The Ball is Round

Do Kester

SRON Netherlands Institute for Space Research,
Landleven 12, 9747 AD Groningen.
email do@sron.nl

Abstract. The qualification matches for the European Championship of 2008 and the World Championship of 2010 for national football[1] teams are analysed using a number of different models. Friendly matches between national teams in the same period were added to provide connectivity between the qualification pools. That model for which the highest evidence is obtained is used to predict the outcome of the championship.

As the division into groups at the tournament is always heavily debated, it is also investigated whether and how this draw into groups influences the championship.

The World Championship will be taking place at the time this poster is presented so that some of the predictive powers of the models can be judged immediately.

Keywords: Model selection, Nested Sampling

DATA

Before entering a tournament like the World Championship or the European Championship, a national football team has to play qualification matches. The organization of the qualification matches depends on which regional grouping the country is in: Europe, Africa, Asia, North America, South America and Oceania. In Europe all national teams are distributed over 7 qualification groups. Each team in each group plays against every other team in the group twice, once at home and once abroad. In South America there is a single group where all countries compete. North America and Asia has prequalification knock-out matches, followed by a group phase. etc. The host countries do not play qualification matches; they get into the tournament automatically.

As well as qualification matches there are also so-called friendly matches played during the same period. They are used to test various formations of the teams, to keep up strength and to maintain match routine. Unlike the qualification matches, these friendly matches do not really count; they have no consequences on the participation in the tournament. But they *do* provide connectivity between the qualification groups; it is the only way to gauge the relative strength between the groups.

In Table 1 we list a few key numbers about the tournaments we analysed. The data obtained from the various sources are the results of the matches i.e. the number of goals each team had scored at the end of the match. No other data were used in this paper.

[1] This is about what most people call “football”, except in the USA and Australia where it is called “soccer”. We will use football throughout this paper.

TABLE 1. Qualification (Q) and friendly (F) matches

tournament	host	nr of teams		matches	source
EC 2008	Austria & Switzerland	52	Q	305	wikipedia
			F	146	www.fcupdate.nl
WC 2010	South Africa	145	Q	735	www.fifa.com
			F	559	www.fcupdate.nl

MODELS

Several models were designed to analyse the data, from a very simple one in which each team gets a number of shots at its opponent's goal, to more elaborate ones including defense, home advantage, midfield and/or strategy. In Table 2 the terms are further explained, along with their priors, ranges and default values. The default values are chosen such that algorithmically the particular item seems to be absent.

The basic formulae are

$$g_1 = a_1(1-d_2) \qquad g_2 = a_2(1-d_1) \tag{1}$$

where the subscript refers one of the two teams in a match. g_1 is the predicted number of goals for team 1, a_1 is its attack strength and d_2 is the opponents defense success rate.

Other parameters modify the defense and/or the attack, so that both a and d are some function of the other parameters, h, m, s. After applying this function Equation 1 is used. The modification caused by the home advantage, h, looks like

$$a = a \times h \qquad d = d^{1/h^2} \tag{2}$$

The ratio between the midfields modifies the attack and defense of both teams in a similar way, proportional to their strategy. All modifications are such that the defense remains within its required range of $0 < d < 1$.

The predicted number of goals is a real number; the data, the actual scores, are of course always integral.

From the set of models defined in the previous paragraphs we shall select the best one, having the highest Bayesian evidence. As usual we write Bayes' rule:

$$\begin{aligned} \Pr(\theta|M) \times \Pr(D|\theta M) &= \Pr(D|M) \times \Pr(\theta|DM) \\ \text{prior} \times \text{likelihood} &= \text{evidence} \times \text{posterior} \end{aligned} \tag{3}$$

TABLE 2. Various models

name	explication	prior	range	default
attack	number of attacks a team launches	Jeffreys	[1, 20]	n/a
defense	fraction of attacks which is countered	uniform	[0, 1]	0
home	advantage of playing at home	uniform	[0.5, 1.5]	1
midfield	overall strength of the team	uniform	[0.5, 1.5]	1
strategy	from defensive to aggressive	uniform	[0, 1]	0.5

where M is a model as defined above, θ stands for the parameters and D is the data, the number of goals scored per team and per match. For the likelihood we have to use here a Poisson distribution; football typically produces just a few goals per match. The logarithm of the likelihood reads

$$\log L = \sum n \log g - g - \log(n!) \tag{4}$$

where n is the number of goals of one team in one match. The summation is over the scores of both teams in all matches. The priors for the parameters are listed in Table 2.

NESTED SAMPLING

The algorithm called Nested Sampling [1, 2] is set up to calculate the evidence directly, incidentally also yielding the (average) model parameters. It starts with an ensemble of N (100) multidimensional points, θ, chosen randomly from the distribution on the prior domain of the parameters. For each of the N points the likelihood is calculated. An iterative cycle is started in which the point with the lowest likelihood is replaced by a copy of one of the others. The parameters of the copy are randomly moved around over the prior distribution; provided that the new likelihood remains larger than the original one. A new (random) point is found which is higher in likelihood. This way we slowly climb the likelihood mountain. The evidence follows as an integral of the likelihood mountain, while the optimal parameters are weighted sums of all the points.

In this description of Nested Sampling, the most important part is the "randomization of the new point". Special care needs to be taken to get it right or the fit will be suboptimal and the evidence will (most likely) be too low. We use three engines to move the points around.

StepEngine: One by one and in random order the parameters are changed. Each time, the likelihood is calculated to see if the step can be accepted.

CrossEngine: Select from the ensemble another point at random. Take the parameters at random from one or the other. Calculate the likelihood to see if the new point is above the low likelihood limit.

FrogEngine: Select a few (up to 5) other points from the ensemble and calculate their average. Move at random along the line defined by the original point and the average. Calculate the likelihood. This engine will move the point efficiently in problems where the parameters are (highly) correlated.

The first engine works very locally, on one parameter at a time. The other two change more parameters or even all simultaneously. It is absolutely necessary to have global engines too, to move out of awkward regions of the likelihood landscape. At each step in the Nested Sampling cycle these engines are called three times in a random order.

RESULTS

We calculated the evidences for a selection of models using 4 different datasets: qualification matches alone, and qualification and friendly matches, for the EC 2008 and for

TABLE 3. Log Evidence results for various models

attack	defense	midfield	strategy	home	EC-q	EC-q&f	WC-q	WC-q&f
x	x	x	x	x	-824.1	-1122.1	-1787.9	-3285.0
x	x	x	0.5	x	-838.1	-1143.2	-1830.0	-3351.7
x	x	1.0	0.5	x	-837.4	-1142.2	-1830.4	-3358.1
x	x	x	x	1.0	-850.2	-1154.6	-1888.5	-3393.6
x	x	x	0.5	1.0	-855.5	-1164.3	-1913.2	-3447.7
x	x	1.0	0.5	1.0	-857.2	-1166.0	-1915.7	-3454.0
x	0.0	1.0	0.5	1.0	-1028.6	-1369.8	-2233.8	-3862.1
x	0.0	1.0	0.5	x	-1036.6	-1374.3	-2260.5	-3879.9

the WC 2010. The results are listed in Table 3. Where the table has a parameter entry "x", the model fit included that parameter. Otherwise the parameter was fixed at the (default) value listed.

In all cases the evidence is highest for the most complex model, containing all 5 parameters. This model will be used in the remaining analysis. Evidence values across columns cannot be compared as the calculations involve different datasets.

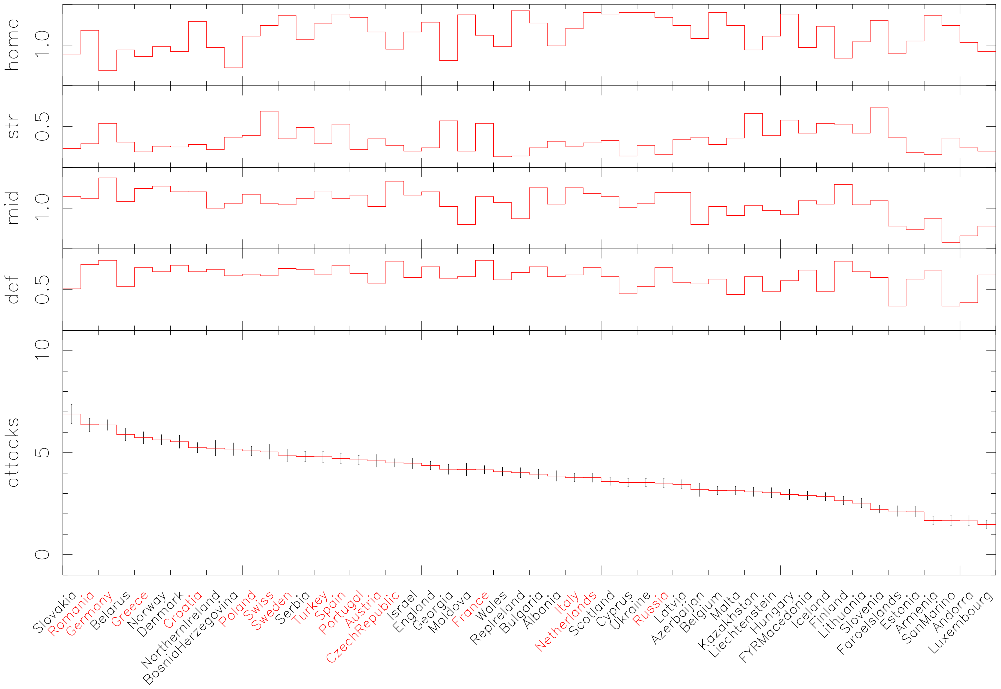

FIGURE 1. Parameters for all national teams, using qualification matches only

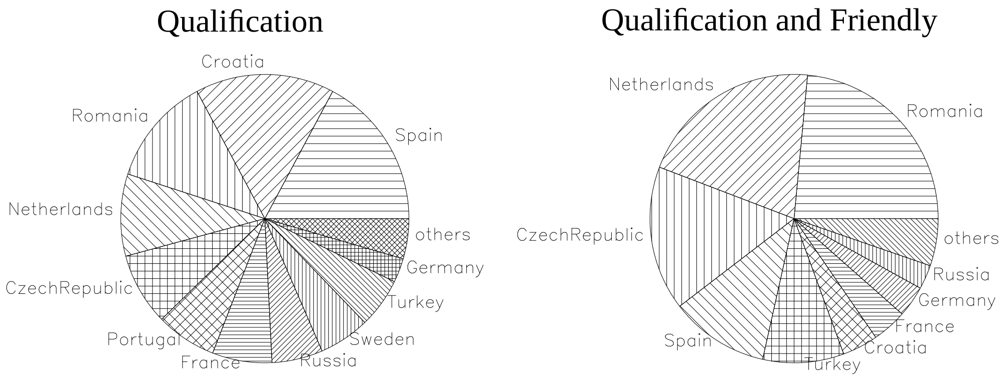

FIGURE 2. Chances to win

European Championship 2008

In Figure 1 we present the parameters for all national teams participating in the qualification rounds for the EC 2008, ordered according to their attack strength. At the bottom of the figure the names of the teams are listed; in red the teams which actually made it into the tournament. Obviously a good attack is not the only thing needed to get into the tournament. In fact there is no simple ordering that would separate those teams which made it into the tournament from the others. We also see that most teams prefer a defensive strategy to a aggressive one (str < 0.5). The home advantage for most teams is larger than 1, meaning that they play better at home than abroad. There are, however, a few teams where it is smaller than one. We leave it to the reader to decide what this means.

The tournament begins with a group phase in which the teams are divided into four groups, more or less at random. Each team plays the other members of its group once. The two best teams from each group enter a knock-out phase, in which only the winner of the game progresses.

Given the results of an analysis, we can run the tournament a large number of times (1000), calculating the predicted number of goals for each match and drawing a random number from the relevant Poisson distribution. Following this to the end, through the group phase and the knock-out phase, we can predict what the chances are for each team to win.

During these runs only the hosts (Switzerland and Austria) play all their matches at home; they always have a home advantage. For all other teams the home advantage has been disabled.

In Figure 2 we display the two sets of probabilities for national teams to win the EC 2008, one using the qualification matches only, the other using also friendly matches. The fact that Spain did actually win the European Championship for 2008, shows clearly in the qualification matches only. Even though it is only a 17 % chance, it is still the largest. Addition of the friendly matches into the dataset, reduces the chances of Spain by 6 %, overtaken by three other teams. Surprisingly, the lack of connectivity between the qualification matches has little influence on the final result. Perhaps the fact that the qualification matches are really important is more significant than the connectivity

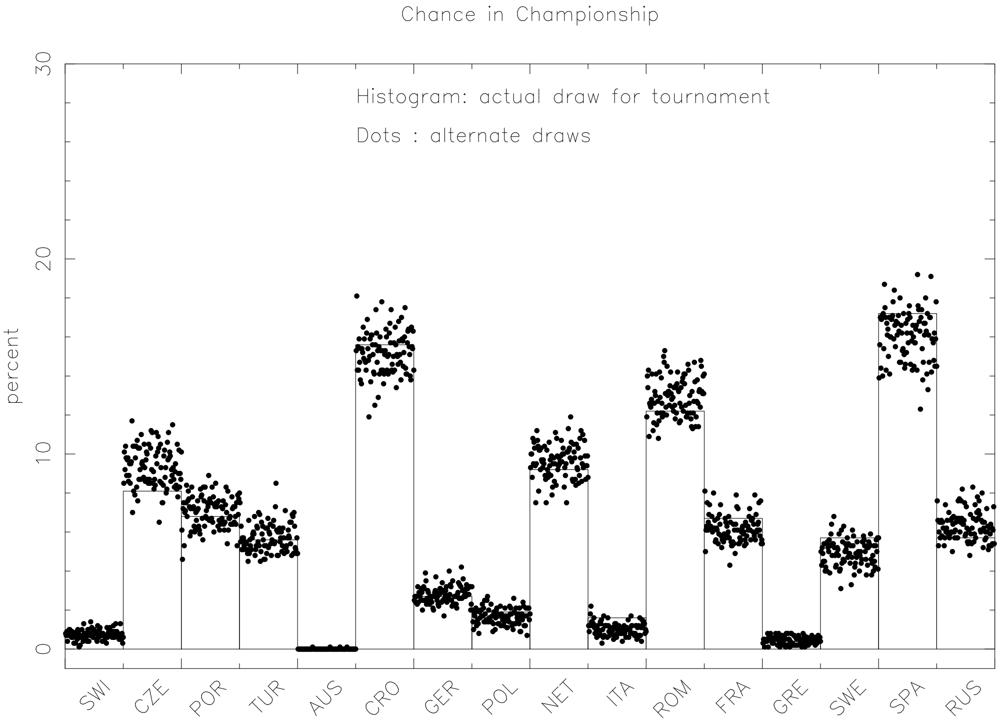

FIGURE 3. Finalists in alternate draws of the tournament groupings.

provided by the friendly matches. Or the quality of the groups is sufficiently close for the lack of connectivity not to matter. The runner-up, Germany, and both other semi-finalists, Turkey and Russia, do not figure clearly in either of the two pie-charts.

Alternate Tournaments

The division into tournament groups is more or less random. Based on certain criteria the teams are assigned a ranking from one to four. Every group randomly obtains a team of each ranking.

The grouping always attracts comments. In the 2008 tournament the third group, including the Netherlands, Italy, Romania and France, was colloquially known as the “group of death”. All teams were considered “strong”, but only two of them could survive the group phase.

Would another draw into groups have a large influence on the outcome of the tournament. Figure 3 shows the influence of other drawings in the groups on the chances of the teams to win. The histogram shows the chances of the national teams with the groups as they were actually played. The dots show the chances for 100 alternate draws from the same rankings. The uncertainty in the histogram is about 3 %, so that the alternate tournament configurations scarsely influences the chances of the teams to win. Some influence remains but not much. The best teams still have the best chance to win.

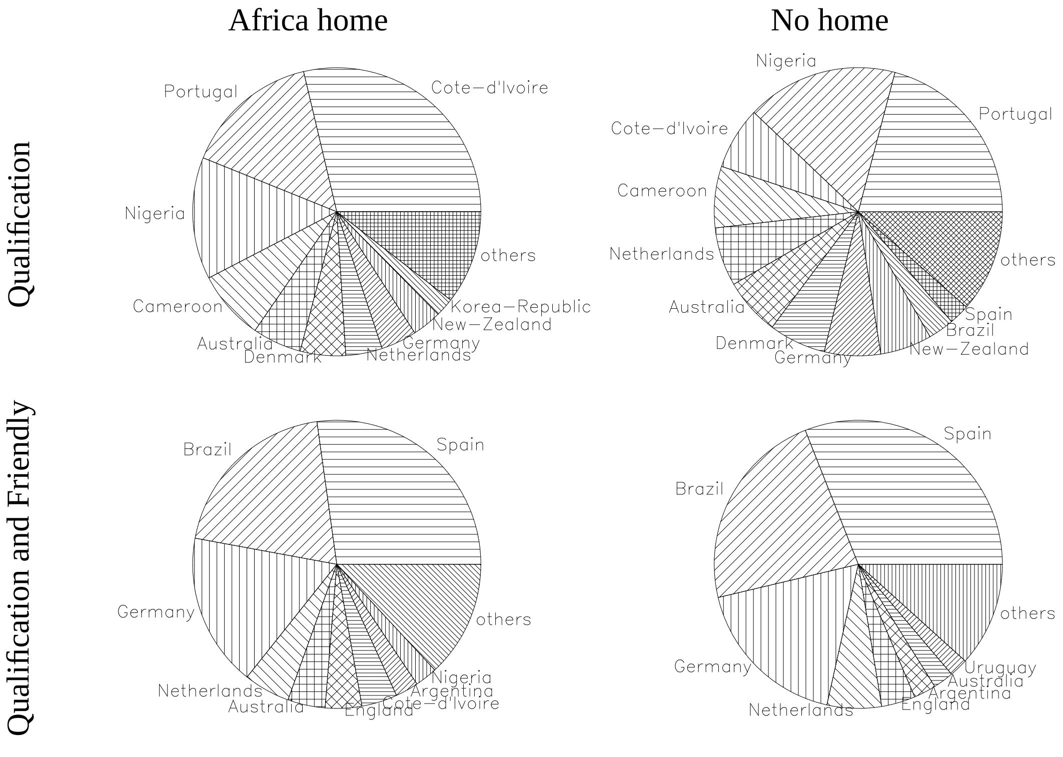

FIGURE 4. Chances to win

World Championship 2010

The data used for the World Championship is listed in Table 1. The strengths of the national teams were calculated in the same manner as for the European Championship of 2008. The results for all 145 national teams are not shown because they will not fit into a legible figure.

In Figure 4 we present predictions for four situations. On the one hand we have the use of qualification matches only, or using also additional friendly matches. We further observed that all African participants have a quite substantial home advantage. As they might feel more “at home” in South Africa than the other, non-African teams, we also calculated the case in which African teams keep half of their home advantage (Africa home). In the “no home” case *all* teams lose their home advantage, except of course South Africa itself.

The top panels in Figure 4 shows that there is a fair chance that the cup will stay on the African continent, even more so when we accept the African home advantage. In the latter case it is over 50 %. However this finding could very well be due to the lack of connectivity between the qualification groups. If we add the friendly matches into the fray, the usual suspects have the largest chances to win: Spain, Brazil, Germany etc., independent of any African home advantage.

For the lower right situation of Figure 4 a full prediction is shown for the results of each team. In Figure 5 the teams are listed from top to bottom. Each four teams in this list form a group at the tournament. On the horizontal axis it is gray-coded what are the

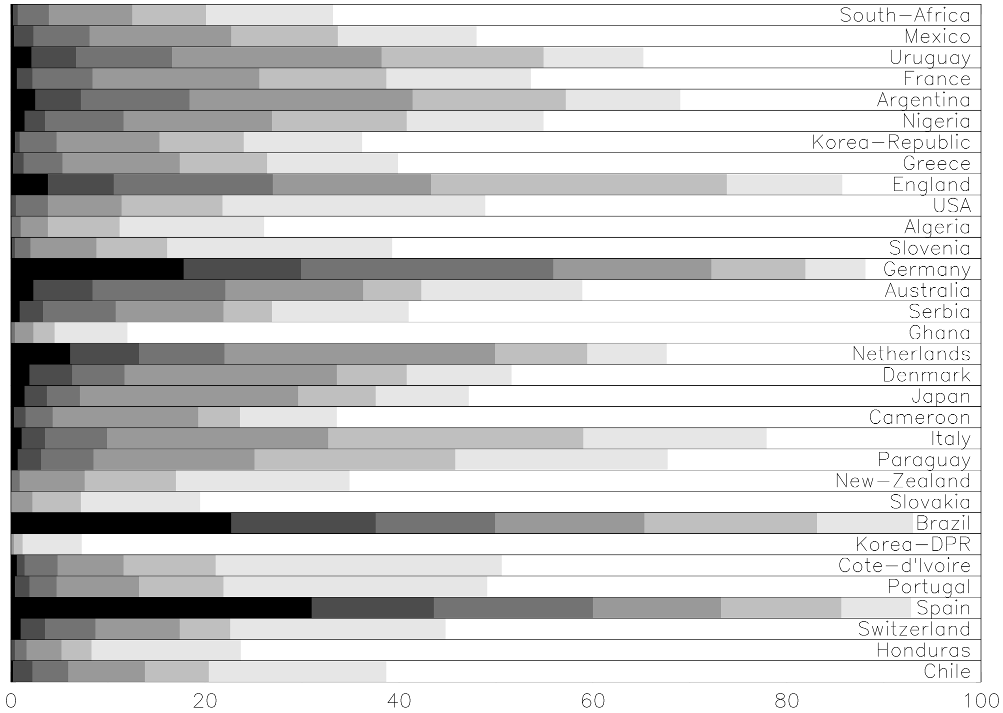

FIGURE 5. Predicted results for all participants.

chances in % for reaching a certain level. It starts with black, indicating the chance to win the cup; then in succesively lighter gray tones, the chances for reaching resp. the finals, the semi-finals and the quarter finals. The lightest two shades of gray indicate the chances of becoming first resp second in the group phase. It is clear that the first teams in each group are most likely to survive the group phase, except for South Africa itself. As a host South Africa is selected as a group leader, where as the other group leaders are selected as the strongest teams. This is all well known and it shows.

CONCLUSIONS

1. The outcome of a tournament does not depend much on the selection into groups. Independent of the grouping the same teams have the best chances to win.
2. The WC 2010 will most likely be won by one of Spain, Brazil or Germany. Unsurprisingly.
3. The ball is round.

REFERENCES

1. J. Skilling, “Nested Sampling for General Bayesian Computation,” in *Bayesian Analysis 4*, 2006, pp 833-860.
2. D.S. Sivia, and J. Skilling, “Data Analysis, A Bayesian Tutorial, 2nd Edition,” Oxford University Press, 2006.