Chilean Journal of Statistics
Vol. 9, No. 1, April 2018, 59-65

Short Note

The next winner of the 2018 FIFA World Cup will be...: An illustration of the use of statistical simulation to make a prediction in a complex tournament

Juan Carlos Correa M.[1,*], Carlos Barrera-Causil[2], and Fernando Marmolejo-Ramos[3]

[1]School of Statistics, National University of Colombia, Medellín, Colombia,
[2]Faculty of Applied and Exact Sciences, Metropolitan Technological Institute, Medellín, Colombia,
[3]School of Psychology, The University of Adelaide, Adelaide, Australia

(Received: February 20, 2018 · Accepted in final form: March 11, 2018)

Abstract

Statistical simulations are used in both teaching and research contexts and are a powerful tool for solving complex problems for which there are no exact solutions or which require too many resources. Furthermore, statistical simulations are a powerful computational tool to predict events. In this study, statistical simulations are used to analyse the progressive results of the 2018 FIFA World Cup, what teams are the most likely to play in the finals, and to predict the team which has the highest probability of being the champion. The ideas outlined in the discussion are likely to encourage readers to modify the R code provided in order to meet their own data and interests.

Keywords: 2018 FIFA World Cup · Statistical prediction · Statistical simulation · Teaching of statistics.

Mathematics Subject Classification: Primary 65C60.

1. Introduction

1.1 Structure of the tournament

Football is by far the most popular sport in the world. The Fédération Internationale de Football Association (FIFA) is the institution that regulates this sport and coordinates all the leagues worldwide, at both the professional league and country levels. The World Cup is the biggest sports tournament in the world, which is held every four years. The national teams play regional qualifiers organised by confederations. Those qualifiers determine the 32 teams that will progress to the final tournament. FIFA makes the final draw for the tournament by ballot, and the 32 teams are allocated into eight groups of four. The top two teams from each group move into the knockout stages (i.e. round 16, quarter finals, semi-finals and the final).

*Corresponding author. Email: fernando.marmolejoramos@adelaide.edu.au

ISSN: 0718-7912 (print)/ISSN: 0718-7920 (online)

http://www.soche.cl/chjs

1.2 Figuring the odds

On a regular basis, FIFA determines a global ranking of national teams. The rankings are based on the teams' results in international competitions (Lasek et al., 2013). Months before the tournament, fans speculate about each team's chances of reaching the World Cup final and who will be the ultimate champion. Obviously, there are favoured teams that traditionally have been rivals in the tournament. A key to gauge the possible success of a team is to determine its potential rivals. Because the draw is random, the calculation of an exact probability is impossible. Luckner et al. (2007) applied concepts of money movements in stock exchanges to the 2006 World Cup using the Soccer program. Hoffmann et al. (2002) built a model to predict the performance of a national team given such socio-economic and geographical conditions as a country's per capita income and climate of the capital city in which the games will be played.

By using the points of the teams as ranked by FIFA[1], the match schedules of the 2018 World Cup, and statistical simulations, we aimed to determine which team has the best chance of being the new world champion. Statistical simulation is a clearly defined and constantly developing area that allows complex problems to be solved (Lewis and Orav, 1989; Liu, 2004; Fishman, 1996; Kroese et al., 2014). This is an area with its own problems such as the development of random number generators (Deng and Linn, 2000). Applications of this technique range from elementary problems such as the calculation of the e number (Ripley, 1987; Russell, 1991) to simulations used in video games.

2. Methods

Table 1 presents the teams that will be in the 2018 World Cup, the group they belong to and the points as of April 8, 2018.

Table 1. *Countries participating in the 2018 FIFA World Cup. Each row represents the group stage. Countries are sorted row-wise from left to right according to their FIFA points*

	FIFA points							
Group	Country	Points	Country	Points	Country	Points	Country	Points
A	Uruguay	931	Egypt	687	Russia	531	Saudi Arabia	494
B	Portugal	1360	Spain	1228	Iran	792	Morocco	694
C	France	1185	Peru	1128	Denmark	1108	Australia	740
D	Argentina	1359	Croatia	1053	Iceland	1026	Nigeria	609
E	Brazil	1489	Switzerland	1197	Costa Rica	872	Serbia	780
F	Germany	1609	Mexico	1038	Sweden	1002	South Korea	554
G	Belgium	1337	England	1047	Tunisia	920	Panama	605
H	Poland	1228	Colombia	1106	Senegal	862	Japan	593

For the simulations, it is assumed that the probability that a team will win an elimination tournament match is a binomial process. Specifically, the probability of success of a particular team is given by the division between the team's points and the sum of the points of that team and the other team involved. In the first group stage, the simulation consisted of obtaining samples of size two without replacement of each group with probabilities proportional to the points of the four teams. This was carried out 200,000 times via the R software (R Core Team, 2017).

The complexity of the process is reflected in the fact that each winning team has a different probability of winning in subsequent matches. Hence, the probability depends on the results of the other competitors. Thus, each team makes a random walk, and the

[1]The points are available at http://www.fifa.com/fifa-world-ranking/ranking-table/men/

length of the paths depends on the difficulties encountered during the competition (see Appendix).

3. Results

Table 2 presents the probability of each participating team winning the 2018 World Cup. In theory, all the teams have the chance of being the champion. However, it is clear that there are teams that have much higher odds than others. These probabilities are also represented in Figure 1, where the size of the circles corresponds to the probability that each country has to win the 2018 FIFA World Cup (this figure was produced via the `leaflet` R package (Cheng et al., 2017)).

Table 2. *Probability of a participating country winning the 2018 FIFA World Cup. Countries are sorted row-wise from left to right according to their probability.*

Probability of Winning the Cup					
Germany	Brazil	Portugal	Belgium	Argentina	Spain
0.0880	0.0718	0.0676	0.0650	0.0602	0.0564
Poland	France	Switzerland	Colombia	Denmark	Peru
0.0504	0.0458	0.0406	0.0366	0.0362	0.0360
England	Mexico	Uruguay	Croatia	Iceland	Sweden
0.0340	0.0316	0.0310	0.0306	0.0300	0.0244
Senegal	Tunisia	Costa Rica	Iran	Serbia	Australia
0.0232	0.0218	0.0218	0.0168	0.0128	0.0118
Morocco	Egypt	Panama	Russia	Japan	South Korea
0.0116	0.0116	0.0064	0.0058	0.0056	0.0052
Nigeria	Saudi Arabia				
0.0050	0.0044				

Germany's position as the team favoured to win the 2018 World Cup is ratified in the result of the simulations. However, its favoritism is only 22.6% higher than Brazil (i.e. $(p_A - p_B)/p_B \cdot 100\% = (0.0880 - 0.0718)/0.0718 \cdot 100\%$), 30.2% higher than Portugal and 35.4% greater than Belgium.

A surprising feature of the fixture used for this championship is that it is possible to find any pair of teams from the 32 that can be in the final match. The simulations indicated that the most probable final will be between Brazil and Germany with a probability of 0.0132. The second and third most probable finals are Spain-Portugal ($p = 0.0128$) and Germany-Argentina ($p = 0.0108$). (The complete results can be obtained by running the R code shown in the Appendix.)

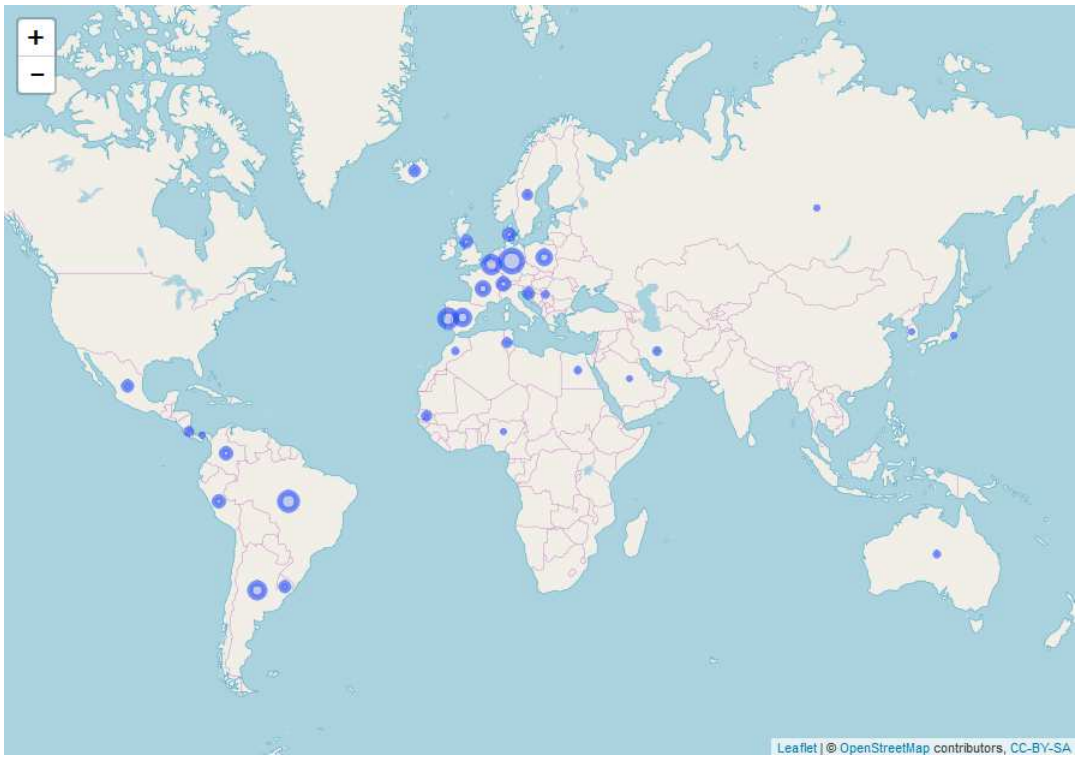

Figure 1. Leaflet map plot for the probability that each team will win 2018 FIFA World Cup.

4. Conclusions

Statistical simulation is a useful tool for solving complex problems and has been used in several areas (see Kroese et al., 2014). Indeed, simulations are so vital to the testing of statistical theories that teaching it in foundation statistical courses has been promoted by developing specific, free R packages (e.g. 'SimDesign' by Sigal and Chalmers, 2016). Other R packages add simulation functions on top of those used for data description, visualisation, and modelling (e.g. 'mosaic' by Pruim et al., 2017).

It is important to remember that simulations entail variability, randomness, and uncertainty. Hence, they are stochastic, not deterministic. Although the results of the simulations shown above enable probabilities to be assigned, we are satisfied knowing that such-and-such an event will happen with certain probability (e.g. even though the probability of raining in Adelaide during summer is very low, it does rain or it does not). The issue of probability is attached to simulations, but is beyond the scope of this article.

The ultimate goal of a simulation is to predict. Good predictions, however, depend on information availability and quality. The simulations reported herein are based merely on the FIFA points. Other information can be included or weighted to come up with more accurate predictions. For example, one could create measurements representing the skills of each player in each team, the quality of the teams' coaches, etc.[2] Those measurements can be combined into a single number for each team via composite indicators; this approach is commonly used in social sciences (see Marozzi, 2016).

We have presented a situation of general interest where, via simulation, it is possible to obtain answers in probabilistic terms to questions that arise between amateurs and

[2]There are many attempts to make predictions of local tournaments by statistical modeling or by using bets. One of the most difficult problems to model is the nationalist condition that any tournament possesses, since players from the same country possess a synergy that can overcome their own conditions. Another situation is the condition of short time that the tournament has (see Goddard and Asimakopoulos, 2004; Pachur and Biele, 2007).

gamblers.

Acknowledgements

The authors thank Susan Brunner and Trevor Jones for proofreading this manuscript. We also thank Rosie Gronthos for professionally proofreading this manuscript (rosie.gronthos@gmail.com). The last author thanks Iryna Losyeva and Alexandra Marmolejo-Losyeva for their unconditional support.

References

Cheng, J., Karambelkar, B., and Xie, Y. (2017). leaflet: Create interactive web maps with the JavaScript 'Leaflet' library. R package version 1.1.0. URL: https://CRAN.R-project.org/package=leaflet.

Deng, L., and Linn, D.K.J. (2000). Random number generation for the new century. The American Statistician 54, 145-150.

Fishman, G.S. (1996). Monte Carlo: Concepts, Algorithms, and Applications. Springer, New York.

Goddard, J., and Asimakopoulos, I. (2004). Forecasting football results and the efficiency of fixed-odds betting. Journal of Forecasting 23, 51-66.

Hoffmann, R., Ging, L.C., and Ramasamy, B. (2002). The socio-economic determinants of international soccer performance. Journal of Applied Economics V, 253-272.

Kroese, D.K., Brereton, T., Taimre, T., and Botev, Z.I. (2014). Why the Monte Carlo method is so important today. WIREs Computational Statistics 6, 386-392.

Lasek, K., Szlávik, Z., and Bhulai, S. (2013). The predictive power of ranking systems in association football. International Journal of Applied Pattern Recognition 1, 27-46.

Lewis, P.A.W., and Orav, E.J. (1989). Simulation Methodology for Statisticians, Operations Analysts, and Engineers, Volumen I. Chapman and Hall, London.

Liu, J.S. (2004). Monte Carlo Strategies in Scientific Computing. Springer, New York.

Luckner, S., Schröder, J., and Slamka, C. (2007). On the forecast accuracy of sports prediction markets. In H. Gimpel, N.R. Jennings, G.E. Kersten, A. Ockenfels, and C. Weinhardt (Eds.): Negotiation and Market Engineering, LNBIP 2. Springer-Verlag, Berlin, 227-234.

Marozzi, M. (2016). Construction, robustness assessment and application of an index of perceived level of socio-economic threat from immigrants: A study of 47 european countries and regions. Social Indicators Research 128, 413-437.

Pachur, T., and Biele, G. (2007). Forecasting from ignorance: The use and usefulness of recognition in lay predictions of sports events. Acta Psychologica 125, 99-116.

Pruim, R., Kaplan, D., and Horton, N.J. (2017). The mosaic package: Helping students to 'think with data' using R. The R Journal 9, 77-102.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org

Ripley, B.D. (1987). Stochastic Simulation. Wiley, New York.

Russell, K.G. (1991). Estimating the value of e by simulation. The American Statistician 45, 66-68.

Sigal, M.J., and Chalmers, R.P. (2016). Play it again: Teaching statistics with Monte Carlo simulation. Journal of Statistics Education 24, 136-156.

Appendix A. R code

```
# R version 3.4.3 interfaced with R Studio version 1.1.423

# Data Entry: Country Name and FIFA points
#             in groups of four teams conformed by the real groups.
#             the first four teams comprise group A, the second four
#             teams comprise group B, and so on.
team <- scan(what = list("",0))
Russia 531 Uruguay 931 Egypt 687 Saudi_Arabia 494
Portugal 1360 Spain 1228 Morocco 694  Iran 792
France 1185 Denmark 1108  Peru 1128 Australia 740
Argentina 1359 Croatia 1053   Nigeria 609 Iceland 1026
Brazil 1489 Switzerland 1197 Serbia 780 Costa_Rica 872
Germany 1609 Mexico 1038 Sweden 1002 South_Korea 554
Belgium 1337  England 1047 Panama 605 Tunisia 920
Poland 1228 Colombia 1106 Senegal 862   Japan 593

# Producing the labels for each group.
group <- rep(1:8, each=4)
position <- rep(1:4,8)
points <- team[[2]]
name <- team[[1]]
number <- 1:32
data <- cbind(number,points,group)
rownames(data) <- name

# simulation begins
Nsim <- 5000 # use 200000
winner <- NULL
Final <- NULL
for (simula in 1:Nsim) {
  resu <- NULL

  # Initial elimination
  # Two teams move onto the next part of the competition
  for (i in 1:8) {
    temp <- data[group==i,]
    sample <- sample(1:4,2,prob=temp[,2])
    resu <- rbind(resu,temp[sample,])
    names <- rownames(resu)
  }

  # Simple elimination
  # The matches are given by the fixture

  # Eighth finals
  p49 <- sample(c(1,4),1,prob=resu[c(1,4),2])
  p50 <- sample(c(5,8),1,prob=resu[c(5,8),2])
  p51 <- sample(c(3,2),1,prob=resu[c(3,2),2])
  p52 <- sample(c(7,6),1,prob=resu[c(7,6),2])
  p53 <- sample(c(9,12),1,prob=resu[c(9,12),2])
  p54 <- sample(c(13,16),1,prob=resu[c(13,16),2])
  p55 <- sample(c(11,10),1,prob=resu[c(11,10),2])
  p56 <- sample(c(15,14),1,prob=resu[c(15,14),2])
```

```
  # Quarter finals
  p57<-sample(c(p49,p50),1,prob=resu[c(p49,p50),2])
  p58<-sample(c(p53,p54),1,prob=resu[c(p53,p54),2])
  p59<-sample(c(p51,p52),1,prob=resu[c(p51,p52),2])
  p60<-sample(c(p55,p56),1,prob=resu[c(p55,p56),2])

  # Semifinals
  p61 <- sample(c(p57,p58),1,prob=resu[c(p57,p58),2])
  p62 <- sample(c(p59,p60),1,prob=resu[c(p59,p60),2])

  # Third and fourth places
  p63 <- sample(c(p61,p62),1,prob=resu[c(p61,p62),2])

  # Big final
  Final <- rbind(Final,c(names[p61],names[p62]))
  winner <- c(winner,names[p63])
} # End simulation

# table of probabilities
table(winner) / Nsim

# probabilities of the final
temp <- table(Final[,1],Final[,2])
temp2 <- (temp+t(temp))/Nsim
temp2
```