A New FIFA?

by Nick Powell and Seyla Ou

Background:

This paper is an investigation into the FIFA ranking system, used by the organization to seed international teams for tournament qualifying rounds as well as tournaments themselves. It was originally put in place in 1992 and underwent major changes in 1999 and 2006 to reflect what FIFA believed were important considerations for international football. The first iteration was extremely simple, awarding 3 points to teams for wins and 1 point for draws. Further iterations became increasingly complex; the current system takes into account four different factors, which will be explained in greater depth later in the paper.

The FIFA model has been heavily criticized on many fronts. In early years, when rankings considered team performances over an 8 year period, critics made the charge that they depended too heavily on old performances. The methodology used today has fixed that; however, it has introduced a host of other problems, namely the fact that a) playing too many friendly matches can bring down a team's rating, and b) team points stack infinitely. Friendly matches are essentially risk-free exhibition matches in which coaches can experiment with different strategies, formations, and players. Although these glorified scrimmages have no effect on tournament eligibility or status, they are nonetheless considered in FIFA's ranking system, and due to the intricacies of the ranking model, teams that play friendlies - even if they win - can be penalized for doing so. The Romanian national team hired a rankings consultant in 2014 who advised against playing friendly matches; as a result, Romania rose into the top 10 and became seeded for the 2018 World Cup qualifying campaign. This is a clear systemic flaw. As for the second point about points stacking infinitely we can see that tournament hosts, who qualify automatically and thus do not play any qualifiers, always fall in the rankings as they are unable to gain points that teams around them are getting from the qualifiers. France, who are hosting Euro 2016, are currently in 21st place - below teams like Hungary and Romania, who are almost certainly inferior teams.

Other ranking systems have been introduced that many argue are more relevant. The Elo system, developed in the 1960s for chess, can be used to model international football matches as well. It passes the eye test of being at least as accurate, if not more accurate, than the FIFA rankings, and Elo is not even a soccer specific model. If valid ranking procedures can be borrowed from completely unrelated fields, then certainly the FIFA model can be improved upon.

Research Question:

With the FIFA ranking system's flaws in mind, we set out to answer the question of whether or not we could devise a better ranking system than the one currently used in practice. To be more precise, we set out to determine whether or not we could create a system that was more fair, less prone to being manipulated, and ultimately free of the flaws present in the current system.

Before we could begin investigating the possibility of creating an improved ranking system, it was necessary for us to select a translatable metric with which we could objectively compare each model's performance. Below we outline the approach we took in choosing this metric, as well as the ways in which we applied the models we used.

Approaches / Models:

In order to acquire the data needed to build our models and test them against each other, we began by scraping match result data from soccerlotto.com. The site has amassed match result data from every international fixture dating as far back as the 1950s. We scraped the data from the 10,000 most recent games and partitioned them into two groups. We used the first group of 9000 games to train each model. After this training was complete, each model had produced its own unique ranking. We used the results from these rankings to predict the outcomes of the final 1000 games in the second group. From there we were able to laterally compare the percent of games each model was able to correctly predict. Of course, we couldn’t simply compare rankings and choose a winner - many football games finish in a tie. Therefore, we had to devise a method of predicting ties. To this end, we used the following R command:

$$ranking = lg(match(team, model) + 5)$$

This took the numerical rankings of each team (of the ~210 teams in each model), added 5, and took the base 2 log. ***This was done with all models to ensure uniformity***. We did not compare the numerical ratings produced by our models, because doing so would introduce biases; it was imperative that all models predicted draws with the same likelihood. Draws were predicted when

$$|logrank_a - logrank_h| < 0.5$$

In all other matchups, our model predicted the team with the higher rank to win.

We began with the FIFA model in order to get a sense of how accurate that model was in predicting match results. FIFA’s equation for rewarding points is [**P = M x I x T x C**], where **P** stands for points, **M** stands for match result (i.e. win or loss), **I** stands for importance of match (i.e. friendly, qualifier, etc.), **T** stands for opposing team’s strength, and **C** stands for confederation strength. We did not have to train the FIFA model in our experiment, since the ranking at the end of the 9000 matches (November 2013) already existed and was publicly available. Therefore, all we had to do was scrape that table and convert it into a vector in our script.

Once this was done, we began to experiment with creating our own ranking system. Our first attempt at this came by implementing the regularized normal Bradley-Terry model discussed in class. Since our data failed to show whether a match was played at a neutral field or not, we made the assumption that the home team actually played at home for each match. After training the B-T model on the first 9000 games, we were able to run two types of predictions. The first was the estimated beta match prediction approach in which we relied on the betas alone to predict each outcome. The second approach was exactly identical to the ranking-based predictions we made with the FIFA model.

Finally, we decided to devise our own strategy for ranking teams. The primary concern we wanted our model to address was infinite stacking. As such, we decided on the following model. At the beginning, every team begins level at zero points. For each of the 9000 games, we came up with a "magic number", which was in fact simply the following equation

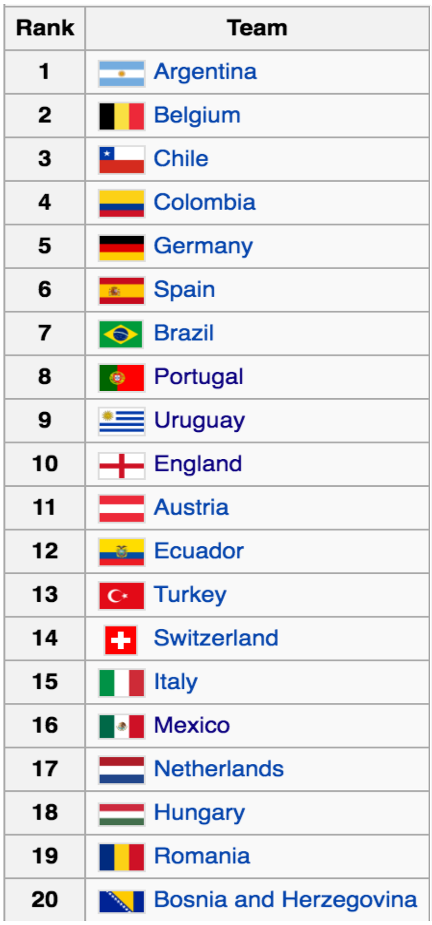
$$2 - e^{(} - log_{10}(ptdiff + 1))$$

This number was arrived at through trial and error. To calculate the number of points gained and lost, ***which was equal for EVERY game played (to maintain the 0 point mean),*** we took a base 2 log of the difference of the teams' scores and added 1 - we can call this number "x". If the "better" team (more points) won the game, we divided "x" by the "magic number" and added that many points to the winning team (and subtracted an equal number of points from the losing team). If the "worse" team won, we multiplied "x" by the "magic number". This was done so that a poor team beating a good team by X goals was worth more than in the rankings than a good team beating a poor team by X goals. At the end, we multiplied the score exchanged by an average of the strength of the two confederations represented, which we chose arbitrarily at the beginning.

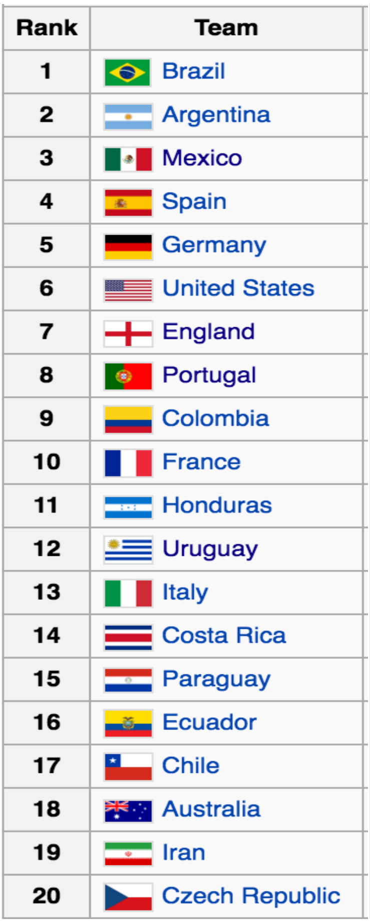
Results:

The top 20 teams in each of the rankings were as below.

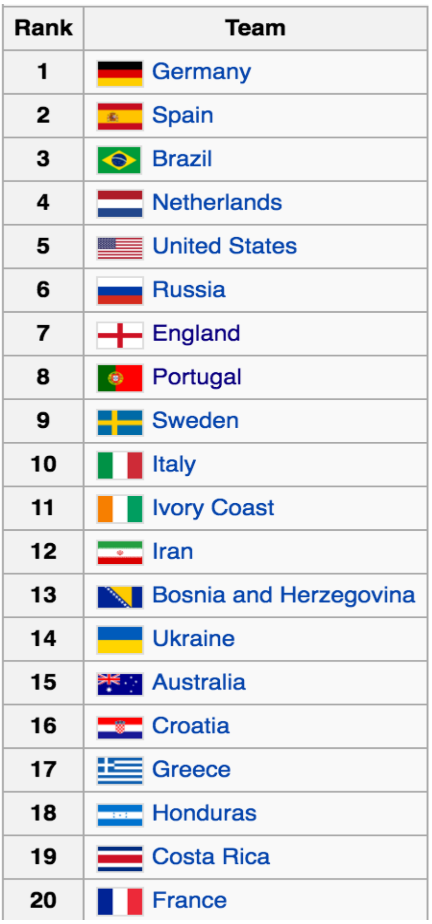
FIFA Rankings	**Bradley-Terry model**	**Our custom model**



Rank	Team
1	Argentina
2	Belgium
3	Chile
4	Colombia
5	Germany
6	Spain
7	Brazil
8	Portugal
9	Uruguay
10	England
11	Austria
12	Ecuador
13	Turkey
14	Switzerland
15	Italy
16	Mexico
17	Netherlands
18	Hungary
19	Romania
20	Bosnia and Herzegovina



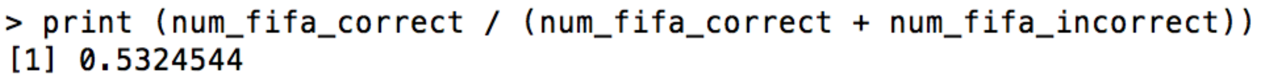
Rank	Team
1	Brazil
2	Argentina
3	Mexico
4	Spain
5	Germany
6	United States
7	England
8	Portugal
9	Colombia
10	France
11	Honduras
12	Uruguay
13	Italy
14	Costa Rica
15	Paraguay
16	Ecuador
17	Chile
18	Australia
19	Iran
20	Czech Republic



Rank	Team
1	Germany
2	Spain
3	Brazil
4	Netherlands
5	United States
6	Russia
7	England
8	Portugal
9	Sweden
10	Italy
11	Ivory Coast
12	Iran
13	Bosnia and Herzegovina
14	Ukraine
15	Australia
16	Croatia
17	Greece
18	Honduras
19	Costa Rica
20	France

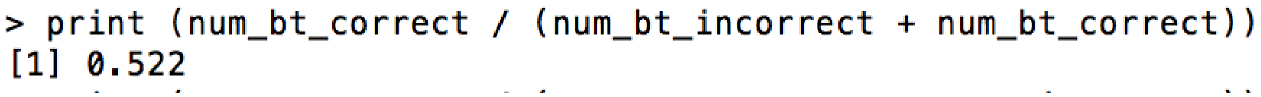
All three seemingly pass the eye test. There are no crazy outliers; no entirely unexpected teams. Russia at #6 in our model is a bit of a mystery, given that they do not appear in the top 20 of either of the other models, but the surprise isn't extraordinary. To get a better sense of how well each of these models performed, we have to look at the prediction success numbers.

FIFA

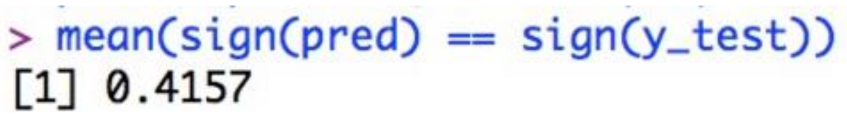


```
> print (num_fifa_correct / (num_fifa_correct + num_fifa_incorrect))
[1] 0.5324544
```

Bradley-Terry

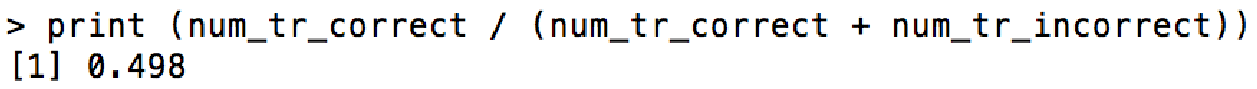


```
> print (num_bt_correct / (num_bt_incorrect + num_bt_correct))
[1] 0.522
```



```
> mean(sign(pred) == sign(y_test))
[1] 0.4157
```

Ours



```
> print (num_tr_correct / (num_tr_correct + num_tr_incorrect))
[1] 0.498
```

The litmus test for success in this project was twofold. First, we wanted to see if we could do better than the FIFA model at predicting the results of international matches in the set of games held out. Second, to determine whether the negative effects of tournament hosts not playing matches was nullified, we would compare the position of the French team among the three statistical models. The results of the tests were somewhat inconclusive, although there did seem to be a trend indicating that our model performed worse than the other two models. The FIFA rankings correctly predicted just over 53% of games; the normal Bradley-Terry model predicted 52% correctly (when we ran the Bradley-Terry-specific prediction, it predicted 41% of games correctly, though as explained before this number cannot be judged against the other models' numbers). However, our model predicted less than 49% of games correctly.

Our original model predicted draws differently than the eventual model. A draw was the predicted result for a match if, as described above in the **approaches** section,

$$lg(ranking_a + 5) - lg(ranking_h + 5) > 0.5$$

However, in our first iteration, we assumed a greater likelihood of draws:

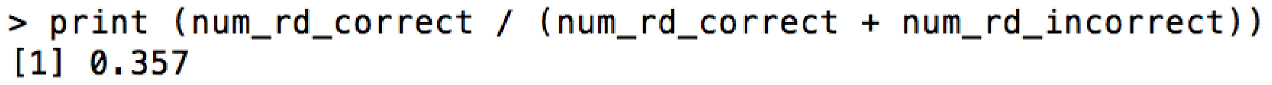
$$lg(ranking_a + 5) - lg(ranking_h + 5) > 1$$

Of note is the fact that, in the original version, the results of each of the models were much lower than the final results - around 35% - but in relation to each other, they were almost exactly the same as the final results.

Conclusions:

There are several conclusions that we are able to draw from these results. First, we are able to see, conclusively, that our model was ***not*** a better predictor of international team performance than the much maligned FIFA rankings. Second, and perhaps of greater import, it does seem as though no ranking system is likely to be a more statistically more successful metric for team performance than the FIFA ranking. Some rankings certainly might look better on paper, and this might indicate that they ***are*** better. However, in terms of their predictive abilities, team rankings are unlikely to ever provide particularly good insights.

After running this experiment, we decided to create one final model. Looking to see how much of a factor random chance was in predicting games correctly, we took the ranking that we had arrived at after training on the 9,000 matches, randomized it completely, and ran the test on it. The result is below.



```
> print (num_rd_correct / (num_rd_correct + num_rd_incorrect))
[1] 0.357
```

Luckily for us and for FIFA, the random ranking correctly predicted a much lower percentage of games than any of our carefully constructed models. However, at almost 36%, this is still a fairly significant number. It is to be expected, of course; with three different outcomes possible, a good prediction function (where all three outcomes are equally likely) can expect to get 33% of results correct.

Our conclusion was that creating a rating that has predictive as well as descriptive qualities is deceptively difficult. The Bradley-Terry model, which does not take into account any football-specific variables, performed better than our model and at a similar level to the FIFA model. With more time and some more robust statistical mechanisms it is possible we could have improved our numbers. However, even then, probably not by too much.

Limitations / Future Work:

While the strength of the models created in this project did not perform as well as we would have liked them to, we believe this is due to a number of limitations. Several of these limitations were imposed on us by the data itself. First and foremost, the data failed to relate the type of each match. As suggested earlier, tournament qualifiers and tournament games should receive a much heavier weight than friendly matches due to the nature of the competitiveness of the games. Without this information, our models treated every game equally in terms of their importance. The other major limitation imposed by the data was the home and away data. While the dataset showed which team was labeled “home” and which team was labeled “away”, it failed to show when a game was played on a neutral field. Without this information, our Bradley-Terry model assumed that every home team was truly home and every away team was truly away. This caused our model to give a home field advantage to the home team even in the games they did not play at home.

The data was not completely to blame for the limitations however. There were several other limitations that we posed on the project ourselves. One of these limitations came in the form of our method for predicting draws based on ranking. As described in the “approaches/methods” section, the way our models decided to predict draws based off of rank was fairly arbitrary. The equation for determining draws was developed through reasoning and common soccer sense, not rigorous mathematical methods.A second limitation was our decision to compare the models’ abilities to make accurate match predictions as a metric for evaluating their efficacy. We decided to use this single metric to determine the quality of each model when no such single metric exists. Instead, the evaluation of each ranking system should be a holistic evaluation that considers fairness, accuracy, and absence of edge cases. Match prediction is only an estimate of these qualities.

Looking to improve our process for the future, we could modify our model to make it a “live” model. The internals would remain the same, but after each prediction it made, it could include that match’s data and update the ladder as it went. This would ensure that the model would grow as the squads, and thus their strength, changed over time. The model would then make predictions with up to date information instead of trying to look 1000 games into the future with an outdated ranking. Another potential area for experimentation is data size. Instead of training on 10,000 games we could instead train on smaller sets of games. This would prevent historically strong teams from being overrated, and historically weak teams from being underrated.