The European tennis season is in full swing, with Roland Garros starting today and Wimbledon taking place in a few weeks. For a sports buff like me, it is the essence of summer (together with the Tour de France). Time to dive into some tennis data. As a follower of both the men’s and the women’s tour it occurred to me that in the latter the tournaments are less predictable. My gut feel was that in the men’s matches the favourite wins the match more frequently than at women’s matches. Of course gut feels are what the world makes go round, unless you are a data scientist. So lets analyse all the matches that were played at the four Grand Slam tournamants for the past fourty years.

**The Seeding System**

But first we need to think of a way to measure if the favourite indeed won. Who the favourite is to win a match might be dependent on who you ask, how do we determine who is the favourite to win? Luckily, there is an element of objectivity in each tournamant that is played. The top players of that moment according to the general rankings are *seeded*. The world number one is seeded first, the number two as second, up until the world number 32. The seeds are placed in the schedule in such a way that the strongest players only meet in the final rounds of the tournamant (if they beat the weaker ones first of course), The numbers one and two cannot meet before the final. They cannot meet the numbers three and four before semi-finals, and will not play against the numbers five up until eight before the quarterfinals. The higher your seed, the more weaker players you’ll meet during the tournamants before you meet a stronger player. From your seed you are expected to win all the rounds until you meet a stronger player, we could say a seeded player is ought to loose in a certain round (except the number one seed who is ought to win the tournamant).

**Building an upset score**

If seeded players loose in a round that is before the “ought-to-loose round” belonging their seed, we can say they underperformed. A way to measure this underperformance is to award an upset point for every round a player loses earlier than its “ought-to-loose round”. The number one seed is ought to win the tournament, if he or she loses the final it will results in one upset point, a loss in the semi-final yields two upset points, a loss in the quarter-final three etcetera. For the number two seed a loss in the semi-final yields one upset point, a loss in the quarter-final two, the round before three etcetera. In the table below all the possible upset points for the sixteen highest seeded players are displayed.

| **Seed** | **F** | **SF** | **QF** | **R16** | **R32** | **R64** | **R128** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 2 | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| 3-4 | 0 | 0 | 1 | 2 | 3 | 4 | 5 |
| 5-8 | 0 | 0 | 0 | 1 | 2 | 3 | 4 |
| 9-16 | 0 | 0 | 0 | 0 | 1 | 2 | 3 |

The total predictability of a tournament is simply the sum of all the individual upset points. Because until 2001 there were only sixteen seeds, instead of 32, we keep using the first sixteen seeds after it as well. Then the tournamant upset scores can range from 0 (every seeded player lost in its “ought-to-loose round” and the first seeded won the whole thing) to 63 (every seeded player lost in the first round).

**The data**

Jeff Sackman maintains a database of all the tennis matches played on both the atp (men’s) and wta (women’s) tour. It is on github so I simply cloned the git repositories on which they are maintained. Using the following code.

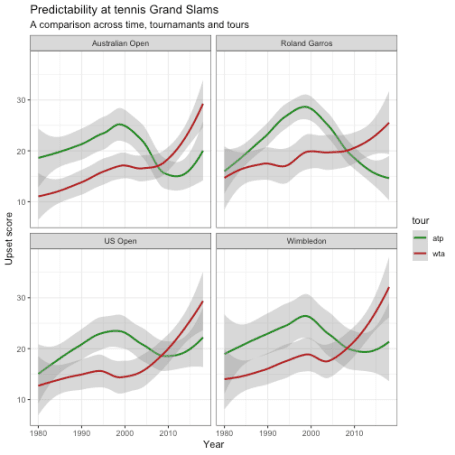
git clone https://github.com/JeffSackmann/tennis\_atp

git clone https://github.com/JeffSackmann/tennis\_wta

Resulting in two folders with a csv file for every year. Many thanks to Jeff for making our lives so easy here.

**The results**

In the graph below the trends are shown for the four Grand Slam tournaments, since 1980, both men (atp) and women (wta). Rather than looking at individual events, I am interested in overall trends. That is why a smoothed curve is used, rather than a line or dot plot.



Indeed the men’s events are more predictable than the women’s, but too my great surprise this is only the case for the last ten years. In the thirty years before that the top women players consistently outperformed the top men players at the big tournaments. However, with players like Federer, Nadal and Djokovic starting to dominate on every surface, the probability of seeing an upset rapidly decreased at the men’s events the last years. Whereas, the women’s tour rapidly grew in unpredictability over the last decade. Many things to explore here. What happened in the late nineties on the men’s tour? Why do we see so many upsets at the women’s? Can we forecast the number of upsets we will see the upcoming tournaments? I leave those questions for someone else to answer, or maybe future-me. Here is the full code of this analysis (if you want to reproduce set the data\_path to the folder where you stored the data first).

library(tidyverse)

gs\_data\_one\_year <- function(year, tour = c("atp", "wta")) {

read\_csv(str\_glue("{data\_path}tennis\_{tour}/{tour}\_matches\_{year}.csv"),

col\_types = cols(loser\_seed = col\_integer())) %>%

filter(tourney\_level == "G") %>%

mutate(year = year, tour = tour) %>%

select(loser\_seed, round, tourney\_name, year, tour) %>%

mutate(tourney\_name = case\_when(

tourney\_name == "French Open" ~ "Roland Garros",

tourney\_name == "Us Open" ~ "US Open",

TRUE ~ tourney\_name

))

}

the\_data <- bind\_rows(

map\_dfr(1980:2018, ~gs\_data\_one\_year(.x, "atp")),

map\_dfr(1980:2018, ~gs\_data\_one\_year(.x, "wta"))

)

upset\_scores\_df <- function(round\_char,

upset\_points,

top\_only = 16) {

stopifnot(length(upset\_points) == 6)

tibble(

round = round\_char,

loser\_seed = 32:1,

upset\_points = rep(upset\_points, c(16, 8, 4, 2, 1, 1))

) %>%

filter(loser\_seed <= top\_only)

}

all\_rounds <- unique(the\_data$round)

all\_upset\_points <- list(2:7, 1:6, 0:5, c(rep(0, 2), 1:4), c(rep(0, 3), 1:3),

c(rep(0, 4), 1:2), c(rep(0, 5), 1))

upset\_df <- map2\_dfr(all\_rounds, all\_upset\_points, upset\_scores\_df,

top\_only = 16)

upset\_set <- left\_join(the\_data, upset\_df) %>%

group\_by(tourney\_name, year, tour) %>%

summarise(upset\_score = sum(upset\_points, na.rm = TRUE)) %>%

ungroup()

upset\_set %>%

ggplot(aes(year, upset\_score, col = tour)) +

geom\_smooth() +

facet\_wrap(~tourney\_name) +

labs(x = "Year",

y = "Upset score",

title = "Predictability at tennis Grand Slams",

subtitle = "A comparison across time, tournamants and tours") +

theme\_bw() +

scale\_color\_manual(values = c("forestgreen", "firebrick"))