During the last years We have become a huge fan of tennis (both, on court and in TV). This year I followed (in TV) with attention Roland Garros and Wimbledon.

Although clay and grass are very different surfaces I’ve found the game very similar. What happened with the serve-volley game? We think it is my duty (as data freak) to found some data and play around with it.

So, I’ve done some code to find out.

library(RCurl)

matches\_file\_m <- "<https://raw.githubusercontent.com/JeffSackmann/tennis_MatchChartingProject/master/charting-m-matches.csv>"

matches\_file\_w <- "<https://raw.githubusercontent.com/JeffSackmann/tennis_MatchChartingProject/master/charting-w-matches.csv>"

points\_file\_m <- "<https://raw.githubusercontent.com/JeffSackmann/tennis_MatchChartingProject/master/charting-m-points.csv>"

points\_file\_w <- "<https://raw.githubusercontent.com/JeffSackmann/tennis_MatchChartingProject/master/charting-w-points.csv>"

download\_matches <- function(){

if(!exists("df\_matches")){

matches\_m <- read.csv(file = matches\_file\_m,header = TRUE,sep = ",",quote = "")

matches\_w <- read.csv(file = matches\_file\_w,header = TRUE,sep = ",",quote = "")

df\_matches <- bind\_rows(matches\_w,matches\_m)

}

return(df\_matches)

}

download\_points <- function(){

if(!exists("df\_points")){

points\_m <- read.csv(file = points\_file\_m,header = TRUE,sep = ",",quote = "")

points\_w <- read.csv(file = points\_file\_w,header = FALSE,sep = ",",quote = "")

colnames(points\_w) <- colnames(points\_m)

points\_m$TB. <- as.integer(points\_m$TB.)

df\_points <- bind\_rows(points\_w,points\_m)

}

return(df\_points)

}

get\_players <- function(){

df\_matches <- download\_matches()

players <- unique(c(df\_matches$Player.1,df\_matches$Player.2))

return(players)

}

get\_tournaments <- function(){

df\_matches <- download\_matches()

tournaments <- unique(df\_matches$Tournament)

return(tournaments)

}

cols <- c("RolandGarros" = "#D35221", "Wimbledon" = "#1B7649")

df\_points <- download\_points()

## Warning in download\_points(): NAs introduced by coercion

df\_matches <- download\_matches()

id\_matches <- df\_matches %>% dplyr::filter(Tournament %in% c("Roland Garros","Wimbledon")) %>% dplyr::select(match\_id,Tournament) %>% dplyr::group\_split(Tournament)

l\_point <- vector("list",2)

for (i in 1:2){

l\_point[[i]] <- df\_points %>% dplyr::filter(match\_id %in% id\_matches[[i]]$match\_id)

}

l\_point[[1]]$Tournament <- "RolandGarros"

l\_point[[2]]$Tournament <- "Wimbledon"

l\_point <- bind\_rows(l\_point)

#porcentaje de Aces

resultado1 <- l\_point %>% dplyr::group\_by(Tournament) %>% dplyr::summarise(PorcentajeAces=100\*sum(isAce)/n())

resultado1

## # A tibble: 2 x 2

## Tournament PorcentajeAces

##

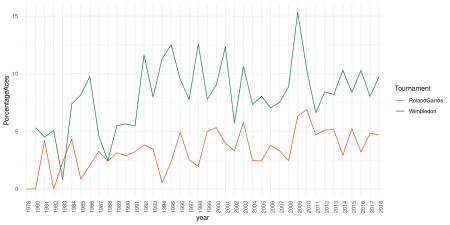
## 1 RolandGarros 4.03

## 2 Wimbledon 8.77

resultado2 <- l\_point %>% dplyr::group\_by(Tournament,year=substring(l\_point$match\_id,1,4)) %>% dplyr::summarise(PorcentageAces=100\*sum(isAce)/n())

g1 <- ggplot(resultado2) + geom\_line(aes(x=year,y=PorcentageAces,color=Tournament,group=Tournament)) + theme\_minimal() + scale\_colour\_manual(values = cols) + theme(axis.text.x =element\_text(angle = 90, hjust = 1))

plot(g1)



Wimbledon still observes higher percentage of aces, but in both tournaments percentaje of aces has been growing.

#longitud del punto

resultado3 <- l\_point %>% dplyr::group\_by(Tournament) %>% dplyr::summarise(AverageLengthRally=mean(rallyLen))

resultado3

## # A tibble: 2 x 2

## Tournament AverageLengthRally

##

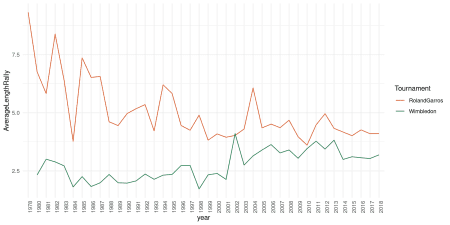
## 1 RolandGarros 4.49

## 2 Wimbledon 2.89

resultado4 <- l\_point %>% dplyr::group\_by(Tournament,year=substring(l\_point$match\_id,1,4)) %>% dplyr::summarise(AverageLengthRally=mean(rallyLen))

g2 <- ggplot(resultado4) + geom\_line(aes(x=year,y=AverageLengthRally,color=Tournament,group=Tournament)) + theme\_minimal() + scale\_colour\_manual(values = cols) + theme(axis.text.x =element\_text(angle = 90, hjust = 1))

plot(g2)



Here you have, Wimbledon still has shorter rallies but the difference is getting smaller. It has surprised me the length of the points in Roland Garros in the 70s and 80s.

#Errores no forzados

resultado5 <- l\_point %>% dplyr::group\_by(Tournament) %>% dplyr::summarise(UnforcedErrors=100\*sum(isUnforced)/n())

resultado5

## # A tibble: 2 x 2

## Tournament UnforcedErrors

##

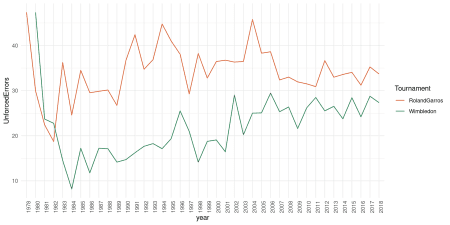
## 1 RolandGarros 33.9

## 2 Wimbledon 23.6

resultado6 <- l\_point %>% dplyr::group\_by(Tournament,year=substring(l\_point$match\_id,1,4)) %>% dplyr::summarise(UnforcedErrors=100\*sum(isUnforced)/n())

g3 <- ggplot(resultado6) + geom\_line(aes(x=year,y=UnforcedErrors,color=Tournament,group=Tournament)) + theme\_minimal() + scale\_colour\_manual(values = cols) + theme(axis.text.x =element\_text(angle = 90, hjust = 1))

plot(g3)



Unforced errors in Wimbledon have been increasing, I guess this is related with the length of the rallies

#Puntos que acaban en volea

resultado7 <- l\_point %>% dplyr::mutate(lastshot=str\_sub(string = rallyNoDirection,start = -1)) %>% dplyr::mutate(lastisvolley=ifelse(lastshot %in% c("v","z","o","p"),TRUE,FALSE )) %>% dplyr::group\_by(Tournament) %>% dplyr::summarise(LastVolley=100\*sum(lastisvolley)/n())

resultado7

## # A tibble: 2 x 2

## Tournament LastVolley

##

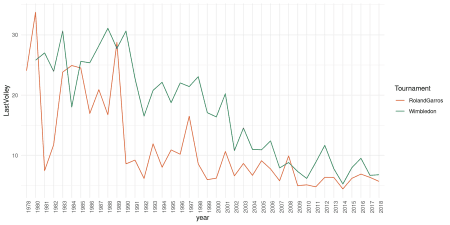
## 1 RolandGarros 8.37

## 2 Wimbledon 13.3

resultado8 <- l\_point %>% dplyr::mutate(lastshot=str\_sub(string = rallyNoDirection,start = -1)) %>% dplyr::mutate(lastisvolley=ifelse(lastshot %in% c("v","z","o","p"),TRUE,FALSE )) %>% dplyr::group\_by(Tournament,year=substring(l\_point$match\_id,1,4)) %>% dplyr::summarise(LastVolley=100\*sum(lastisvolley)/n())

g4 <- ggplot(resultado8) + geom\_line(aes(x=year,y=LastVolley,color=Tournament,group=Tournament)) + theme\_minimal() + scale\_colour\_manual(values = cols) + theme(axis.text.x =element\_text(angle = 90, hjust = 1))

plot(g4)



This is striking, points finishing in volley have been decreasing abruptly and nowadays are difficult to see.

It seems that serve-volley game is outdated even in its sanctuary and baseline tennis reigns in Wimbledon