FIFA WC outcomes prediction

Thomas Fernandes

2023-04-12

1. Introduction

1.1 Ce que dit l'article

Cet article examine la prévisibilité des résultats de la Coupe du monde de football en utilisant deux modèles empiriques. Depuis 1993, le système de classement mensuel (couramment appelé coefficient des pays) de la FIFA pour les équipes nationales de football seniors est devenu une source fiable pour juger du potentiel d'une équipe dans la compétition. Cependant, l'article montre que, malgré le nombre de surprises dans les résultats des matchs, les équipes les mieux classées ont généralement de meilleures performances. Les auteurs analysent les résultats des quatre dernières Coupes du monde de la FIFA de 1994 à 2006 et utilisent les classements mensuels de la FIFA pour évaluer les performances des équipes. Deux modèles empiriques sont proposés : un modèle probit et un modèle de régression linéaire MCO. Le premier modèle évalue la probabilité de victoire pour chaque équipe en fonction de son classement FIFA, tandis que le second modèle utilise la différence de buts comme variable dépendante pour voir si les équipes les mieux classées ont généralement des performances supérieures. Les auteurs trouvent des preuves empiriques que, malgré les surprises occasionnelles, les équipes les mieux classées ont généralement de meilleures performances.

1.2 Ce qu'on cherche à faire

A rédiger

2. Traitement des données

Les données liées aux coupes du monde n'étant officiellement plus disponible gratuitement sur le site de la FIFA, il faut se baser sur des données d'utilisateurs qui ont scrap le site de la FIFA dans l'objectif de répliquer ces bases de données.

Classement FIFA : https://www.kaggle.com/datasets/cashncarry/fifaworldranking — Alex Zabrodin Ces données comprennent le classement fifa de 1992 (date de création) à nos jours.

Affluence : https://www.kaggle.com/datasets/abecklas/fifa-world-cup — Andre Becklas Détails des affluences ainsi que des scores des matchs de coupe du monde de 1a première en 1930 à celle de 2014

Coupe du monde 2022 : https://www.kaggle.com/datasets/die9origephit/fifa-world-cup-2022-complete-dataset — Diego Farchione Database complète de toutes les statistiques de chaque match de la coupe du monde 2022

Matchs : https://github.com/jfjelstul/worldcup — Josh Fjelstul Base de données comprenant les scores de tous les matchs de coupe du monde de sa créaion à nos jours

All : https://www.kaggle.com/datasets/piterfm/fifa-football-world-cup — Petro Ivaniuk Base de données complète reprenant les contenant le score de tous les matchs de coupe du monde (1930-2022), les cartons (jaunes et rouges), l'affluence

ALL2: https://www.kaggle.com/datasets/kaito510/fifa-world-cup-match-stats — Kaito G. Contient les données de possession de balle, carton, tir cadrés de 2001 à 2021.

2.1 Traitement des données de match

```r
#On importe ce csv
WC_matches <- read.csv("WC Data/matches_1930_2022 (All).csv",
                       header = TRUE,
                       sep = ",",
                       stringsAsFactors = FALSE)

WC_matches <- WC_matches[c("home_team",
                           "away_team",
                           "home_score",
                           "away_score",
                           "Attendance",
                           "Round",
                           "Year",
                           "home_red_card",
                           "away_red_card",
                           "home_yellow_card_long",
                           "away_yellow_card_long",
                           "home_substitute_in_long",
                           "away_substitute_in_long")]

#On renomme les colonnes
colnames(WC_matches) <- c("Home",
                          "Away",
                          "Home.Goals",
                          "Away.Goals",
                          "Attendance",
                          "Round",
                          "Year",
                          "Home.Red.Card",
                          "Away.Red.Card",
                          "Home.Yellow.Card",
                          "Away.Yellow.Card",
                          "Home.Substitute.In",
                          "Away.Substitute.In")

#On ajoute une colonne de différence de buts
WC_matches$Goals_Diff <- WC_matches$Home.Goals - WC_matches$Away.Goals

#Home Red Card Count
WC_matches$Home.Red.Card.Count <- sapply(strsplit(WC_matches$Home.Red.Card, "\\|"), function(x) length(

#Away Red Card Count
WC_matches$Away.Red.Card.Count <- sapply(strsplit(WC_matches$Away.Red.Card, "\\|"), function(x) length(

#Red card diff
WC_matches$Red.Card.Diff <- WC_matches$Home.Red.Card.Count - WC_matches$Away.Red.Card.Count

#Home Yellow Card Count
WC_matches$Home.Yellow.Card.Count <- sapply(strsplit(WC_matches$Home.Yellow.Card, ","), function(x) leng
```

```r
#Away Yellow Card Count
WC_matches$Away.Yellow.Card.Count <- sapply(strsplit(WC_matches$Away.Yellow.Card, ","), function(x) leng

#Yellow card diff
WC_matches$Yellow.Card.Diff <- WC_matches$Home.Yellow.Card.Count - WC_matches$Away.Yellow.Card.Count

#Home Substitute In Count
WC_matches$Home.Substitution.Count <- sapply(strsplit(WC_matches$Home.Substitute.In, ","), function(x) l

#Away Substitute In Count
WC_matches$Away.Substitution.Count <- sapply(strsplit(WC_matches$Away.Substitute.In, ","), function(x) l

#Substitution diff
WC_matches$Substitution.Diff <- WC_matches$Home.Substitution.Count - WC_matches$Away.Substitution.Count

#On renomme WC_matches en WC_database
WC_database <- WC_matches

#On ne garde que les données après 1994
WC_database <- subset(WC_database, Year >= 1994)

#On supprime les colonnes Home.Red.Card, Away.Red.Card, Home.Yellow.Card, Away.Yellow.Card, Home.Substi
WC_database <- WC_database[,-c(8,9,10,11,12, 13)]
```

2.2 Traitement des données de match supplémentaires

Pour les matchs les plus récents (à partir de la cdm 2014), nous avons des données supplémentaires sur la possession de balle, le nombre de tir, de tir cadré, de fautes et d'arrêt

```r
WC_matches_stats <- read.csv("WC Data/FIFAallMatchBoxData (Possesion 2001-2021).csv",
                    header = TRUE,
                    sep = ",",
                    stringsAsFactors = FALSE)

#On ne garde que les années de coupe du monde
WC_matches_stats <- subset(WC_matches_stats, year %in% c(2006,
                                                         2010, 2014, 2018, 2022))

#Rename south korea as korea republic
WC_matches_stats <- WC_matches_stats %>%
  mutate(hname = ifelse(hname == "South Korea", "Korea Republic", hname))
WC_matches_stats <- WC_matches_stats %>%
  mutate(aname = ifelse(aname == "South Korea", "Korea Republic", aname))

#On ajoute les différences des nouvelles stats
WC_matches_stats <- WC_matches_stats %>%
  mutate(Possession.Diff = hPossesion - aPossesion)
WC_matches_stats <- WC_matches_stats %>%
  mutate(Shots.Diff = hshots - ashots)
WC_matches_stats <- WC_matches_stats %>%
      mutate(Shots.On.Target.Diff = hshotsOnTarget - ashotsOnTarget)
WC_matches_stats <- WC_matches_stats %>%
      mutate(Fouls.Diff = hfouls - afouls)
```

```r
WC_matches_stats <- WC_matches_stats %>%
      mutate(Saves.Diff = hsaves - asaves)

# Merge des dataframes WC_matches_stats et WC_database en conservant toutes les observations de WC_data
WC_database_2 <- merge(WC_database, WC_matches_stats, by.x = c("Home", "Away", "Year"), by.y = c("hname
WC_database <- WC_database_2

# Supprimer le dataframe temporaire
rm(WC_database_2)


#On enlève les colonnes qui sont doublées
WC_database <- subset(WC_database, select = -c(hgoals,
                                               agoals,
                                               hyellowCards,
                                               ayellowCards,
                                               hredCards,
                                               aredCards))
```

2.3 Traitement des données de la coupe du monde 2022

Les données de la dernière coupe du monde ne sont pas incluses dans les base de données précédemment utilisé. On va donc se servir d'une nouvelle base de données et la même autre même format que WC_database pour pouvoir ensuite merge les deux.

```r
#On importe le csv : "WC Data/Fifa_world_cup_matches_2022 (Cards).csv"
WC_matches_2022 <- read.csv("WC Data/Fifa_world_cup_matches_2022 (Cards).csv",
                    header = TRUE,
                    sep = ",",
                    stringsAsFactors = FALSE)

#On ajoute une colonne Year pour pouvoir merge avec les autres données
WC_matches_2022$Year <- 2022

#On ajoute une colonne de différence de buts
WC_matches_2022$Goals_Diff <- WC_matches_2022$number.of.goals.team1 - WC_matches_2022$number.of.goals.te

#On ajoute une colonne de différence de cartons jaunes
WC_matches_2022$Yellow_Card_Diff <- WC_matches_2022$yellow.cards.team1 - WC_matches_2022$yellow.cards.te

#On ajoute une colonne de différence de cartons rouges
WC_matches_2022$Red_Card_Diff <- WC_matches_2022$red.cards.team1 - WC_matches_2022$red.cards.team2

#On ajoute une colonne de différence de remplacements~
#Il n'y en a pas dans la base de données

#On ajoute une colonne de différence de possession de balle
#Il faut d'abord supprimer le % de la colonne
WC_matches_2022$possession.team1 <- gsub("%", "", WC_matches_2022$possession.team1)
WC_matches_2022$possession.team2 <- gsub("%", "", WC_matches_2022$possession.team2)
WC_matches_2022$possession.team1 <- as.numeric(WC_matches_2022$possession.team1)
WC_matches_2022$possession.team2 <- as.numeric(WC_matches_2022$possession.team2)
WC_matches_2022$Possession.Diff <- WC_matches_2022$possession.team1 - WC_matches_2022$possession.team2
```

```r
#On ajoute une colonne de différence de tir (total.attacks.team1)
WC_matches_2022$Shots.Diff <- WC_matches_2022$total.attempts.team1 - WC_matches_2022$total.attempts.tea

#On ajoute une colonne de différence de tir cadré (on.target.attempts.team1)
WC_matches_2022$Shots.On.Target.Diff <- WC_matches_2022$on.target.attempts.team1 - WC_matches_2022$on.t

#Les colonnes fouls sont en "against" et non "committed by" donc on les renomme
names(WC_matches_2022)[names(WC_matches_2022) == "fouls.against.team1"] <- "fouls.team1"
names(WC_matches_2022)[names(WC_matches_2022) == "fouls.against.team2"] <- "fouls.team2"

#On ajoute une colonne de différence de fautes
WC_matches_2022$Fouls.Diff <- WC_matches_2022$fouls.team1 - WC_matches_2022$fouls.team2

#On garde que les colonnes qui nous intéressent
WC_matches_2022_22 <- subset(WC_matches_2022, select = c(1:7, 11, 12, 21, 22, 57:62, 89:96))

#On renomme les colonnes pour qu'elles soient identiques à celles de WC_database
names(WC_matches_2022_22)[names(WC_matches_2022_22) == "team1"] <- "Home"
names(WC_matches_2022_22)[names(WC_matches_2022_22) == "team2"] <- "Away"
names(WC_matches_2022_22)[names(WC_matches_2022_22) == "number.of.goals.team1"] <- "Home.Goals"
names(WC_matches_2022_22)[names(WC_matches_2022_22) == "number.of.goals.team2"] <- "Away.Goals"
names(WC_matches_2022_22)[names(WC_matches_2022_22) == "yellow.cards.team1"] <- "Home.Yellow.Cards.Coun
names(WC_matches_2022_22)[names(WC_matches_2022_22) == "yellow.cards.team2"] <- "Away.Yellow.Cards.Coun
names(WC_matches_2022_22)[names(WC_matches_2022_22) == "red.cards.team1"] <- "Home.Red.Cards.Count"
names(WC_matches_2022_22)[names(WC_matches_2022_22) == "red.cards.team2"] <- "Away.Red.Cards.Count"
names(WC_matches_2022_22)[names(WC_matches_2022_22) == "total.attempts.team1"] <- "hshots"
names(WC_matches_2022_22)[names(WC_matches_2022_22) == "total.attempts.team2"] <- "ashots"
names(WC_matches_2022_22)[names(WC_matches_2022_22) == "on.target.attempts.team1"] <- "hshotsOnTarget"
names(WC_matches_2022_22)[names(WC_matches_2022_22) == "on.target.attempts.team2"] <- "ashotsOnTarget"
names(WC_matches_2022_22)[names(WC_matches_2022_22) == "fouls.team1"] <- "hfouls"
names(WC_matches_2022_22)[names(WC_matches_2022_22) == "fouls.team2"] <- "afouls"
names(WC_matches_2022_22)[names(WC_matches_2022_22) == "possession.team1"] <- "hPossession"
names(WC_matches_2022_22)[names(WC_matches_2022_22) == "possession.team2"] <- "aPossession"

#On va maintenant cherche à faire une base de données avec les données de la Coupe du Monde 2022 et les

# remplacer les valeurs manquantes dans WC_database avec les valeurs correspondantes dans WC_matches_20
WC_database$Home <- tolower(trimws(WC_database$Home))
WC_database$Away <- tolower(trimws(WC_database$Away))
WC_database$Year <- as.integer(WC_database$Year)

WC_matches_2022_22$Home <- tolower(trimws(WC_matches_2022_22$Home))
WC_matches_2022_22$Away <- tolower(trimws(WC_matches_2022_22$Away))
WC_matches_2022_22$Year <- as.integer(WC_matches_2022_22$Year)



for (i in 1:nrow(WC_database)) {
  if (is.na(WC_database$hshots[i]) && WC_database$Year[i] == 2022) {
    home_team <- WC_database$Home[i]
    away_team <- WC_database$Away[i]
    year <- WC_database$Year[i]
    matching_row <- match(paste(home_team, away_team, year),
```

```
                          paste(WC_matches_2022_22$Home, WC_matches_2022_22$Away, WC_matches_2022_22$Yea
    WC_database$hshots[i] <- WC_matches_2022_22$hshots[matching_row]
    WC_database$ashots[i] <- WC_matches_2022_22$ashots[matching_row]
    WC_database$hPossesion[i] <- WC_matches_2022_22$hPossession[matching_row]
    WC_database$aPossesion[i] <- WC_matches_2022_22$aPossession[matching_row]
    WC_database$hshotsOnTarget[i] <- WC_matches_2022_22$hshotsOnTarget[matching_row]
    WC_database$ashotsOnTarget[i] <- WC_matches_2022_22$ashotsOnTarget[matching_row]
    WC_database$hfouls[i] <- WC_matches_2022_22$hfouls[matching_row]
    WC_database$afouls[i] <- WC_matches_2022_22$afouls[matching_row]
    WC_database$Home.Yellow.Cards.Count[i] <- WC_matches_2022_22$Home.Yellow.Cards.Count[matching_row]
    WC_database$Away.Yellow.Cards.Count[i] <- WC_matches_2022_22$Away.Yellow.Cards.Count[matching_row]
    WC_database$Home.Red.Cards.Count[i] <- WC_matches_2022_22$Home.Red.Cards.Count[matching_row]
    WC_database$Away.Red.Cards.Count[i] <- WC_matches_2022_22$Away.Red.Cards.Count[matching_row]
    WC_database$Possession.Diff[i] <- WC_matches_2022_22$Possession.Diff[matching_row]
    WC_database$Shots.Diff[i] <- WC_matches_2022_22$Shots.Diff[matching_row]
    WC_database$Shots.On.Target.Diff[i] <- WC_matches_2022_22$Shots.On.Target.Diff[matching_row]
    WC_database$Fouls.Diff[i] <- WC_matches_2022_22$Fouls.Diff[matching_row]
    WC_database$Yellow.Cards.Diff[i] <- WC_matches_2022_22$Yellow.Cards.Diff[matching_row]
    WC_database$Red.Cards.Diff[i] <- WC_matches_2022_22$Red.Cards.Diff[matching_row]
  }
}
```

2.4 Traitement du classement FIFA

```
#On importe le deuxième csv : "fifa_ranking-2022-12-22 (Rank)"
FIFA_rank <- read.csv("WC Data/fifa_ranking-2022-12-22 (Rank).csv",
                      header = TRUE,
                      sep = ",",
                      stringsAsFactors = FALSE)

#On garde uniquement les classements datés au mois de mai de chaque année et on ajoute une nouvelle col
FIFA_rank <- subset(FIFA_rank, ifelse(year(FIFA_rank$rank_date) == 2022, month(FIFA_rank$rank_date) == 
FIFA_rank$rank_date <- as.Date(FIFA_rank$rank_date)
FIFA_rank$Year <- format(FIFA_rank$rank_date, "%Y")

#On ne garde que les années de la Coupe du Monde
FIFA_rank <- subset(FIFA_rank,year(FIFA_rank$rank_date) %in% c(1994, 1998, 2002, 2006,
                                                               2010, 2014, 2018, 2022))

#On renomme les entrées de "country_full" USA en "United States" pour le merge
FIFA_rank$country_full <- gsub("USA", "United States", FIFA_rank$country_full)
#Turkey en Türkiye
FIFA_rank$country_full <- gsub("Turkey", "Türkiye", FIFA_rank$country_full)
#Yugoslavia en fr Yugoslavia
FIFA_rank$country_full <- gsub("Yugoslavia", "fr Yugoslavia", FIFA_rank$country_full)

#Colonne Year en numérique
FIFA_rank$Year <- as.integer(FIFA_rank$Year)
FIFA_rank$country_full <- tolower(FIFA_rank$country_full)

#On ajoute les colonnes "FIFA_Rank.Home" et "FIFA_Rank.Away" dans le dataframe "WC_database"
WC_database$FIFA_Rank.Home <- 0
WC_database$FIFA_Rank.Away <- 0
```

```r
#Fonction qui récupère le classement FIFA d'une équipe à une année donnée
get_rank <- function(year, team){
  rank <- FIFA_rank$rank[which(FIFA_rank$country_full == team & FIFA_rank$Year == year)]
  if(length(rank) > 0){
    return(rank)
  } else {
    return(NA)
  }
}

#On ajoute les classements FIFA dans le dataframe "WC_database"
for (i in 1:nrow(WC_database)) {
  WC_database$FIFA_Rank.Home[i] <- get_rank(WC_database$Year[i], WC_database$Home[i])
  WC_database$FIFA_Rank.Away[i] <- get_rank(WC_database$Year[i], WC_database$Away[i])
}
```