Could an Independent Yorkshire Win the World Cup - Rest of the World/UK

Robert Hickman

2018-06-10

Recently, a Yorkshire national football team appeared in a league of national teams for stateless people. This got me wondering how the historic counties of the UK would do at the world cup. Could any of them compete with full international teams?

I published the complete code for that article on this blog this week. However, one question which I kept being asked was how a 'All of the UK' team would do (i.e. if the country wasn't split up into England, Wales, Scotland, and Northern Ireland). Listening to the latest Double Pivot Podcast, drafting plyers not going to the World Cup, I also wondered what a 'Rest of the World' 11 would look like/fare.

```
library(dplyr)
library(magrittr)
library(data.table)
library(ggplot2)
```

Building Teams

To save time, I'm gonig to used saved versions of the datasets I built up over the 5 blog posts.

```
#world rankings
world_rankings <- readRDS("national_rankings.rds")

#player data
all_players_data <- readRDS("all_players_position_data.rds")
#all British players
british_player_birthplaces <- readRDS("british_player_birthplaces.rds")

#the countries going to the world cup
world_cup_countries <- c("Russia", "Saudi Arabia", "Egypt", "Uruguay",
                         "Portugal", "Spain", "Morocco", "Iran",
                         "France", "Australia", "Peru", "Denmark",
                         "Argentina", "Iceland", "Croatia", "Nigeria",
                         "Brazil", "Switzerland", "Costa Rica", "Serbia",
                         "Germany", "Mexico", "Sweden", "Korea Republic",
                         "Belgium", "Panama", "Tunisia", "England",
                         "Poland", "Senegal", "Colombia", "Japan")

#load data to save having to recalculate optimal teams
optimal_national_teams <- readRDS("optimal_national_teams.rds")
national_teams <- readRDS("national_teams.rds")

#the formations for selecting teams
formations_df <- readRDS("formations_df.rds")
formation_coords <- readRDS("player_position_coords.rds")
```

I won't include the functions in this blog post either, but the article uses (at most very slight modified) functions from the previous 5 posts.

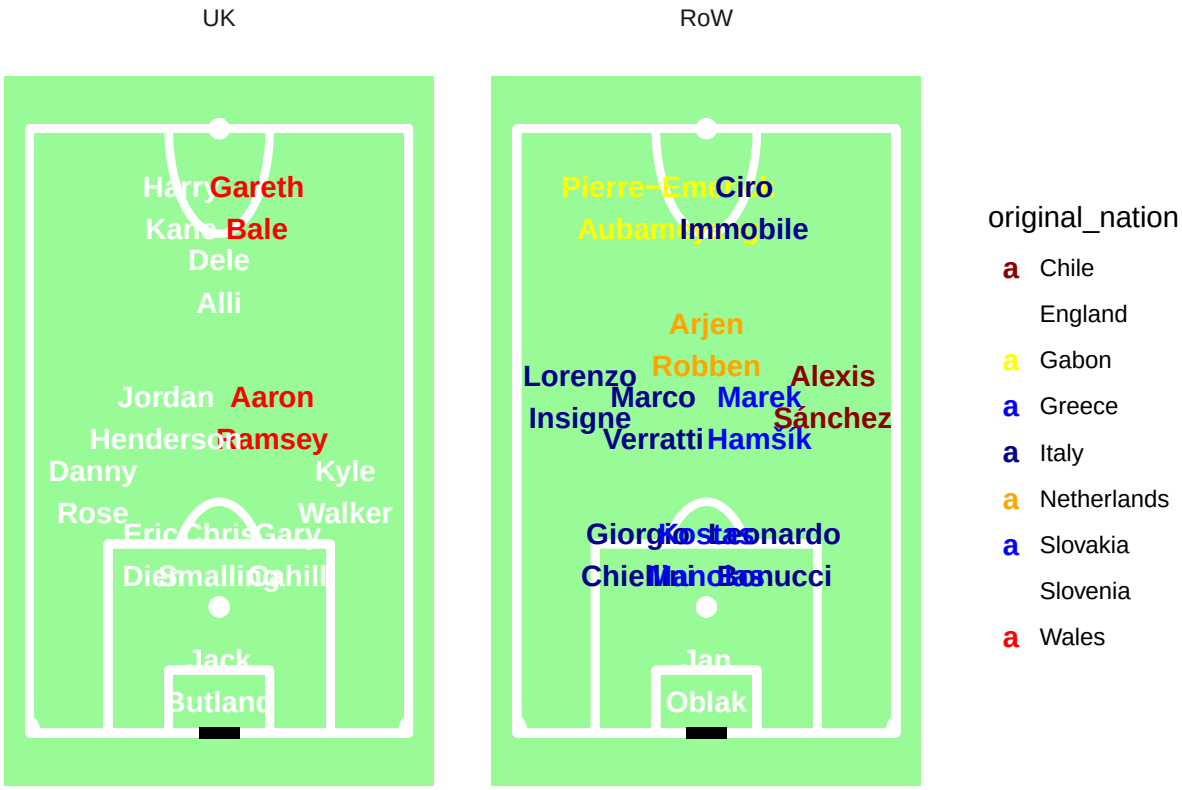
We first need to sort the players into either the UK vs. the rest of the World* and finding the optimal teams for each, as we did prviously.

*it's possible Welsh (especially Gareth Bale), Northern Irish, or Scottish players might make the rest of the World team, but I'll ignore that possibility for simplicity

```
#get the names of each player to merge in
player_lookup <- all_players_data %>%
  select(id, name, nationality) %>%
  mutate(original_nation = as.character(nationality))

#sort the data for finding teams
nationalised_players <- all_players_data %>%
  setDT() %>%
  #convert british players nationality to UK
  .[id %in% british_player_birthplaces$id, nationality := "UK"] %>%
  #filter out players from countries at the world cup
  .[!nationality %in% world_cup_countries] %>%
  #convert non-UK players nationality to "Rest of World"
  .[!id %in% british_player_birthplaces$id, nationality := "RoW"]

#find the optimal teams for both these nations
extranational_teams <- rbindlist(lapply(unique(nationalised_players$nationality), find_optimal_team,
                                        select(nationalised_players, id, nationality, 49:60), replica
```

These can then be plotted to show the teams as before.

```
#select the best 4 county teams by total ability
extranational_teams %<>%
  setDT() %>%
  .[, unique_position := make.unique(as.character(position)), by = "nation"] %>%
  merge(., formation_coords, by = c("formation", "unique_position")) %>%
  merge(player_lookup, by = "id")

#plot the data
p <- ggplot(data = extranational_teams)
p <- p %>%
  #custom pitch aesthetic function
  draw_pitch()
p <- p +
  geom_text(aes(x = player_x, y = player_y, label = gsub(" ", "\n", name), colour = original_nation), f
  scale_colour_manual(values = c("darkred", "white", "yellow", "blue", "darkblue", "orange", "blue", "w
  facet_wrap(~nation)

plot(p)
```



Calculating Ability

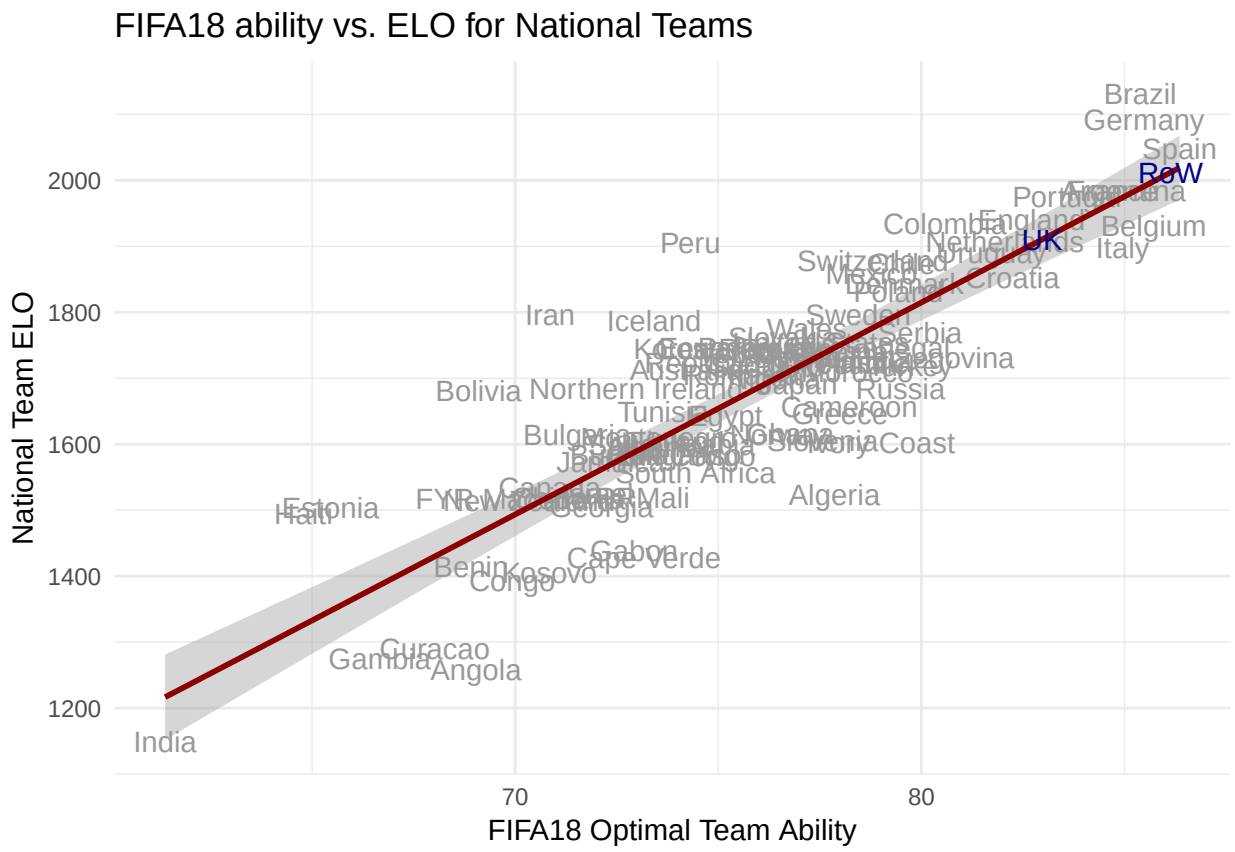
As previously, we can calculate the expected ELO of such teams via linear regression of the FIFA18 ability vs. ELO of actual national teams.

This time, let's predict the ability of the extranational teams based on this regression before plotting, just to save on plots/time/code/etc.

```
#merge in the world rankings for each fieldable national team
national_teams %<>% merge(., world_rankings, by = "nation") %>%
  #merge in the optimal team total_ability for each nation
  merge(., unique(select(optimal_national_teams, nation, total_ability)), by = "nation")

#regress ELO against total_ability (as judged by selection of FIFA18 players)
ability_regression <- lm(data = national_teams, ELO ~ total_ability)

#munge the extranational teams df to predict the ELO
extranational_teams <- data.frame(nation = c("UK", "RoW")) %>%
  merge(., select(extranational_teams, nation, total_ability), by = "nation") %>%
  #predict the ELO of each county using the previous regression
  mutate(predicted_ELO = predict(ability_regression, .)) %>%
  unique()
```

Then we can plot this regression and see where the RoW and UK fall in terms of actual nations

```
#plot ELO vs. total_ability
p <- ggplot(data = national_teams, aes(x = total_ability, y = ELO)) +
  geom_text(aes(label = nation), colour = "grey60") +
  #add in the linear regression line + confidence intervals
  stat_smooth(method = "lm", colour = "darkred") +
  geom_text(data = extranational_teams, aes(label = nation, x = total_ability, y = predicted_ELO), colou
  xlab("FIFA18 Optimal Team Ability") +
  ylab("National Team ELO") +
  ggtitle("FIFA18 ability vs. ELO for National Teams") +
  theme_minimal()

plot(p)
```



What's quite nice about the graph is it shows the limitation of this approach. By definition, a UK team should be *at least* as good as the English national team, but because England overperform their 'FIFA ability', the UK is actually ranked a fair bit lower in terms of ELO

```
#show the ELOs of the English national football team
#and predicted ELO of a UK team
national_teams$ELO[national_teams$nation == "England"]
```

```
## [1] 1941
```

```
extranational_teams$predicted_ELO[extranational_teams$nation == "UK"]
```

```
##       12
## 1910.421
```

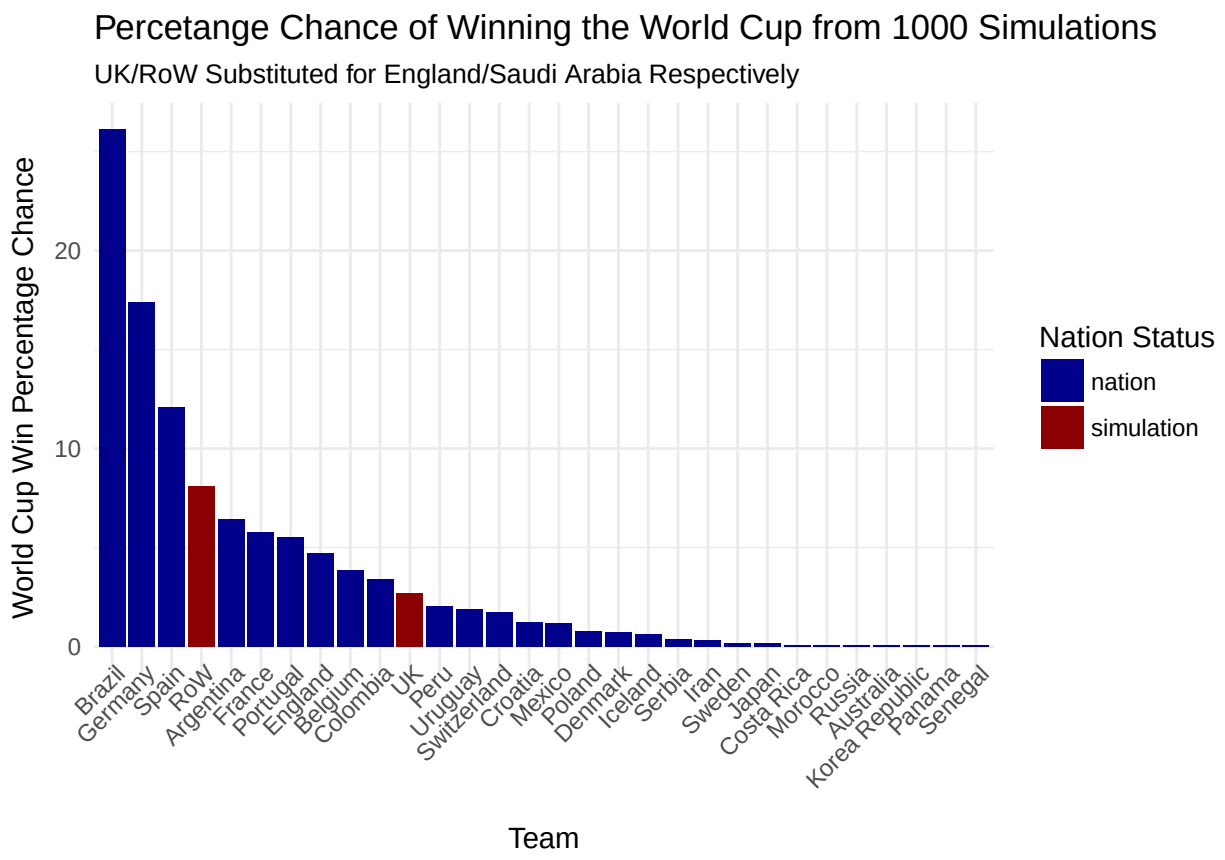
The RoW team is similarly probably undervalued in terms of ELO. FIFA18 ranks the players as a lot better than teams like Germany and Brazil, but with much lower ELO

We can then run the simulations, swapping the UK/RoW in for countries. The obvious substitute for the UK is England. For the RoW I decided to remove the team with the lowest ELO, which turns out to be Saudi Arabia

```
#merge the ELOs with the world cup draw information
wc_teams %<>% merge(., select(national_teams, nation, ELO) %>%
                      rbind(., data.frame(nation = "Panama", ELO = 1669)), by = "nation")
wc_teams$nation <- as.character(wc_teams$nation)

simulate_counties <- function(extranation, simulations, replace_country) {
  #replace Englands ELO with that of the county team replacing them
  wc_teams$ELO[wc_teams$nation == replace_country] <- extranational_teams$predicted_ELO[extranational_t
  wc_teams$nation[wc_teams$nation == replace_country] <- extranation

  #run x number of simulations
  for(simulation in 1:simulations) {
    winner <- simulate_tournament(wc_teams, knockout_matches, group_matches)
    if(simulation == 1) {
      winners <- winner
    } else {
      winners <- append(winners, winner)
    }
  }

  #spit out a df with each winner and the number of times they win
  simulation_df <- data.frame(table(winners))
  names(simulation_df) <- c("nation", "championships")

  #work out the percentage chane of each nation/county winning
  simulation_df$percentage <- simulation_df$championships / (simulations/100)
  return(simulation_df)
}

#run the simulations
UK_simulation <- simulate_counties("UK", 1000, "England") %>%
  mutate(simulation = "UK")
RoW_simulation <- simulate_counties("RoW", 1000, wc_teams$nation[which.min(wc_teams$ELO)]) %>%
  mutate(simulation = "RoW")
```

```
simulation_results <- rbind(UK_simulation, RoW_simulation) %>%
  setDT() %>%
    .[, perc_chance := mean(percentage), by = "nation"] %>%
  .[, c("nation", "perc_chance")] %>%
  unique(.) %>%
  .[nation %in% c("RoW", "UK"), nation_status := "simulation"] %>%
  .[!nation %in% c("RoW", "UK"), nation_status := "nation"] %>%
  #order by percentage chance of winning the WC
  .[, nation := factor(nation, levels = nation[order(-.$perc_chance)])]

#plot the results
p <- ggplot(data = simulation_results, aes(x = nation, y = perc_chance)) +
  geom_bar(stat = "identity", aes(fill = nation_status)) +
```

```
  scale_fill_manual(values = c("darkblue", "darkred"), name = "Nation Status") +
  xlab("Team") +
  ylab("World Cup Win Percentage Chance") +
  ggtitle("Percetange Chance of Winning the World Cup from 1000 Simulations",
          subtitle = "UK/RoW Substituted for England/Saudi Arabia Respectively") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, vjust = 1.2))

plot(p)
```



The team for the RoW tend to do fairly well. In fact only Brazil, Germany, or Spain (3 of the tournament favourites) tend to win more simulated World Cups than them. The team for the whole of the UK disappoints as much as the English national team, winning about the same as the original, and other similarly ranked nations, such as Colombia, or Peru.