour="black",size=10),axi
s.text.y = element_text(colour="black", size=10))
bar <- bar + theme(legend.text = element_text(colour= "black",size=13))
bar
```

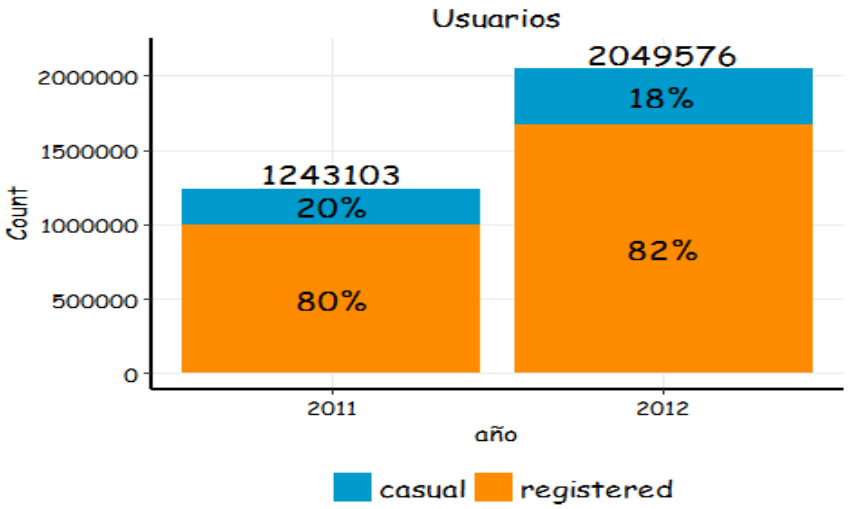

Fig 2: Comparación de la demanda anual de bicicletas

- La proporción de la demanda de usuarios casual y registered es de 1 a 4
- Existe un incremento del 50% de la demanda en el 2012

3.Demanda promedio diario:

El objetivo es visualizar las diferencias de la demanda por estación y de acuerdo al tipo de dia, es decir si es laborable o festivo. Y como se ha observado en que la demanda ha variado de un año en un 50%; por tanto para que no genere errores siginficativos, se está realizando una demanda promedio por hora a lo largo del dia respecto al ultimo año.

Codificación inicial:

```
datos2012 <- filter(datos[,-1],año == "2012")
datos2012 <- within(datos2012, rm(date,año))
i <- 1
for(S in 1:4){ season <- filter(datos2012, season == S)
               for (H in 0:23) { hora <- filter(season, hr == H)
               for(wk in 0:1){workingday <- filter(hora, workingday == wk)
              if ( i ==1){ datosprom2012 <-  round(colMeans(workingday),0)
               i <- 2
                           }
 else { datosprom2012 <- rbind(datosprom2012, round(colMeans(workingday),
0))}
                               }
                             }
               }
datosprom2012 <- as.data.frame(datosprom2012)
rownames(datosprom2012) <- NULL
datosprom2012$season <- as.factor(datosprom2012$season)
```

Ploteo:

```
datosprom2012$workingday <- ifelse(datosprom2012$workingday == 0, "h","w"
)
names <- list( '1'="Springer",'2'="Summer",'3'="Fall",'4'="Winter",'w'="w
orkingday",'h'="holiday")
labeller <- function(variable,value){ return(names[value])}
p <- ggplot(datosprom2012, aes(hr)) + geom_line (aes(y= casual, colour="c
asual")) + geom_line(aes(y= registered,colour="registered"))

p <- p + geom_line(aes(y= casual + registered, colour= "total"))
p <- p + scale_colour_manual("",values= c("casual"= "deepskyblue3","regis
tered"= "darkorange","total"= "firebrick4"))
p <- p + facet_grid(workingday ~ season, labeller = labeller, scale= "fre
e_y" ) + theme_bw()

## Warning: The labeller API has been updated. Labellers taking `variable
`and
## `value` arguments are now deprecated. See labellers documentation.
```

```
p <- p + theme(strip.background = element_rect(colour = "paleturquoise",
fill = "paleturquoise"))
p <- p + theme(legend.position = "bottom",legend.direction = "horizontal"
,legend.title = element_blank())
p <- p + theme(plot.title = element_text(hjust = 0.5), axis.line = elemen
t_line(size= 1, colour= "black"))
p <- p + theme(panel.grid.minor = element_blank())
p <- p + theme(plot.title = element_text(family = "Comic Sans MS"), text
= element_text(family= "Comic Sans MS"))
p <- p + theme(axis.text.x = element_text(colour="black",size=10),axis.te
xt.y = element_text(colour="black", size=10))
p <- p + theme(legend.text = element_text(colour= "black",size=13))
p <- p + ggtitle("Demanda horaria promedio 2012")
p <- p + ylab("Count") + xlab("hours")
p
```

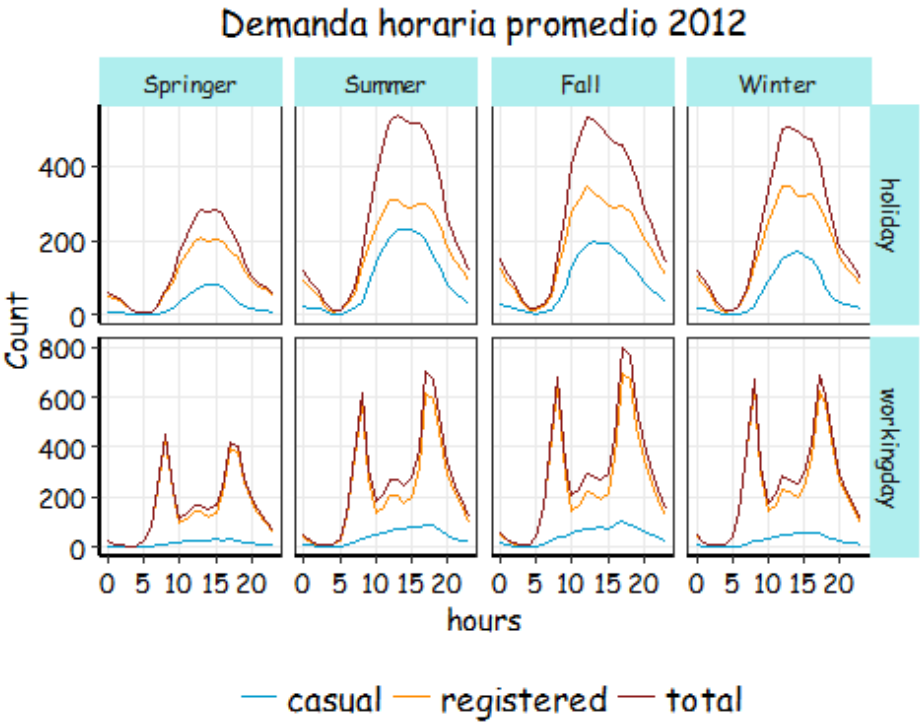

Figura 3: Comparación de la demanda promedio por hora de acuerdo al dia y estación en 2012

- En los dias festivos, la gráficas de la demanda de bicicletas de usuarios tanto casual como registered se asemejan a la campana de Gauss, además ambas funciones tienen similares horas punta al mediodia.

- En los dias laborables, las demanda entre ambos tipos de usuarios varian drásticamente, la demanda de usuarios casuales tienen una menor preponderancia que presenta una función creciente y decreciente con pendientes casi llanas donde el punto máximo viene a ser a las 15 horas. Mientras que la demanda de usuarios registered se asemeja a dos triangulos

obtusos próximos entre si, donde las dos puntos máximos se registrán a las 9 y 18 horas.

- Existe una menor demanda de bicicletas en general en la estación de Primavera, y especialmente en los dias festivos.

4.Análisis temporal y predicción de la demanda:

Por último se han realizado dos análisis:

- Análisis Temporal: Que viene a ser un estudio de descomposición aditiva del total de los datos temporales en funcion estacional, tendencia y aleatorio.
- Predicción de la demanda: De acuerdo al conjunto total de datos temporales, se ha estimado la demanda futura por mes para el siguiente año.

Ambos análisis se han procedido por separado la demanda de bicicletas por usuarios casual, registered y la demanda total.

Codificación previa:

```
datosmensual <- within(datos, rm(datetime,season,hr,workingday))
datosmensual$mes <- format(as.Date(datosmensual$date, "%Y-%m-%d"),"%m")
datosmensual$año <- as.numeric(datosmensual$año)
datosmensual$mes <- as.numeric(datosmensual$mes)
demandamensual <- matrix(c(rep(0,24*2)),24,2)
colnames(demandamensual) <- c("casual","registered")
i <- 1
for(y in 2011:2012){datoaño <- filter(datosmensual,año == y)
                    for(m in 1:12){datomes <- filter(datoaño, mes == m)
                        demandamensual[i,] <-  colSums(datomes[,2:3])
                                        i <- i +1
                                }
                    }
demandamensual <- as.data.frame(demandamensual)
demandamensual$total <- demandamensual$casual + demandamensual$registered
```

4.1 Casual Users:

```
casual <- ts(demandamensual$casual,frequency = 12,start= c(2011,1))
casualcomponents <- decompose(casual)
plot(casualcomponents)
```

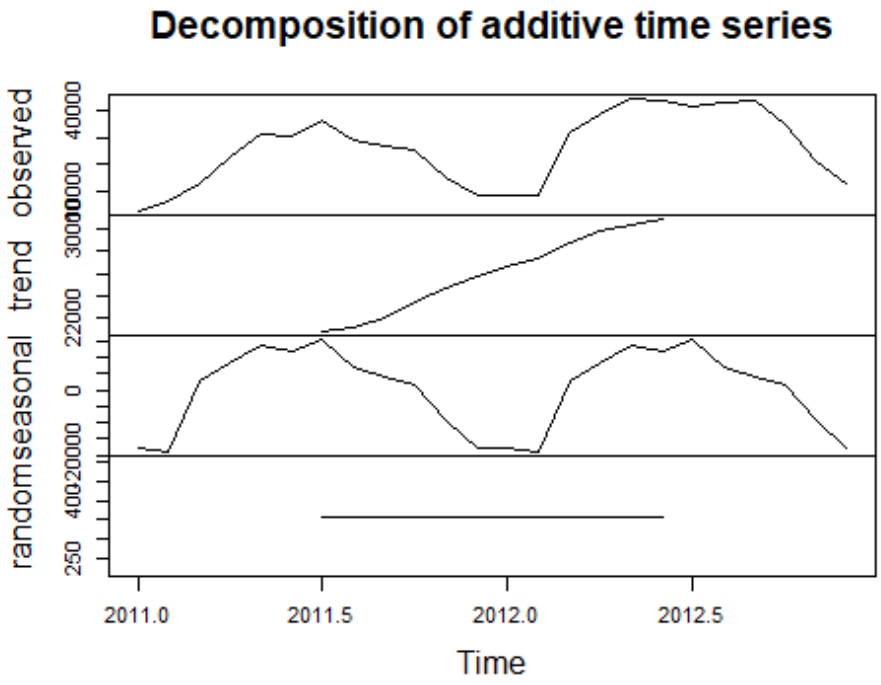

Figura 4: Descomposición temporal de la demanda de usuarios casual

- No existe la aleatoriedad en la demanda
- La tendencia es creciente

```
casualforescast <- HoltWinters(casual)
casualforescast2 <- forecast.HoltWinters(casualforescast, h=12)
plot(casualforescast2,main= "Forecast of casual demand from HoltWinters")
```

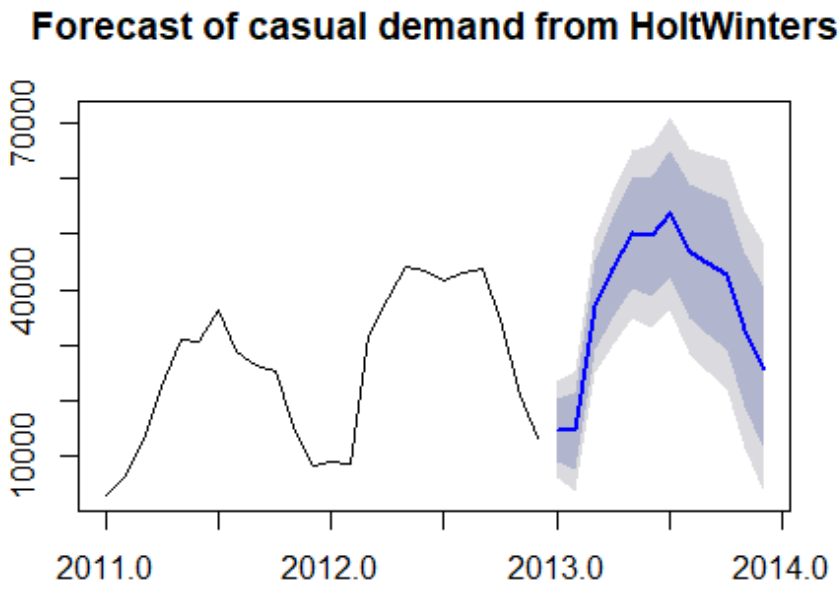

Figura 5: Previsión de la demanda de usuarios casual

- Se prevé una mayor demanda, donde la demanda pico seria a mitad de año

4.2Registered Users:

```
registered <- ts(demandamensual$registered,frequency = 12,start= c(2011,1
))
registeredcomponents <- decompose(registered)
plot(registeredcomponents)

## Warning in plot.window(...): amplitud relativa de valores = 60 * EPS,
es
## pequeño (eixo 2)
```

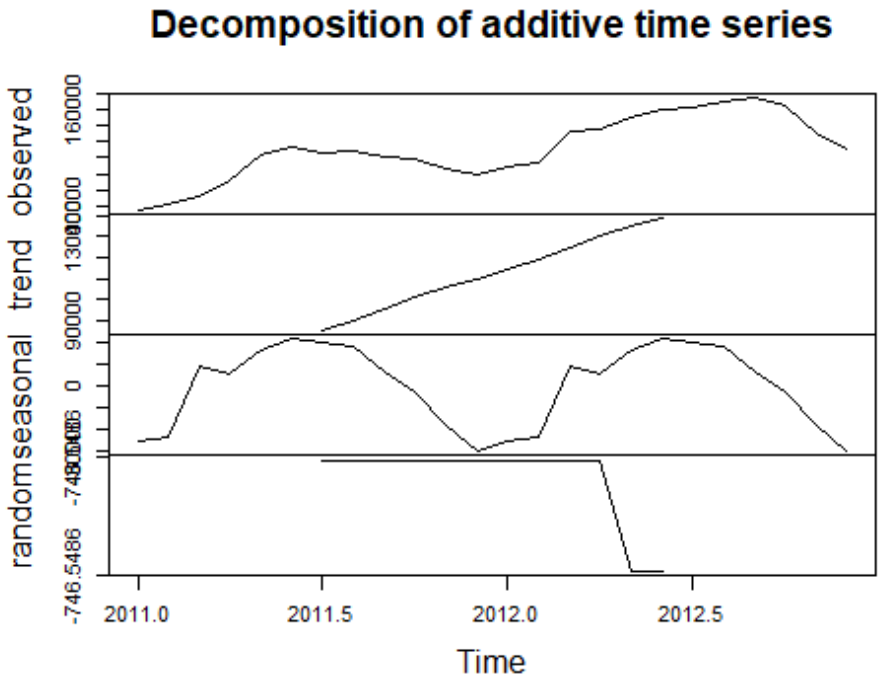

Figura 6: Descomposición Temporal de la demanda de usuarios registered

- Existe aleatoriedad en los usuarios registered
- La tendencia es creciente

```
registeredforescast <- HoltWinters(registered)
registeredforescast2 <- forecast.HoltWinters(registeredforescast, h=12)
plot(registeredforescast2,main= "Forecast of registered demand from HoltW
inters")
```

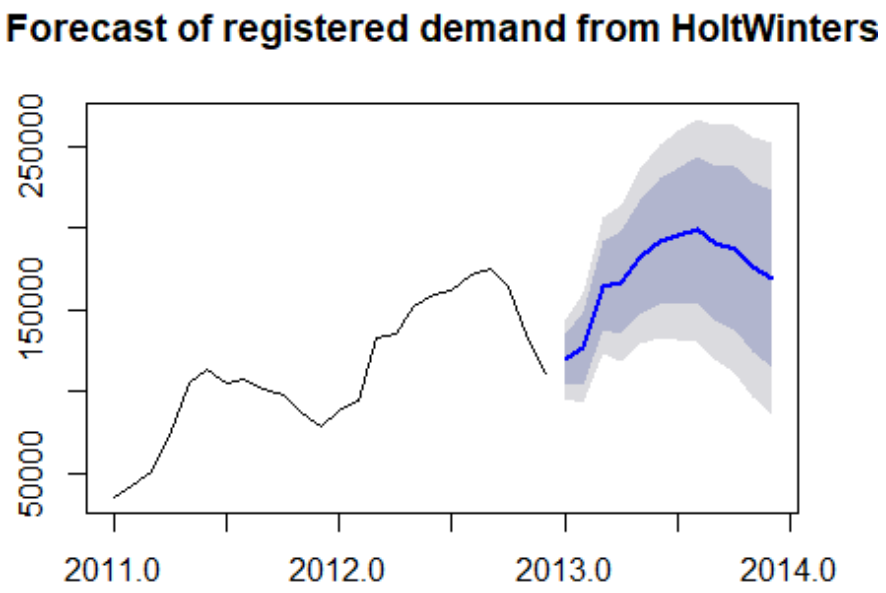

Figura 7: Predicción de la demanda de usuarios registered

- Similar al caso en la demanda de usuarios casual, se prevé una mayor demanda, donde la demanda pico seria a mitad de año

4.3Total Users:

```
total <- ts(demandamensual$total,frequency = 12,start= c(2011,1))
totalcomponents <- decompose(total)
plot(totalcomponents)
```

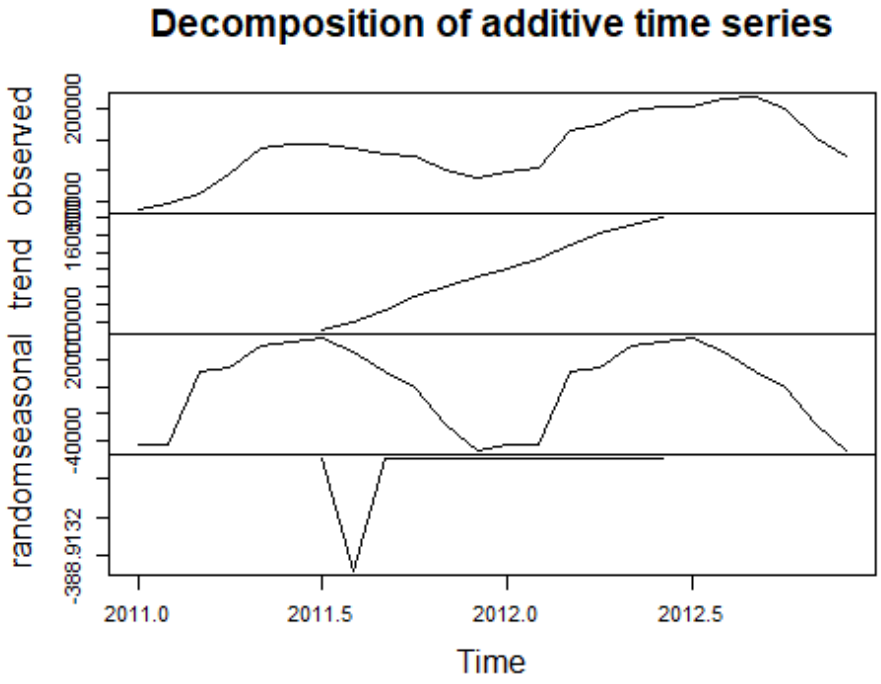

Figura 8: Descomposición Temporal de la demanda total

- Presenta identicas características respecto a la demanda de usuarios registered

```
totalforescast <- HoltWinters(total)
totalforescast2 <- forecast.HoltWinters(totalforescast, h=12)
plot(totalforescast2,main= "Forecast of total demand from HoltWinters")
```

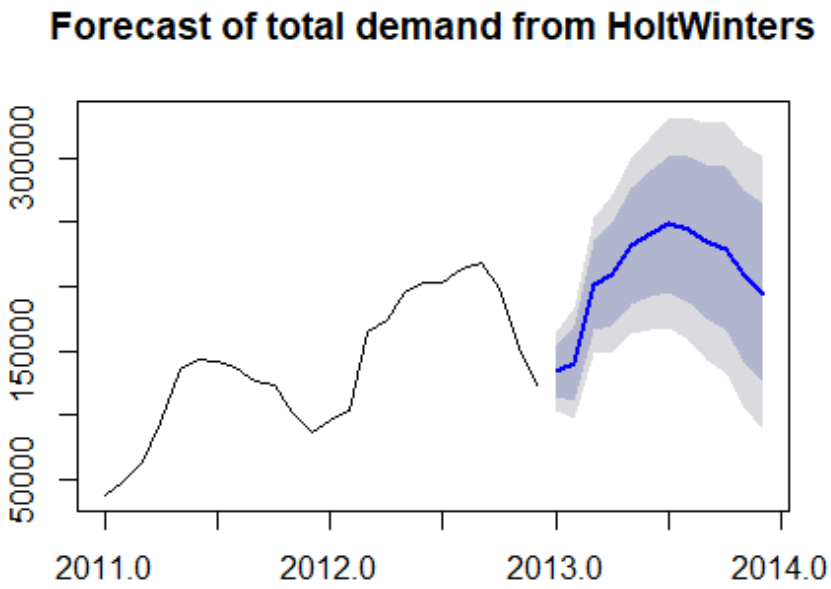

Figura9: Prevision mensual en la demanda total

- Presentan identicas características respecto a la demanda de usuarios registered