

Time-varying ratings for international football teams

Rose D. Baker¹ and Ian G. McHale^{*2}

¹Centre for Sports Business, Salford Business School, University of Salford, UK.

²Centre for Sports Business, Management School, University of Liverpool, UK.

November 17, 2017

Abstract

We present a model for rating international football teams. Using data from 1944 to 2016, we ask ‘which was the greatest team?’. To answer the question requires some sophisticated methodology. Specifically, we have used k-fold cross-validation, which allows us to optimally down-weight the results of friendly matches in explaining World Cup results. In addition to the central aim of the paper, we also discuss, from a philosophical perspective, situations in which model over-fitting is perhaps desirable. Results suggest that Hungary in 1952, is a strong candidate for the all-time greatest international football team.

Keywords: sports, cross validation; empirical Bayes; ranking; rating; soccer; shrinkage.

1 Introduction

Football is the world’s most popular sport¹, and its flagship competition is the FIFA World Cup. Although international football dates back to 1888 when England played Scotland in the first international match, the World Cup started in 1930 and has been held every four years since, barring breaks during the Second World War. Nowadays over 200 countries play in qualifying tournaments to make it through to the World Cup finals, hoping to be crowned world champions. Most fans have memories of watching great players and teams at World Cups, and recalling great sides like England 1966, Brazil 1970 and Netherlands 1974 may make many of us feel nostalgic.

Although each World Cup competition (and other continental competitions like the UEFA European Championship) offers a candidate for the best team in the world for that year, the passing of time means that we cannot *know* the answer to a question that most football fans find themselves asking at some point: ‘which is the best team of all time?’. This is the main question we attempt to answer in this paper.

The problem of estimating how strong a team was at a time-point with hindsight is different from the problem of forecasting results, as we can also use data on matches played after the match in question. For example, to estimate how good a team was around the year 1990, we

*email address: ian.mchale@liverpool.ac.uk

¹<http://www.britannica.com/EBchecked/topic/550852/football>

should use results from 1989 and 1991 to get the best estimate of the team’s strength. Using forecasting techniques, one can only use data from 1989 in this example, and the estimate of how good the team was near 1990 would be less accurate.

To address the ratings problem here, we use as a starting point the time-varying ratings model presented in Baker and McHale (2014) which was used to rank football teams in domestic football in England. However, a key difference between international football and domestic football lies in the number of games played by teams. A domestic team will play in the region of 60 games per year whilst an international team will play around eight, with the number even smaller in earlier years. A consequence of this is that although the Baker and McHale model can be used to estimate each international team’s ‘ability’ at a specific moment in time, the standard errors on the estimated abilities are often large and can vary widely from one team to another. As such, it is sometimes the case that two competitors will have similar estimated strengths but differing standard errors on the estimated strengths. This may be caused by the competitors playing different numbers of games, or may be a consequence of higher volatility in results. Nevertheless, the result that two teams are rated as similar when in fact one has a larger standard error on the rating than the other is somewhat unsatisfactory.

This is a general problem in ranking, not specific to estimating ratings models in football. In women’s tennis for example, Baker and McHale (2017) present an empirical Bayes procedure for identifying the all-time greatest player; more variable results were shrunk further towards the mean. But the problem is not confined to sport. One can imagine new web pages achieving a high rating with corresponding high standard error, or competing products achieving similar ratings despite one product having been rated highly in only a small number of trials. In this paper we present a solution to such difficulties which can be adopted in rankings problems in general. We demonstrate the methodology in a sporting context: an empirical Bayes time-varying ratings model for football teams. We also present a solution to the resulting computational problems which gives fast and efficient estimation of auxiliary model parameters and of the ratings themselves.

The paper is organised as follows. Section 2 presents the model we use for estimating team strengths in football, and the empirical Bayes extension. In section 3 we present our method for obtaining rankings from the model. In section 4 we describe the data. Results of model fitting are given in section 5 with some additional findings discussed in section 6. The paper concludes with our closing remarks in section 7.

2 Time-varying model for estimating strengths of football teams

The model we use as the basis for our ratings is a slightly modified form of the model first presented in Maher (1982). The numbers of goals scored by the two teams in any particular match are assumed to be independent Poisson random variables, whose means are determined by the attack and defensive abilities of each side. Specifically, the interaction between the two teams’ abilities is specified so that in match j , played at time t , between teams s_{j1} and s_{j2} , the

number of goals scored are g_{j1} and g_{j2} respectively, and are given by

$$g_{j1} \sim \text{Poisson}(\alpha_{s_{j1}}(t)/\beta_{s_{j2}}(t)) \text{ and } g_{j2} \sim \text{Poisson}(\alpha_{s_{j2}}(t)/\beta_{s_{j1}}(t)), \quad (1)$$

where $\alpha_{s_{ji}}(t)$ is a measure of team s_{ji} 's attack ability ($i = 1, 2$) at the time of the j th match and $\beta_{s_{ji}}(t)$ is a measure of its defensive ability at the time of the j th match, and $\alpha, \beta > 0$.

The (log) likelihood function is the basis for inference about the model parameters, hence the task is to compute this, given the model parameters. Let $a_{s_{j1}}(t) = \ln(\alpha_{s_{j1}}(t))$ and $b_{s_{j1}}(t) = \ln(\beta_{s_{j1}}(t))$, then, discarding a constant, the log-likelihood for the j th match is

$$\ell_{0j} = g_{j1}(a_{s_{j1}}(t) - b_{s_{j2}}(t)) + g_{j2}(a_{s_{j2}}(t) - b_{s_{j1}}(t)) - \exp(a_{s_{j1}}(t) - b_{s_{j2}}(t)) - \exp(a_{s_{j2}}(t) - b_{s_{j1}}(t)), \quad (2)$$

and the total log-likelihood is $\ell_0 = \sum_{j=1}^m \ell_{0j}$, where m is the total number of matches played.

This specification does not allow for any 'home advantage' effect. Unlike for domestic league football, an international team can play a match at home, away from home, or at a neutral venue. To allow for this in the model we experimented with letting $a \rightarrow a + \xi + \gamma$ for a home match, $a \rightarrow a + \xi - \kappa$ for an away match, and $a \rightarrow a + \xi$ for a neutral match. Here γ is the home advantage, κ the away disadvantage, and ξ is a parameter to allow for the downward trend in number of goals scored with time. However, when we estimated γ and κ they were not statistically significantly different in magnitude. Thus, in our final specification, we use just one parameter, so that neutral matches can be interpreted as being half a home match and half an away match.

To allow for any dependence between the number of goals scored by the two teams, we follow a similar approach to that adopted in Baker and McHale (2014) and inflate/deflate the probabilities of certain scores occurring. Letting \mathbf{P} be a matrix of score probabilities with elements p_{uv} for final game scores u and v for each team, we use inflation parameters so that draw probabilities for u 'goals all' were scaled up by a factor $1 + c_u$, for $u = 0, 1, 2$. If p_{uv} is the uninflated Poisson model probability of home and away teams scoring u, v goals respectively (equation 2), p_{uu} becomes $(1 + c_u)p_{uu}/(1 + \sum_{k=0}^2 c_k p_{kk})$ for $u \leq 2$, otherwise p_{uv} becomes $p_{uv}/(1 + \sum_{k=0}^2 c_k p_{kk})$. The resulting matrix of diagonally inflated score probabilities still sums to one. The log-likelihood and its derivatives change accordingly, and the parameters c_u can be estimated along with the other parameters of the model. There was no improvement to the model in terms of log-likelihood from adding a scaling parameter for 3 goals and so we stopped at $u = 2$. The estimated values of c_0 , c_1 , and c_2 were 0.253, 0.089 and 0.142 respectively. The predicted off-diagonal probabilities could also have been adjusted, but our experiments with this did not improve the model fit and was omitted.

For inference, some measure of error on fitted model parameters, or on quantities derived from them, is required. The approach adopted was to bootstrap the dataset, and to refit the model to the bootstrapped data. Doing this 100 times was computationally feasible, and enabled standard errors to be found on the 'points' awarded to teams when ranking them. The bootstrap method used was the 'parametric bootstrap' (e.g. Hastie *et al*, 2009, p.264). Here

fitted strengths were used to randomly generate match results from the inflated Poisson model.

2.1 Time-varying team strengths

Baker and McHale (2014) present a method to estimate the time-varying strengths in equation 2. Their approach is to use barycentric rational interpolation to interpolate between strengths estimated at nodes positioned at evenly spaced intervals over the period of a team’s existence. An alternative method to barycentric rational interpolation would be to use splines. However, barycentric interpolation has several advantages over splines. Barycentric curves are simpler than splines to compute, and are infinitely differentiable, meaning that finding extrema of a fitted curve is much easier than for splines. The approximation error is also often smaller. The interested reader can refer to Baker and Jackson (2014) for a general account of barycentric rational interpolation in Statistics, and to Baker and McHale (2014) for a more detailed discussion of barycentric rational interpolation in sport.

The logarithm of the time-varying strength is modelled using the barycentric rational interpolant so that the logged attack strength of team s at time t is given by

$$a_s(t) = \frac{\sum_{k=1}^{n_s} w_{sk} A_{sk} / (t - t_{sk})}{\sum_{k=1}^{n_s} w_{sk} / (t - t_{sk})} \quad (3)$$

where A_{sk} is the k th tabulated log-attack strength of team s , i.e. the log-strength at time t_{sk} (the k th node). Similarly, B_{sk} denotes the k th tabulated log-defence strength of team s and is used to calculate the interpolated logarithm of the time-varying defensive strength, b_s , in the way described in equation (3). The A_{sk} and B_{sk} parameters are estimated and used to calculate the log-strengths at any moment in time using barycentric interpolation. There are n_s tabulated strengths for team s and the choice of the number of nodes and node position, which is analogous to the placement of knots in spline interpolation is discussed below. The parameters w_k are weights which control the amount of curvature of the interpolation. As found in Baker and McHale (2014) using weights given by $w_k = (-1)^k$ proved to give the best results in terms of model fit as measured by the cross validation procedure described below.

2.2 Empirical Bayes model extension

The model described above can be used to estimate time-varying team strengths so that, for example, the Brazil team of 1970 can be compared with the England team of 1966. However, teams that have a relatively short run of good results can achieve an unrealistically high estimated strength. For example, when this model is fitted to the data on results of international football matches, the Costa Rica team in the early 1950s is rated in the top 3 teams in history! Any football fan knows that this is a spurious result. It seems to occur because around that time, the Costa Rica team won almost all of its matches by large margins against lowly opposition. Notably, although the point estimate of strength is very high, it has an equally high standard error. This is a general problem in sports rankings: that teams, or competitors, without a great reputation (deservedly) can appear near the top of the ranking, based on a few

good results.

A solution here is to assume that the strength parameters are drawn from some prior distribution. Indeed, previous authors have adopted this approach. Graves et al. (2003), for example, assume that racing drivers' abilities come from a common distribution when ranking drivers over a season of races. Baker and McHale (2017) use a prior when rating women's tennis players. Here we use a similar approach, but one designed specifically for football. Another difference from previous work is the use of k-fold cross-validation to estimate some model parameters.

The benefits of using the prior are that the estimated strengths are shrunk towards a grand mean, and the amount of shrinkage depends on the evidence (or lack of it) of being different from average. As a result, teams winning all of a small number of matches will have their estimated strengths shrunk back drastically towards the grand mean because there is not as much evidence of merit as for a team that has a lower win rate over a larger number of games.

Let there be N teams, with team s having n_s tabulated (logged) attack strengths A_{ks} and corresponding defence strengths B_{ks} . There are $M = \sum_{s=1}^N n_s$ strength pairs. Our initial thought was to put priors on A_{ks} and B_{ks} directly. However, we obtained better results when putting a prior distribution on the total strength $x_{ks} = A_{ks} + B_{ks}$ and the attack-defence strength difference $y_{ks} = A_{ks} - B_{ks}$, and assuming these to be uncorrelated. One can think of x_{ks} as a measure of how good the team is overall, whilst y_{ks} can be thought of as a measure of how attacking the team is. For a fixed x_{ks} , a larger y_{ks} suggests the team is exerting more effort in attack and less in defence. One can think of x and y as the two principal components of strength. Analysis bore out the wisdom of this intuitive modelling step, because y shrank much more than x .

We assume that $x_{ks} \sim N(\mu_x, \phi^2)$ and $y_{ks} \sim N(\mu_y, \psi^2)$. Writing $\ell_0 = \ell_0(\mathbf{x}, \mathbf{y})$, we maximise

$$\ell(\mathbf{x}, \mathbf{y}) = \ell_0(\mathbf{x}, \mathbf{y}) - (1/2) \sum_{s=1}^N \sum_{k=1}^{n_s} \{(x_{ks} - \mu_x)^2 / \phi^2 + (y_{ks} - \mu_y)^2 / \psi^2\},$$

i.e. we multiply the likelihood by the pdf that the strengths come from a common distribution, so giving the log-likelihood a shrinkage term.

To estimate the strengths and the home advantage parameters and trend ξ we maximise ℓ . This is MAP (maximum a posteriori) estimation, because ℓ is the logarithm of the posterior probability, shorn of some factors that are held constant, and in MAP estimation we find the mode of ℓ instead of just maximising the likelihood function ℓ_0 . Computational methods of maximising log-likelihoods like ℓ are mentioned in Baker and McHale (2017). In outline, first and (diagonal) second derivatives are computed analytically, and a robustified Newton-Raphson iteration used, ignoring the off-diagonal terms of the Hessian matrix. Because of the crude approximations used, the second-order convergence of the Newton-Raphson method is lost, but the method can still maximise the likelihood for thousands of parameters quickly.

2.3 Estimating ϕ, ψ and M using cross-validation

The strength parameters can be estimated using the MAP estimation described above. However, the estimation of the strength shrinkage parameters ϕ, ψ and the total number of nodes $M = \sum_{s=1}^N n_s$ is more difficult. For example, to estimate ϕ and ψ we should use the full prior pdf, without omitting any terms except constants, to obtain

$$\ell(\mathbf{x}, \mathbf{y}) = \ell_0(\mathbf{x}, \mathbf{y}) - (1/2) \sum_{s=1}^N \sum_{k=1}^{n_s} \{(x_{ks} - \mu_x)^2 / \phi^2 + (y_{ks} - \mu_y)^2 / \psi^2\} - (1/2) \left(\sum_{s=1}^N n_s \right) \ln(\phi^2 \psi^2).$$

Here μ_x, μ_y can be trivially estimated and the strength factors x_{ks} and y_{ks} are estimated by the MAP method. The method of node placement is specified, so the only parameters remaining to be estimated are ϕ, ψ and M . We can see that ℓ is maximised when $\phi = \psi = 0$ and all strengths are equal to the mean strength. Hence maximising ℓ will not enable us to estimate ϕ and ψ sensibly.

One could solve this problem by integrating over the prior distribution to obtain the posterior probability, rather than just finding the posterior mode, and this type of Bayesian analysis can yield estimates of ϕ, ψ and M that are ‘sensible’. However, doing this requires further modelling complexity and approximations, particularly of the integrals required, and when we try to estimate M as well, the complexity increases further. The main problem here is in the further modelling required. Rather than engaging with very complicated methods that are in any case only approximate, we cut the Gordian knot by using the frequentist method of k-fold cross-validation.

The methodology of cross-validation is most commonly used to assess the performance of estimation methods, and its essence is that a model used to predict some observations is not fitted to those observations, so that an unbiased estimate of the probability of correct prediction is obtained. Often the problem is a classification one e.g. of whether a creditor will default on a loan, as in Bastos (2010). With an unbiased estimate of model performance on fresh data, model parameters such as number of regressors can be adjusted to give better performance, and so clearly cross-validation can also be used to estimate model parameters, and not just to evaluate model performance. Sometimes, as in Wang and Zidek (2006), this estimation through cross-validation can be done analytically, but usually it must be done numerically. The method of k-fold cross-validation works well and does not have the major computing problem of the leave-one-out method of cross-validation (e.g. Hastie *et al* (2009), p243), where the likelihood for a match would be computed using data from all other matches. This requires a separate parameter estimation for each match.

To carry out a k-fold cross-validation, one divides the tournament matches into ‘folds’, here by taking the earliest tournament match as belonging to fold 1, and assigning subsequent tournament matches to folds 2...10, then restarting at fold 1. We hold out the tournament matches only, as these are the matches we would most like the estimated strengths to reflect. All the data except for fold n matches are fitted, and the log-likelihood for the n th fold of matches is computed. The procedure is repeated for all folds and the tournament match log-likelihood

for all k folds is computed. One ends up with the log-likelihood for all cup matches, where the likelihood for each cup match has been computed using only data that excluded that match. We call this the ‘out-of-sample log-likelihood’. Here we use 10 folds, which is commonly done. This is computationally feasible, and increasing the number of folds beyond 10 hardly changed the results; already with 10-fold cross validation, 90% of available data is used in computing each contribution to the ‘out-of-sample log-likelihood’. Usually, 5 or 10-fold cross validation is recommended (Hastie *et al*, 2009, p. 243). To estimate ϕ , ψ and M , one maximises this out-of-sample log-likelihood for these parameters.

2.4 Friendly matches vs tournament play: down-weighting different types of matches

In the literature on forecasting the results of international football matches, there is some debate as to the relevance of friendly matches. This is because in such matches teams tend to field weakened sides and experiment with new players. In tournament play however, the time for experimenting should be over and winning is paramount. As such, teams consist of the strongest set of players available. For our purposes, in identifying the greatest teams of all time, it seems unsuitable to assign the same level of importance to friendly matches as to matches in the World Cup Finals. Our solution to this conundrum is to scale the log-likelihood for friendly matches by some factor $\delta < 1$. This means that the likelihood of a friendly match is raised to the power δ and so each friendly match counts as a proportion δ of a tournament match.

One could down-weight the matches by an arbitrary amount but this would be unsatisfactory. Instead, we again resort to cross validation to estimate δ from the data. The out-of-sample log-likelihood will be largest with the optimal weight (value of δ). This use of ‘relevance weighted likelihood’ and the estimation of parameters by cross-validation is described in Wang and Zidek (2006).

Results are that both friendly matches and tournament qualifying matches should have weight 0.45 ± 0.05 . This gives a modest improvement of 8.7 in the out-of-sample likelihood for cup matches.

The ‘out-of-sample log-likelihood’ for cup matches is an excellent statistic for examining the practical usefulness of all aspects of the modelling, for two reasons. It is a statistic computed on ‘fresh’ data, and the log-likelihood is anyway widely used as a measure of model ‘goodness’.

3 Obtaining ratings from our model

Once the model parameters have been estimated, including attack and defensive strengths, in order for the model to be used to generate a single team rating, the strengths must be combined to give one single measure of team strength, which can be used to easily compare teams. Baker and McHale (2014) used $a_s(t) + b_s(t)$ as a measure of team quality, because team 1 is more likely to win than team 2 at time t if $a_1(t) + b_1(t) > a_2(t) + b_2(t)$.

There is a problem here in that over a period of time, e.g. 4 years, strength varies, so

that the average total strength over the four year period $\bar{a}_s + \bar{b}_s$ cannot be used to successfully rank teams. For example, let the log-attack strengths of two teams be a_1, a_2 and log-defence strengths be b_1, b_2 . The expected number of goals scored by team 1 under the independent Poisson model, assuming normality of the logged strengths and independence of strengths from different teams, is

$$E \exp(a_1 - b_2) = \exp(\mu_{a_1} - \mu_{b_2} + 1/2\sigma_{a_1}^2 + 1/2\sigma_{b_2}^2),$$

where μ_{a_1} and μ_{b_1} are the expected values and $\sigma_{a_1}^2, \sigma_{b_2}^2$ are variances of the attack and defence strengths. The variances arise here because of ‘transformation bias’. They reflect changes in strength over time. We also have

$$E \exp(a_2 - b_1) = \exp(\mu_{a_2} - \mu_{b_1} + 1/2\sigma_{a_2}^2 + 1/2\sigma_{b_1}^2).$$

Then for team 1 to score a greater expected number of goals, we have

$$\mu_{a_1} - \mu_{b_2} + 1/2\sigma_{a_1}^2 + 1/2\sigma_{b_2}^2 > \mu_{a_2} - \mu_{b_1} + 1/2\sigma_{a_2}^2 + 1/2\sigma_{b_1}^2.$$

This gives

$$\mu_{a_1} + \mu_{b_1} + 1/2\sigma_{a_1}^2 - 1/2\sigma_{b_1}^2 > \mu_{a_2} + \mu_{b_2} + 1/2\sigma_{a_2}^2 - 1/2\sigma_{b_2}^2,$$

so that we can take

$$\mu_{a_1} + \mu_{b_1} + 1/2\sigma_{a_1}^2 - 1/2\sigma_{b_1}^2$$

as a ranking criterion. Using the estimated means (\bar{a} and \bar{b}) and variances (s_a^2 and s_b^2) of strengths over the period of time (e.g. four years), we thus use

$$\bar{a}_1 + \bar{b}_1 + 1/2s_{a_1}^2 - 1/2s_{b_1}^2 \tag{4}$$

as our ranking criterion.

Curiously, uncertainty in offensive strength gives a better rank, while uncertainty in defensive strength gives a worse one. The normal model for x_1, x_2, y_1, y_2 already introduced assumed that x and y were uncorrelated. Imposing this condition gives $1/2\sigma_{a_1}^2 - 1/2\sigma_{b_1}^2 = \text{cov}(a_1, b_1)$, so that the measure of strength is $\bar{a}_1 + \bar{b}_1 + \text{cov}(a_1, b_1)$. This suggests that simply using $\bar{a}_s + \bar{b}_s$ as a measure of team quality over a period is not as sensible as it may seem at first; teams do better overall if their attacking and defensive strengths are in synchrony.

The criterion given in (4) could be used to rank teams, but it does not take account of another source of uncertainty here - the statistical error on the estimated strengths, which should not be ignored. As such we propose a new method that is more robust. If we imagine many games played between the two countries, with strengths drawn from two sampling distributions (with equal means but non-equal standard errors), the country with the larger standard error on strength will sometimes be much stronger than the other and sometimes much weaker. If we are looking at some observable outcome, such as number of games won minus number lost, the

average outcome will not be identical for the teams; we can use the bootstrap replicates directly to find the expected outcome. Using the ranking derived from considering such contests also allows the deviation from Poisson probabilities to be included in the ranking. It also allows the whole sampling distribution of the strengths to be used, not just the first two moments.

The ranking method we use is thus:

- find 4-year (for example) periods of maximum average strength for the teams;
- choose playing times for each team uniformly through the 4-year period. Here if the two periods were contemporaneous, matches would be played uniformly throughout the period, and hence we imagine matches played at 100 regularly spaced epochs through the periods;
- for (say) 100 simulations, allow each team to play all the others using in turn each bootstrapped set of strengths;
- for each simulated match, we award 1 point for a win, 0.5 point a draw and 0 points for a loss. We then average this quantity to give a pseudo-win probability of winning matches over a set period of time;
- rank teams by this pseudo-win probability.

4 Data

We obtained data on the results of all international football matches from 24th December 1944 to 25th May 2016. In total there are 38,047 matches. The data were scraped from the www.11v11.com website. For matches in which extra-time or penalties occurred, we use the scoreline at the end of normal play (approximately 90 minutes), and thus ignore what happened in extra-time and penalties. We think this is reasonable, as when estimating how strong teams are, if the scores are tied after 90 minutes, the two teams are likely to be relatively close in strength, and the information on which team won a subsequent penalty shoot-out is less relevant than the scoreline after 90 minutes.

Before presenting our results, given the long period of time covered, it is interesting to look at some descriptive statistics.

4.1 Evolution of playing style

A simple way to examine whether, and how, football has changed over the last 70 years is to consider the average number of goals scored per match. A simple analysis shows that this decreased substantially from 1944. It was 4.9 in 1944 and 2.6 in 2016. Figure 1 shows that average goals were as low as 2.2 in 1989, and subsequently has increased slightly, though the overall trend seems to have reached a stable value.

To examine differences in the number of goals per match across match types, we fit a regression of average goals scored with covariates: time (year), time squared, and match status

(tournament finals, tournament qualifier, minor tournament, friendly). The results show that compared to friendly matches, matches in minor tournaments have 0.36 extra goals, tournament qualifying matches have 0.10 goals more, and matches in tournament finals have 0.12 goals per match less. This suggests that playing style is slightly more cautious and defensive in tournament finals.

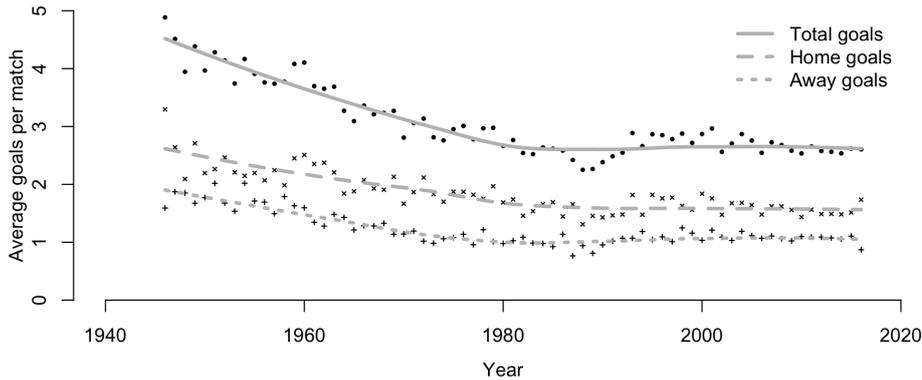


Figure 1: Average goals scored from 1945 to 2016. The grey lines are smoothed loess lines fitted to the data.

Figure 1 also shows the average number of home team goals per match, and the average number of away team goals per match. The difference between the home and away averages reveals information about the evolution of home advantage and it appears to be narrowing such that home advantage is now less than what it once was.

4.2 Team level descriptives

Table 1 shows some summary statistics for the top ten teams according to the percentage of matches won plus one other team which has won the World Cup (Uruguay). In addition to overall win percentage, also shown are the home/away win percentages, the average goals for and against, the number of World Cups won, and the total number of games played.

It is probably no surprise to football fans to see Brazil at the top of the ratings. However, it is wrong to use this type of table to determine which is the best team, as this does not take into account the quality of the opposition. The presence of Korea for example is likely to be a consequence of such a bias since they will play countries from their federation more often, and Korea belongs to a relatively weak federation (compared to Europe or South America). A further complication is that the number of games played by each country differs. Larger sample sizes (like Brazil and England) would result in more precise estimates of the win percentage, whilst teams like the USSR and the Czech Republic have played fewer games and the calculated win percentages are likely to be less precise.

Before moving on to our results, we note that the difference between home and away win

Table 1: Summary statistics for countries playing more than 200 games between 24th December 1944 and 25th May 2016. The first ten countries are the top 10 according to win percentage. Uruguay has been added as a past winner of the World Cup.

| Ranking | Country | Win % | Home win % | Away win % | Neutral win % | Average goals for | Average goals against | World Cups won | Total games |
|---------|----------------|-------|------------|------------|---------------|-------------------|-----------------------|----------------|-------------|
| 1 | Brazil | 64.0 | 72.8 | 50.6 | 66.7 | 2.17 | 0.86 | 5 | 900 |
| 2 | Germany | 61.5 | 64.7 | 55.8 | 71.1 | 2.21 | 1.01 | 3 | 358 |
| 3 | West Germany | 59.8 | 63.1 | 55.1 | 56.9 | 2.10 | 1.00 | 0 | 408 |
| 4 | Spain | 57.3 | 68.6 | 44.8 | 50.9 | 1.91 | 0.89 | 1 | 626 |
| 5 | England | 57.0 | 60.1 | 53.8 | 43.1 | 2.07 | 0.91 | 1 | 797 |
| 6 | USSR | 56.2 | 67.1 | 49.2 | 48.1 | 1.82 | 0.82 | 0 | 406 |
| 7 | Czech Republic | 55.4 | 71.1 | 41.4 | 33.3 | 1.84 | 0.94 | 0 | 251 |
| 8 | Korea Republic | 54.1 | 63.5 | 41.1 | 16.1 | 1.82 | 0.94 | 0 | 1169 |
| 9 | France | 53.9 | 60.2 | 45.3 | 49.1 | 1.82 | 1.00 | 1 | 739 |
| 10 | Italy | 53.6 | 63.9 | 40.4 | 50.7 | 1.62 | 0.87 | 4 | 666 |
| 43 | Uruguay | 43.2 | 57.6 | 31.1 | 34.1 | 1.53 | 1.20 | 2 | 680 |

percentages given in table 1 offers an interesting insight into home advantage. It appears, from this naïve analysis, that the Czech Republic has experienced the greatest home advantage, having a 30% drop in win percentage for away matches.

Similarly, one might consider the win percentage in neutral venues a measure of how well a team plays at tournaments. Germany tops this list with Brazil in second. England drops dramatically, whilst Italy rises the rankings.

5 Results

There is no standard method to assess goodness-of-fit for the model used here. Instead, we adopt an out-of-sample procedure whereby one game in 10 was omitted from the model fitting, and only these omitted games used in the goodness-of-fit testing. The numbers of won, drawn and lost games for 10 years around the peak performance for the top 50 teams were used to compute a chi-squared statistic; expected numbers of games won, drawn or lost were computed from the model. Each team gives a chi-squared statistic with two degrees of freedom, since the total number of games is fixed. Thus there are 100 degrees of freedom, and this yielded $\chi^2[100] = 102.24$, i.e. the model provides a reasonable fit.

To examine the contribution of each model component to the goodness-of-fit of the overall model we can examine the change in the out-of-sample log-likelihood, as each component is added. On largely omitting shrinkage by setting $\psi = \phi = 10$, the log-likelihood decreased by 38.1, showing that shrinkage is important. Omitting inflation of diagonal scores had almost no effect on this log-likelihood, so this model refinement, although it improves the fit to the data, does not improve prediction of cup-match results. Omitting home advantage decreased log-likelihood by 9.8, showing that this aspect of the modelling is important. Moving to 20 folds increased the log-likelihood by only 1.2, showing that 10 folds are enough. According to this measure, the parts of the modelling that improve predictive ability are, in decreasing order of

Table 2: All time greatest international football team: columns 2 to 5 give the ranking according to the maximum strength achieved by a team over a four year period with the value of the points won (1 for a win, 0.5 for a draw and 0 for a loss) and the central year it was achieved. The final four columns give the ranking according to a team’s maximum ten-year points won. Figures in parentheses show standard errors.

| Ranking | Four-year maximum strength | | | | Ten-year maximum strength | | | |
|---------|----------------------------|------|---------------|----------------------------|---------------------------|------|---------------|----------------------------|
| | Country | Year | Points (s.e.) | $\bar{a} - \bar{b}$ (s.e.) | Country | Year | Points (s.e.) | $\bar{a} - \bar{b}$ (s.e.) |
| 1 | Hungary | 1952 | 0.681 (0.05) | 0.642 (0.03) | Hungary | 1951 | 0.692 (0.03) | 0.649 (0.03) |
| 2 | France | 1998 | 0.647 (0.06) | 0.092 (0.04) | Brazil | 1998 | 0.663 (0.03) | 0.156 (0.03) |
| 3 | Spain | 2010 | 0.644 (0.06) | 0.048 (0.06) | Spain | 2008 | 0.629 (0.03) | 0.103 (0.03) |
| 4 | Colombia | 2013 | 0.606 (0.06) | 0.119 (0.05) | Germany | 2009 | 0.607 (0.03) | 0.172 (0.03) |
| 5 | Brazil | 1995 | 0.601 (0.04) | 0.106 (0.04) | France | 2001 | 0.596 (0.03) | 0.09 (0.02) |
| 6 | USSR | 1955 | 0.597 (0.05) | 0.568 (0.01) | Argentina | 2002 | 0.595 (0.04) | 0.14 (0.02) |
| 7 | West Germany | 1974 | 0.596 (0.05) | 0.132 (0.05) | West Germany | 1976 | 0.587 (0.03) | 0.147 (0.02) |
| 8 | Netherlands | 1973 | 0.594 (0.05) | 0.168 (0.04) | England | 1964 | 0.58 (0.03) | 0.408 (0.02) |
| 9 | Ghana | 1965 | 0.593 (0.08) | 0.471 (0.04) | Netherlands | 2001 | 0.577 (0.04) | 0.127 (0.02) |
| 10 | Germany | 2012 | 0.583 (0.04) | 0.179 (0.04) | USSR | 1958 | 0.576 (0.03) | 0.503 (0.02) |

importance: shrinkage, home advantage, down-weighting friendly matches, inflation of diagonal scores.

Our model gives estimates of each team’s attack strength, $\alpha_i(t)$, and defensive strength, $\beta_i(t)$ at time t . As discussed above, we use a pseudo-win probability measure to rate the teams (we call this measure ‘points’ in our results). Table 2 shows the top 10 teams for 4-year and 10-year strength. Hungary in 1952 is the strongest team over both the 4-year and 10-year periods. France in 1998 and Spain in 2010 are second and third in the 4-year ratings, whilst Brazil in 1998 and Spain in 2008 are second and third in the 10-year table. Football aficionados will very much like seeing Hungary 1952 top the ratings list, and undoubtedly it is hard to argue with the presence of serial winners such as France 1998 and Spain 2010. The France team won the 1998 World Cup and 2000 European Championships, whilst a World Cup victory in 2010 for the Spain team was sandwiched between consecutive European Championship victories in 2008 and 2012. Perhaps the most surprising absence is that of the Brazil side around 1970 which is widely regarded as the greatest team of all time. We comment on this in section 6.2 below.

Figure 2 shows the evolution of strengths for three of the sides in our all-time greatest ratings list. It is noticeable how Brazil is very strong throughout the time period and has many high but narrow peaks. France and Spain however, have a much lower average strength across the entire time period, but have longer periods of sustained high strength in 2000 and 2010 respectively.

Figure 3 shows the evolution of strength for Brazil. It is noticeable how for four out of five of Brazil’s World Cup victories, the team is estimated to be at a local maximum and the victory is followed by a slump. Also shown on the plot are the average numbers of goals scored and conceded by Brazil. Other than around the late 1990s, there is very little evidence of a strong relationship between the estimated strength and the goals scored/conceded.

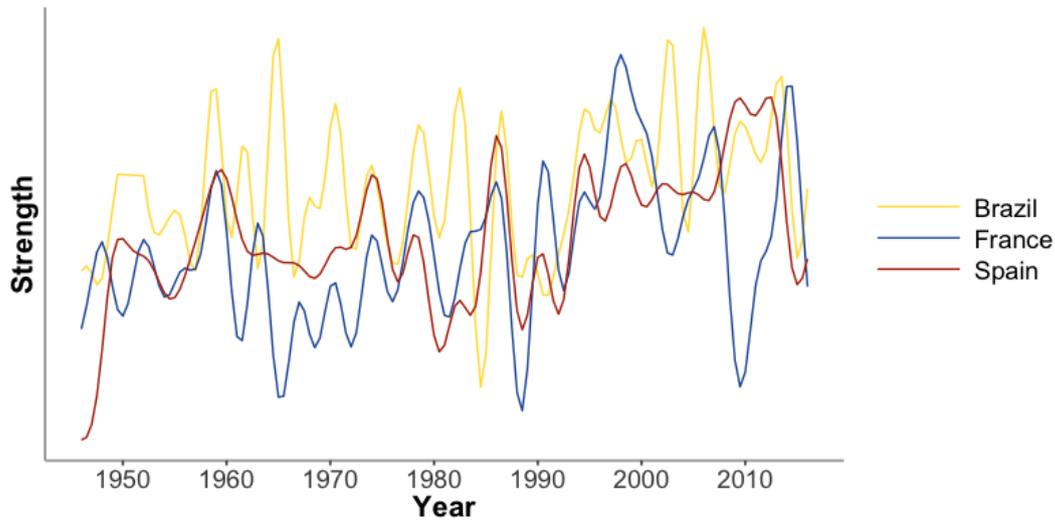


Figure 2: Strength trajectories for Brazil, France and Spain from 1945 to 2016.

6 Extra time: some additional results

6.1 Luckiest teams

Along with the ‘who was the greatest?’ question, one can ask ‘who was the luckiest?’, or ‘which results are most surprising?’. Here we must add ‘with hindsight’, so we mean that, modelling a team’s performance before and after a cup match, which win was luckiest (or which defeat was unluckiest).

A straightforward way to answer this is to omit all tournament matches from a particular year (but include friendlies) and find the model implied probabilities of each team winning the matches it plays in the tournament(s) for that year. This process can be repeated for all years. The ‘luckiest’ team is then the one benefiting from the least likely result.

A now famous result occurred when Germany beat Brazil 7-1 in the 2014 World Cup. The predicted probability of this scoreline happening was less than 0.001. These low probabilities are most likely an underestimate: in reality, it may be that when one team is winning ‘hands down’, the other team basically gives up, which the model does not reflect.

One may wonder whether England’s famous win over West Germany in the final of the 1966 World Cup was a fluke. No: the probability of England winning the match was 0.38 and the probability of the observed goal difference or greater was 0.18.

6.2 Lessons from the media and overfitting

On presenting the results of an earlier version of this model to the BBC it became clear that the football experts were somewhat dissatisfied with the ratings. It soon became clear that the experts ‘knew’ which team should be at the top of the ratings (Brazil 1970). In trying to understand why Brazil 1970 was not top of the ratings (but was in fact 12th), we began experimenting with the model. On increasing the number of nodes the fitted time-varying

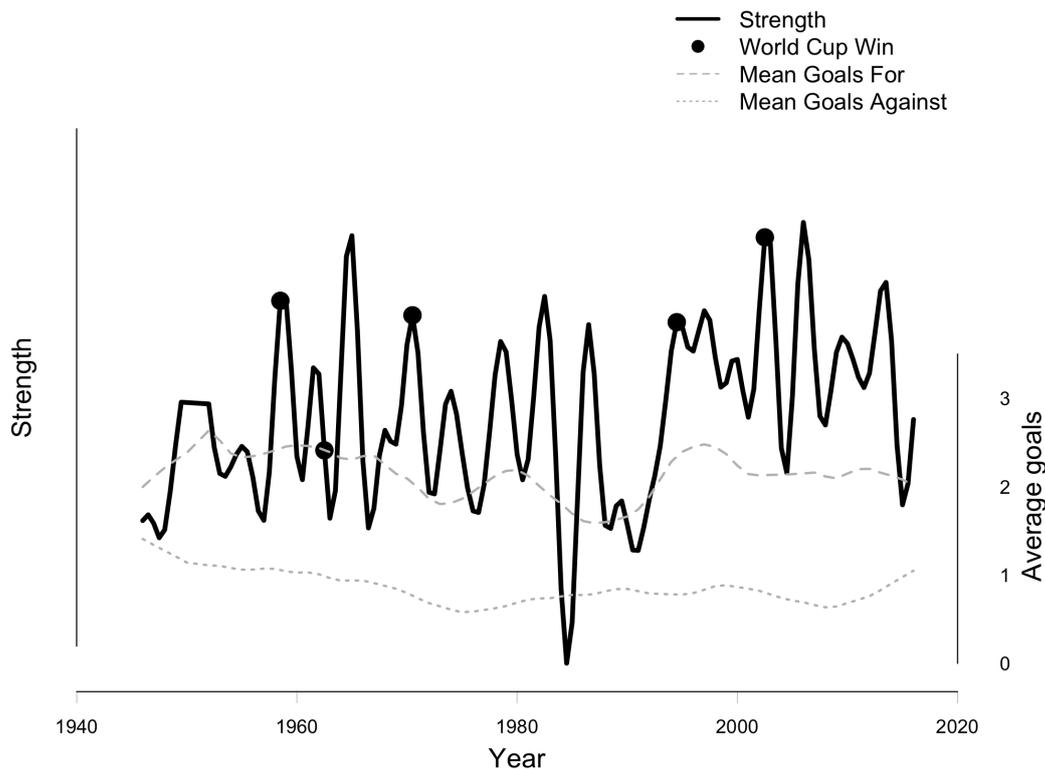


Figure 3: Strength trajectory for Brazil from 1945 to 2016.

strengths become more wriggly and respond to match results more sensitively. In tandem with decreasing the relative importance of friendly matches, Brazil 1970 did indeed rise through the rankings. In doing this experiment, we were in fact over-fitting the data and this raises an important philosophical point.

When using ‘proper’ statistical procedures to ask ‘who was the greatest’ we are using statistical methods to essentially remove noise from the results of football matches and then summarise the findings in the form of rankings. The resulting rankings reflect what should have happened given the recent performances of the team(s) without the element of chance. However, chance did happen and this is what people remember. The practical usefulness of a conventional model is mainly that it would give a much better indication of future performance than the overfitted model. It may also please some football enthusiasts, who can see that their team ‘should’ have won but were unlucky.

However, the overfitted model is more representative of what actually happened and better reflects what people remember. In the end, the BBC decided to use the overfitted model which used double the number of nodes, M , and reduced the importance of friendly matches to 0.1. The article can be viewed at <http://www.bbc.co.uk/sport/football/36387046>.

7 Conclusions

Despite being in the era of ‘big data’, it is still a common problem for statisticians to find themselves dealing with ‘small data’, meaning that fitting complex models and obtaining precise, meaningful parameter estimates is difficult. In seeking an answer to the question ‘which is the all-time greatest team’ in international football, we found ourselves in this situation. Our solution was to use Empirical Bayes methodology, in which a ‘prior’ distribution shrinks team strengths towards a common mean. Estimation of the model parameters was done using a combination of maximum a posteriori estimation and k-fold cross validation.

Besides the relatively small size of the dataset, further problems were optimising the number of ‘nodes’ to be used in the barycentric interpolation of team strengths, and optimally down-weighting the results of ‘friendly’ matches, which give some information about team performance in cup matches, but are less useful in prediction than actual cup-match results. These parameters could (probably) be estimated by seeking to maximise the posterior probability of a complex Bayesian model, but again k-fold cross-validation was used, as it does not require approximations to be made, and is computationally far simpler.

Our results suggest that the Hungary team in 1952, the France team around the turn of the 20th Century, and the Spain team around 2010 are strong contenders for the strongest teams ever.

In addition, we have touched on some other interesting questions not before addressed, such as ‘which team was the luckiest?’, and given our experience of presenting an earlier version of this work to the media. This is part of the experience of doing OR.

We hope this work demonstrates that quantitative methods can be used to answer real-world, interesting problems; that this engages the public in OR, and that we ‘add’ to the debate, not end it!

References

- Baker R and Jackson D (2014). Statistical application of barycentric rational interpolants: an alternative to splines, *Computational Statistics*, **29** (5), 1065-1081.
- Baker, R.D. and McHale, I.G. (2014). Time varying ratings in association football: the all-time greatest team is. *Journal of the Royal Statistical Society: Series A, Statistics in Society*, **178** (2) 481-492.
- Baker, R.D. and McHale, I.G. (2017). An empirical Bayes model for time-varying paired comparisons ratings: Who is the greatest women’s tennis player? *European Journal of Operational Research* **258** (1) 328-333.
- Bastos, J.A. (2010). Forecasting bank loans loss-given-default. *Journal of Banking and Finance* **34** (10) 2510-2517.

- Carlin, B. P and Louis, T. A. (1996). Bayes and Empirical Bayes methods for data analysis, Chapman and Hall, New York.
- Hastie, T., Tibshirani, R. and Friedman, J. (2009), The elements of statistical learning: data mining, inference, and prediction, 2nd ed., Springer, New York.
- Higgins J. P. T. and Thompson S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine* **21**, 1539-1558.
- Dixon, M. and Coles, S. (1997). Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 46(2): 265-280.
- Graves, T.L., Reese, M.A. and Fitzgerald, M.A. (2003). Hierarchical Models for Permutations: Analysis of Auto Racing Results. *Journal of the American Statistical Association*, 98(462): 282-291.
- Hu, F. (1997), The asymptotic properties of the maximum-relevance weighted likelihood estimators, *Canadian Journal of Statistics*, **25**, 45-49.
- Maher, M. J. (1982). Modelling association football scores. *Statistica Neerlandica*, 36(3):109-118.
- Maritz, J. S. and Lwin, T. (1989). Empirical Bayes methods (2nd. ed.), Chapman and Hall, New York.
- Wang, X. and Zidek, J. V. (2005), Selecting likelihood weights by cross-validation, *Annals of Statistics* **33** (2), 463-500.
- Vasicek O. A. (1973), A Note on Using Cross-Sectional Information in Bayesian Estimation of Security Betas. *The Journal of Finance*, **28** (5), 1233-1239