Analysis of Tennis Matches

Matthew Weglicki

December 6, 2015

DATA ANALYSIS BY MATCHES

Using our clean data, we will now go on to investigate the dominance of certain countries within men's singles tennis by analyzing all available match statistics since 2010. We will be using the number of titles each country has won since 2010 as a proxy for "dominance". Tennis is also played on three different court surfaces, Clay, Grass and Hard, as well as three different tour "levels", ATP, Masters, and Grandslams, where each level offers a different number of points. Thus after we have collected data on how many titles each country has won, we will proceed to analyze how many of those titles were won on corresponding surfaces and levels, in order to answer the questions of whether some countries are particularly dominant on certain surfaces, or whether they shine on big occasions (ie at tournaments with more points).

Part I - Function Creation

First we must create the functions which when given the match statistics from certain years, are able to extract those matches which were "title matches", where the winner won the entire tournament, and output the winning nationalities, surface, and level of each tournament in that year.

```
setwd("/Users/williamcheung/Desktop/stats133/Tennis/")
setwd("./code")

library("readr")

# read in the files from clean_data
matches_10 <- read_csv("../clean_data/matches_10_clean.csv")
matches_11 <- read_csv("../clean_data/matches_11_clean.csv")
matches_12 <- read_csv("../clean_data/matches_12_clean.csv")
# exclude an entry from data which was repetitive and hence placed a bug in
# the analysis later on
matches_12 <- matches_12[-2656, ]
matches_13 <- read_csv("../clean_data/matches_13_clean.csv")
matches_14 <- read_csv("../clean_data/matches_14_clean.csv")
matches_15 <- read_csv("../clean_data/matches_15_clean.csv")

# function which produces unique number of tournaments for a given year
uniq_tour <- function(tour_year) {
    return(unique(tour_year$tourney_name))
}

# function when given a name or list of names of tournaments and a given
# year will output the total number of matches in the tournament(s), hence
# allows you to find the index of the final match of the tournament
tour_num_matches <- function(tour_names, tour_year) {
    num_matches <- numeric(0)
    for (i in tour_names) {
        num_matches <- c(num_matches, as.numeric(sum(tour_year$tourney_name ==
```

```r
            i)))
    }
    return(num_matches)
}

# function that when given tour_year, uniq_tour and total matches in each
# tournament, returns the ioc of the title winner
title_winner_ioc <- function(tour_year, tour_names, total_matches) {
    ioc_list <- character(0)
    for (i in 1:length(tour_names)) {
        ioc_list <- c(ioc_list, tour_year[tour_year$tourney_name == tour_names[i] &
            tour_year$match_num == total_matches[i], 6])
    }
    return(ioc_list)
}

# function that combines all of the above ie given a match year, it just
# displays a list of the countries that won the tournaments in a given year
title_winning_iocs <- function(tour_year) {
    tour_names <- uniq_tour(tour_year)
    total_matches <- tour_num_matches(tour_names, tour_year)
    iocs <- title_winner_ioc(tour_year, tour_names, total_matches)
    return(iocs)
}


# last two functions for surface and tourney level surface
title_winner_surface <- function(tour_year, tour_names, total_matches) {
    surface_list <- character(0)
    for (i in 1:length(tour_names)) {
        surface_list <- c(surface_list, tour_year[tour_year$tourney_name ==
            tour_names[i] & tour_year$match_num == total_matches[i], 2])
    }
    return(surface_list)
}

title_winning_surfaces <- function(tour_year) {
    tour_names <- uniq_tour(tour_year)
    total_matches <- tour_num_matches(tour_names, tour_year)
    surfaces <- title_winner_surface(tour_year, tour_names, total_matches)
    return(surfaces)
}

# tourney_level
title_winner_level <- function(tour_year, tour_names, total_matches) {
    level_list <- character(0)
    for (i in 1:length(tour_names)) {
        level_list <- c(level_list, tour_year[tour_year$tourney_name == tour_names[i] &
            tour_year$match_num == total_matches[i], 3])
    }
    return(level_list)
}
```

```
title_winning_levels <- function(tour_year) {
    tour_names <- uniq_tour(tour_year)
    total_matches <- tour_num_matches(tour_names, tour_year)
    levels <- title_winner_level(tour_year, tour_names, total_matches)
    return(levels)
}
```

Part II - Dataframe Creation

Now we move onto applying the functions from part I in order to create dataframes for each year containing information about the winning nationality, surface and level of each tournament in that year. We then consolidate all these dataframes into one comprehensive dataframe, from which we will be able to conduct our appropriate analysis.

```
# create a dataframe for each year with the information in the above description
df_2010 <- data.frame(
  tournament_name = uniq_tour(matches_10),
  winner_ioc = title_winning_iocs(matches_10),
  winning_surface = title_winning_surfaces(matches_10),
  winning_level = title_winning_levels(matches_10),
  year = 2010
)


df_2011 <- data.frame(
  tournament_name = uniq_tour(matches_11),
  winner_ioc = title_winning_iocs(matches_11),
  winning_surface = title_winning_surfaces(matches_11),
  winning_level = title_winning_levels(matches_11),
  year = 2011
)

df_2012 <- data.frame(
  tournament_name = uniq_tour(matches_12),
  winner_ioc = title_winning_iocs(matches_12),
  winning_surface = title_winning_surfaces(matches_12),
  winning_level = title_winning_levels(matches_12),
  year = 2012
)

df_2013 <- data.frame(
  tournament_name = uniq_tour(matches_13),
  winner_ioc = title_winning_iocs(matches_13),
  winning_surface = title_winning_surfaces(matches_13),
  winning_level = title_winning_levels(matches_13),
  year = 2013
)

df_2014 <- data.frame(
  tournament_name = uniq_tour(matches_14),
  winner_ioc = title_winning_iocs(matches_14),
  winning_surface = title_winning_surfaces(matches_14),
  winning_level = title_winning_levels(matches_14),
```

```r
  year = 2014
)

df_2015 <- data.frame(
  tournament_name = uniq_tour(matches_15),
  winner_ioc = title_winning_iocs(matches_15),
  winning_surface = title_winning_surfaces(matches_15),
  winning_level = title_winning_levels(matches_15),
  year = 2015
)

# create a master dataframe of all of the above data
master_df <- merge(df_2010, df_2011, all = TRUE)
master_df <- merge(master_df, df_2012, all = TRUE)
master_df <- merge(master_df, df_2013, all = TRUE)
master_df <- merge(master_df, df_2014, all = TRUE)
master_df <- merge(master_df, df_2015, all = TRUE)
```

Part III - Analysis of Master Dataframe

Now we are able to draw on our data to answer the questions of which are the most dominant countries in men's singles tennis since 2010, and do they have a particular affinity to a court type or to the larger events in the sport?

Total Number of Titles:

```r
# list of countries that have won titles since 2010
winner_ioc_list <- unique(master_df$winner_ioc)
# Check how many countries have won titles
length(winner_ioc_list)
```

```
## [1] 32
```

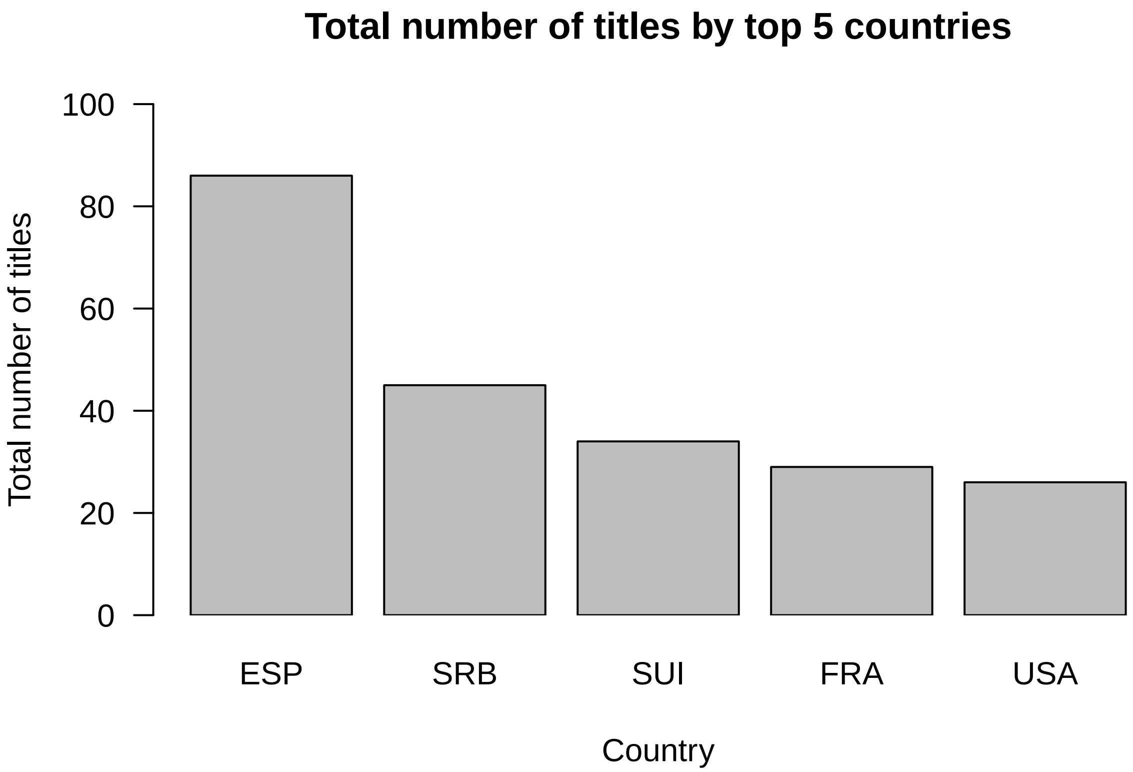
```r
# vector of total titles by country with names
titles_by_ioc <- numeric(0)
for (i in winner_ioc_list) {
    titles_by_ioc <- c(titles_by_ioc, sum(master_df$winner_ioc == i))
}
names(titles_by_ioc) <- winner_ioc_list[1:32]
titles_by_ioc
```

```
## ESP BUL USA CZE SRB SUI FRA GBR CAN JPN ARG SWE URU ITA AUS GER SVK CRO
##  86   4  26  11  45  34  29  22   7   9  24   7   3   6   6   8   3  15
## LAT RSA RUS AUT BRA KAZ POR BEL UKR CYP FIN NED UZB DOM
##   6   3   9   6   3   2   2   2   4   1   1   2   1   1
```

```r
# order the vector
ordered_titles_by_ioc <- sort(titles_by_ioc, decreasing = TRUE)
ordered_titles_by_ioc
```

```
## ESP SRB SUI FRA USA ARG GBR CRO CZE JPN RUS GER CAN SWE ITA AUS LAT AUT
##  86  45  34  29  26  24  22  15  11   9   9   8   7   7   6   6   6   6
## BUL UKR URU SVK RSA BRA KAZ POR BEL NED CYP FIN UZB DOM
##   4   4   3   3   3   3   2   2   2   2   1   1   1   1
```

```
# top 5 countries in terms of titles
top5 <- ordered_titles_by_ioc[1:5]
top5_names <- names(top5)

# construct barplot in terms of top 5 title winning countries
barplot(top5, xpd = F, axes = F, main = "Total number of titles by top 5 countries",
    xlab = "Country", ylim = c(0, 100), ylab = "Total number of titles")
axis(side = 2, at = seq(from = 0, to = 100, by = 20), tick = TRUE, las = 2)
```



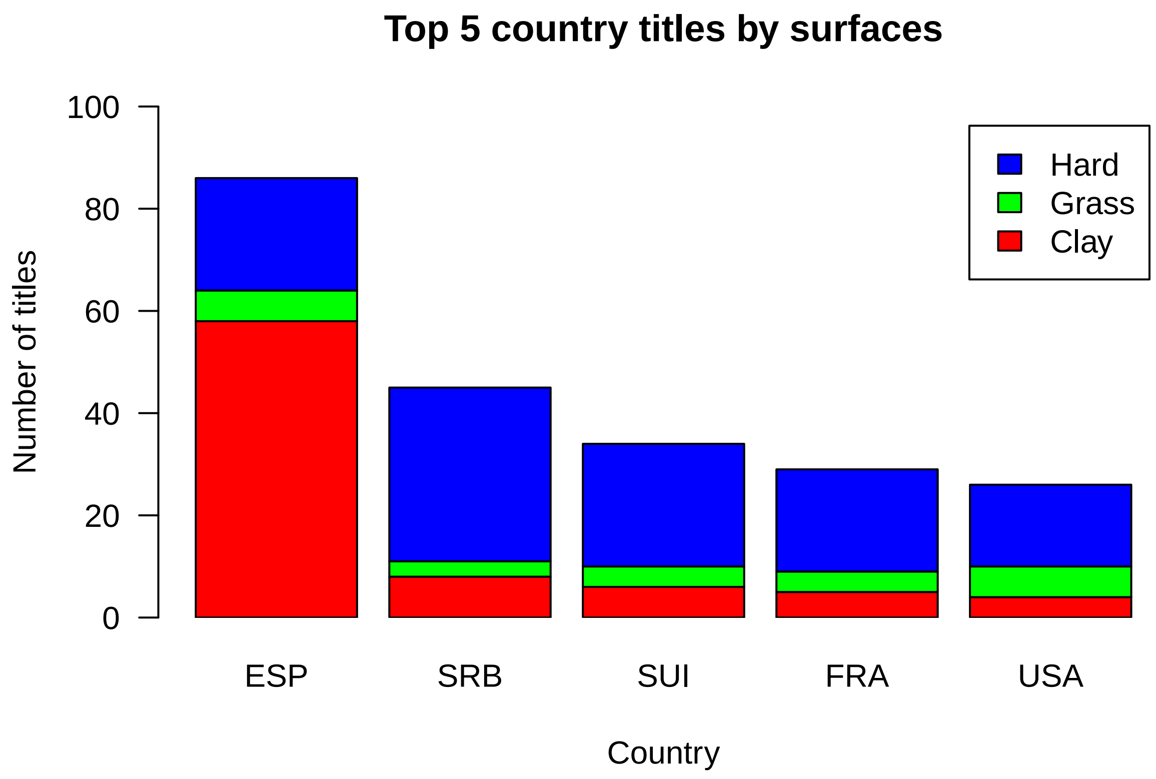
Titles by Surface:

```
# number of titles each country wins on each surface
clay_wins_list <- numeric(0)
for (i in winner_ioc_list) {
    clay_wins_list <- c(clay_wins_list, sum(master_df$winner_ioc == i & master_df$winning_surface ==
        "Clay"))
}

grass_wins_list <- numeric(0)
for (i in winner_ioc_list) {
    grass_wins_list <- c(grass_wins_list, sum(master_df$winner_ioc == i & master_df$winning_surface ==
        "Grass"))
}

hard_wins_list <- numeric(0)
```

```
for (i in winner_ioc_list) {
    hard_wins_list <- c(hard_wins_list, sum(master_df$winner_ioc == i & master_df$winning_surface ==
        "Hard"))
}


surface_df <- data.frame(country = winner_ioc_list, Clay = clay_wins_list, Grass = grass_wins_list,
    Hard = hard_wins_list)

# extracting indices required to obtain the top 5 countries from above for a
# comparison
row_indices_top5 <- c(which(surface_df$country == "ESP"), which(surface_df$country ==
    "SRB"), which(surface_df$country == "SUI"), which(surface_df$country ==
    "FRA"), which(surface_df$country == "USA"))

top5_surfaces <- surface_df[row_indices_top5, ]
top5_surfaces
```

```
##   country Clay Grass Hard
## 1     ESP   58     6   22
## 5     SRB    8     3   34
## 6     SUI    6     4   24
## 7     FRA    5     4   20
## 3     USA    4     6   16
```

```
# plot surfaces information for comparison
barplot(height = t(as.matrix(top5_surfaces[, 2:4])), names.arg = (top5_names),
    xpd = F, main = "Top 5 country titles by surfaces", xlab = "Country", ylab = "Number of titles",
    ylim = c(0, 100), axes = F, col = c(rgb(1, 0, 0), rgb(0, 1, 0), rgb(0, 0,
        1)), legend = colnames(top5_surfaces)[-1])
axis(side = 2, at = seq(from = 0, to = 100, by = 20), tick = TRUE, las = 2)
```



Titles by Level:

```
# number titles each country wins per level what levels are there?
unique(master_df$winning_level)
```

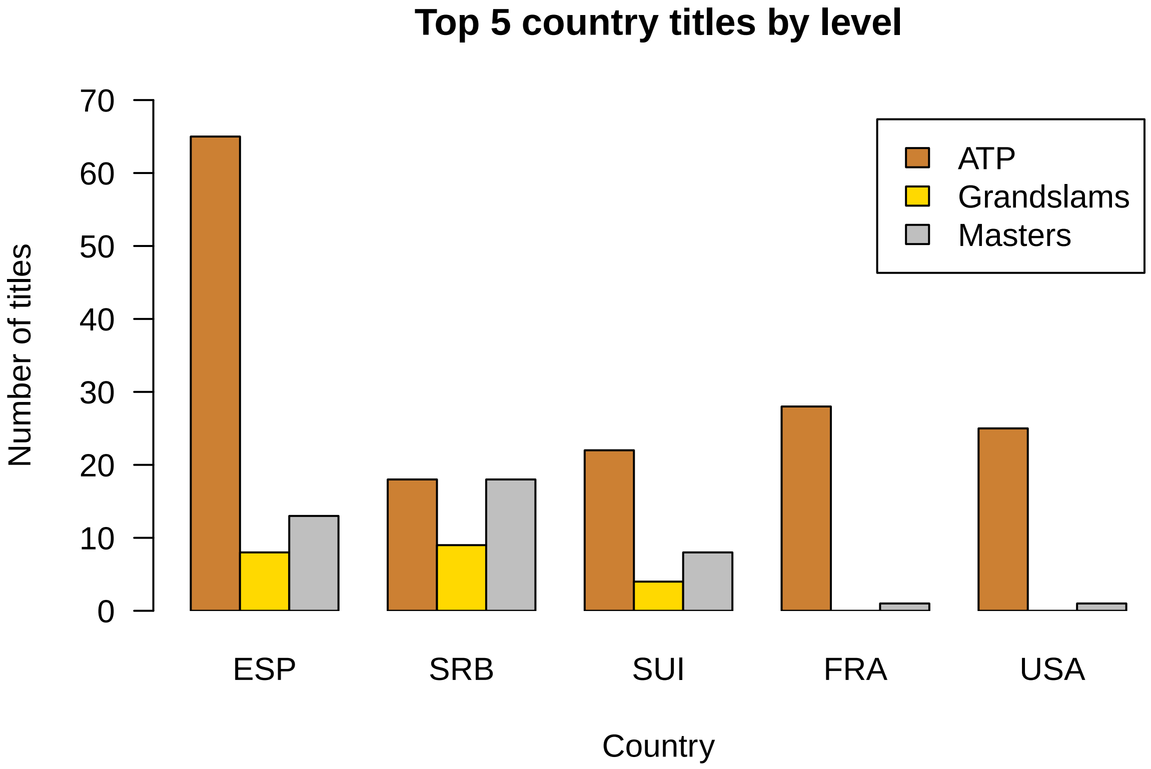
```
## [1] A G M
## Levels: A G M
```

```
# construct a dataframe of levels
A_wins_list <- numeric(0)
for (i in winner_ioc_list) {
    A_wins_list <- c(A_wins_list, sum(master_df$winner_ioc == i & master_df$winning_level ==
        "A"))
}


G_wins_list <- numeric(0)
for (i in winner_ioc_list) {
    G_wins_list <- c(G_wins_list, sum(master_df$winner_ioc == i & master_df$winning_level ==
        "G"))
}


M_wins_list <- numeric(0)
for (i in winner_ioc_list) {
    M_wins_list <- c(M_wins_list, sum(master_df$winner_ioc == i & master_df$winning_level ==
        "M"))
}
```

```r
level_df <- data.frame(country = winner_ioc_list, ATP = A_wins_list, Grandslams = G_wins_list,
    Masters = M_wins_list)

# extract our top 5 countries from earlier for further comparison
top5_levels <- level_df[row_indices_top5, ]
top5_levels
```

```
##   country ATP Grandslams Masters
## 1     ESP  65          8      13
## 5     SRB  18          9      18
## 6     SUI  22          4       8
## 7     FRA  28          0       1
## 3     USA  25          0       1
```

```r
# plot for visual comparison
barplot(height = t(as.matrix(top5_levels[, 2:4])), names.arg = (top5_names),
    xpd = F, main = "Top 5 country titles by level", xlab = "Country", ylab = "Number of titles",
    ylim = c(0, 70), axes = F, col = c(rgb(0.8, 0.5, 0.2), rgb(1, 0.85, 0),
        rgb(0.75, 0.75, 0.75)), legend = colnames(top5_levels)[-1], beside = TRUE)
axis(side = 2, at = seq(from = 0, to = 70, by = 10), tick = TRUE, las = 2)
```


Top 5 country titles by level
Number of titles
ATP
Grandslams
Masters
ESP
SRB
SUI
FRA
USA
Country