Predicting Match Outcomes at the FIFA World Cup

Dylan O'Connell

December 13, 2017

Abstract

We predict the outcome probabilities for each match in the 2018 FIFA World Cup group stage using an Elo ratings model. We initialize team Elo ratings using the 2001 FIFA world ranking, and update these with the results of all World Cup matches played since 2002 (including their qualifying, group stage, and knockout games). We adjust the Elo ratings model through a variety of parameters so that it better fits the observed data, and find that the model efficacy improves when we incorporate the margin of victory and place greater weight on both more recent matches, as well as matches played at the World Cup main event. To account for home field advantage, we estimate its impact and incorporate a flat ratings boost for any home teams to balance out its effects. We search among a broad range of possible parameter adjustments, and select the resulting model which best matches the historical data and our model assumptions. This model calculates a final Elo rating for each of the 32 national teams in the 2018 World Cup, which we use to estimate the probability of each outcome in the 48 group stage matches.

1 Introduction

There are 32 teams in the 2018 FIFA World Cup group stage, divided into 8 groups of four teams (groups "A" through "H"). Each team will play each of the three other members of their group exactly once, but the selection of these groups is not completely random. Qualified teams are divided into four "pots" based on their perceived strength in the FIFA ranking (with the host nation being automatically added to the strongest pot), and each group receives a team from each pot. Group composition is further restricted as multiple teams from the same federation cannot be drawn into the same group (with the exception of Europe's UEFA, as there are more teams than groups) [1]. Our primary dataset is compiled from all matches played in the 2002, 2006, 2010, 2014, and 2018 World Cups (qualifying, group stage, and knockout, with 2018 only including qualifying), and the goal is to use these prior results to predict the likelihood of match outcome for each of the 48 group stage games in 2018.

The selection process outlined above ensures that most group stage matches are played between teams from different federations, which means that we have an exceedingly sparse set of repeated matchups in our dataset. In fact, there are only 14 total matches in our 16 years of previous World Cup data that exactly match any to be played in the 2018 group stage. World Cups are the primary time where teams play outside of their federation, so a direct examination of prior head to head matchups is unproductive. In fact, the fundamental challenge of this task is that past results in soccer are entirely dependent on the context of the strength of the opposition at the time. This creates a circular challenge, as rating team strength based on results relies on prior knowledge of a team's strength.

As an example of the futility of an analysis of a team's raw past results, we consider Tables 1 and 2. We compare the base statistics for Australia and Argentina in the entirety of our dataset (including group stage matches), and solely in group stage matches, respectively. We can see that Argentina has vastly outperformed Australia in the World Cup group stage, despite Australia's strong performance in the totality of our dataset (given that this includes their group stage losses, their qualifying performance is particularly spectacular). Any prediction for group stage results based on these raw statistics for each team would invariably have failed. The issue is simple and fundamental. The difficulty of opposition that each team faces is not intended to be random. In this case, Argentina must qualify in the challenging CONMEBOL South American federation, where they consistently play a large number of matches against world class opponents. Australia has a more unique path to qualification through the ASEAN federation in South East Asia, where they play teams that include relatively small island nations. For instance, Australia's impressive goal differential is heavily boosted by consecutive matches where they defeated American Samoa 31-0, and Tonga 22-0. In short, any analysis of the raw statistics of a team's prior matches is irrelevant unless we have context for the strength of their opposition.

Table 1: All World Cup matches, 2002-2018.

	Wins	Draws	Losses	Mean Goals For	Mean Goals Against
Australia	36	11	10	3.14	0.84
Argentina	39	19	11	1.69	0.91

Table 2: Group stage World Cup matches, 2002-2014.

	Wins	Draws	Losses	Mean Goals For	Mean Goals Against
Australia	2	2	6	1.22	2.22
Argentina	9	2	1	1.92	0.58

1.1 Elo Rating System

Elo ratings provide a broad framework for continuously updating the relative strength of each national team at the time of each successive match, which we will need in order to use the past results for future predictions. The core setup is as follows. Each team has a rating (which we denote R), that is updated as after each match they play. If team A and team B (with ratings R_A and R_B) play each other in a match, we can calculate the Expected Score for team A (E_A, shown in Equation 1) in that match. Here, score is a function of the match result (not the goals scored), which in its base form is generally $E_A = 1$ for a win, $E_A = 1/2$ for a draw, and $E_A = 0$ for loss (we will see that we can adjust this concept of "score" to account for margin of victory in Section 3.2). Then, once we observe the result of the match (which has observed score S_A), we can use this, the expected score (E_A), and a tuning parameter K weighting the match (which we will discuss in detail in later sections) to update the Elo rating for team A to its new value, R_A^* (shown in Equation 2, with a corresponding update for team B).

$$E_A = \frac{1}{1 + 10^{-(R_A - R_B)/400}} \tag{1}$$

$$R_A^* = R_A + K(S_A - E_A) \tag{2}$$

This is a naturally self correcting system, because an upset causes a large shift in score for both teams, while a result that is expected will cause a much smaller shift in ratings. This property is particularly appealing to us, because we know that a primary challenge will be that we naturally expect the strength of national teams to shift in ways different than if we were measuring the rating of an individual. As players retire and new ones join, the team composition itself can change dramatically between competitions, so we want a rating system that works to quickly adjust to data that conflicts with its previous belief. Further, if two teams play each other endlessly with a fixed probability of match outcomes, their Elo ratings will reach equilibrium, rather than endlessly diverging, even if one team has a positive winning percentage. In fact, it will reach the equilibrium where that winning percentage corresponds with the given Expected Score between this pair of Elo ratings. Certain rating systems tend to blindly reward teams for playing additional matches, while this system only cares about the quality of a team's results. The 400 term in Equation 1 determines the scale of the rating system. This is the constant used by the FIDE chess rating system, and it implies that a team with a 100 point advantage in Elo rating has an Expected Score of about 0.64, which is a fairly intuitive scale to work with.

Elo ratings were originally developed for Chess, but they have seen use in a wide variety of fields. American Division I college football famously used an Elo system as part of its computer ranking in order to select the two teams which would play for the national championship (until it was replaced by the Playoff Committee in 2013), and the website "FiveThirtyEight" has adapted Elo rating systems to make predictions for most major American sports. Elo ratings provide a standard framework for estimating the outcomes of head to head competition, and while the system requires some approximations of the behavior of the sport, it continues to see use because of its broad applicability to many topics.

Elo rating systems are not without their flaws. For example, one study of Elo ratings in chess showed that it tends to underrate the chance of an upset in very lopsided matches [2]. A primary reason hypothesized for this is that weaker players tend to improve more quickly between tournaments than stronger players, which means the algorithm will tend to underestimate the weaker opponent's chances in a match. This by itself is not likely to be a major concern for national teams in soccer. In chess, each player tends to improve as their career progresses, while this cannot universally be true among soccer national teams, as the pool of top competitors is largely fixed (besides political shifts in country definitions) and their skill is determined relative to the pool of teams. Thus, we should be concerned that the strength of teams fluctuates between World Cups (which it does), but it is unlikely to be systemically true that all teams tend to improve over time, as team strength is relative to a fixed pool of national teams.

Precise analysis of Elo ratings requires assumptions about the dataset that are unlikely to be exactly true, but the system is somewhat robust against these inaccuracies, due to its self correcting nature. We assume that each team has some true strength at a given moment in time, which we cannot directly measure. The crucial Expected Score calculation (Equation 1) assumes that each team has the same standard deviation for their observed performance in a given match (which is randomly distributed around their true strength at that time) [2]. This is a core assumption that may not precisely fit our data, as it is difficult to prove that some national teams could not have a higher standard deviation of observed performance given such limited data. Further, Elo ratings assume that shifts in the true strength of a team are gradual over time. This depends on the time frame that one considers, but among soccer national teams this is unlikely to always be true. Sometimes a large number of players will retire between World Cups, or for a specific match, a crucial star player may be missing due to injury. Unfortunately, it is entirely possible for a national team to have a rapid shift in true strength. We note that this will prove problematic no matter our approach. It incentivizes us more heavily weighting extremely recent matches rather than taking a broader look at past performance. It is reasonable to place a high weight on recency, but given the sparsity of our dataset, and the inherent randomness involved in soccer, we have to strike a balance. It is trivial to find cases where a team has an excellent match on one day, and plays poorly soon after, with no changes to be found between the games, as we understand that the results of a soccer match have a relatively high variance. Thus, we can see that there are elements of Elo rating assumptions that are not precise fits for our data. However, by and large, similar assumptions are unavoidable for any insightful analysis, and an Elo rating system is well equipped to produce reasonable results even with some violation of assumptions. Ultimately, the way to address these concerns is to carefully examine our resulting model, and ensure that the results are intuitive and accurate along the way. Indeed, much of our work will come from trying a variety of Elo based approaches, and analyzing the results.

Elo ratings simply provide a framework for our analysis, and allow for significant flexibility in their exact implementation. Most of our work will involve fine tuning the parameters of our rating system so that it leads to more suitable predictive outcomes. We can adjust the tuning parameter K (how we weight each individual match), to place higher emphasis on matches we deem to be greater predictive value (such more recent or more important ones), which we will detail in Section 3. Indeed, a crucial advantage of Elo ratings compared to other newer competitors is that they are relatively intuitive (and indeed are computationally very simple). This allows us to better refine the algorithm using domain knowledge than would be possible with a rating system whose inner workings are a "black box" to outside analysis. To simplify the implementation of our model, we use the Elo package in R developed by Ethan Heinzen to compute rating changes [4]. This is largely to improve the readability of the code, as we can see that the Elo system is computationally straightforward.

2 Data Cleaning

Our primary dataset is the collection of all World Cup matches (qualifying, group stage, and knockout) played from 2002-2018, collected by the Rec Sport Soccer Statistics Foundation [3]. Due to the concerns outlined in Section 1.1, we cannot solely examine the matches involving the given 32 teams. We need to consider all matches in the dataset, so that we can continuously update our estimated rating for each team. It is tempting to think that matches from 2002 may not be relevant for predicting match outcomes in 2018, as there are unlikely to be any common to both rosters. However, as discussed further in Section 3.3, we will make use of these early matches both because historical strength is predictive of future success (as the infrastructure to create a top tier national team persists long beyond a single roster), and because our Elo rating system will be more accurate if we allow it to adjust over a longer period of time.

We consider only the match result at the end of regulation time, so that the matches are played with a consistent format (knockout rounds go to extra time, and then penalty kicks, both of which are significant changes to the match structure). We further remove all matches that are listed as abdicated, annulled, abandoned, or not played. Thus, we only consider the 2964 matches that have officially held results at the end of regulation time. For each match, we record names of the two national teams that participated, the number of goals scored by each team, the year of the corresponding World Cup, and whether it was a qualifying match or at the World Cup main event. For qualifying matches, we record which country hosted the match, and for main event group stage matches, we record the group they played in (when a category does not apply to a match, we leave it as NA). After cleaning, we see that there are no missing group stage matches in this dataset (for the qualifying rounds, many matches must be discarded due to being annulled or abandoned).

The primary challenge in assembling the data is that national teams are denoted in a vast variety of ways in this dataset, and it is vital that we ensure consistency in tracking each team. To do this, we create a dictionary that maps all relevant abbreviations and alternative spellings into a consistent format. In total, we require 115 mappings to interpret this dataset, although they only apply to 42 different countries, because certain countries are abbreviated in a number of disparate ways.

The other challenge we face is that there are several countries whose political borders have shifted through the course of our dataset. Most of the changing membership of FIFA comes from small teams at the fringe of our dataset. While these teams largely do not make it to the World Cup, their performance in qualifying rounds has some minor indirect impact on the rating of the World Cup teams that we are ultimately concerned with. We generally initialize newly introduced teams with a rating near the bottom of our scale, as most new additions are small island nations who are just forming the infrastructure for a national team (see Section 2.1 for further discussion of initial ratings). For example, Kosovo and Gibraltar were awarded FIFA membership for the 2018 World Cup, and had not previously competed. Their combined record in the 2018 qualification rounds was 0 wins, 1 draw, and 16 losses, supporting our intuitition that small nations who are just beginning their national teams must essentially start from scratch.

The crucial exception is the Serbian national team, which is one of the 32 teams in the 2018 group stage (and thus of direct interest to us). Through 2006, Serbia & Montenegro competed with a joint national team (in 2002 listed as Yugoslavia, in 2006 listed as Serbia & Montenegro), at which point Montenegro declared independence. Ultimately, the core team became the Serbian national team, and Montenegro formed a new national team. Thus, we can simply count the results of the Serbian & Montenegro national team as being part of the Serbian team's history, and consider the Montenegro national team to be newly formed as of 2006. The Montenegro national team did eventually achieve relevance, but its early results were quite poor, and it wasn't until the 2012 World Cup that they began to rise. Thus, this fits with our assumption that newly formed national teams can be considered weak, and their rise to relevance should be properly tracked by our Elo rating system.

We note that each of the 32 teams in the 2018 World Cup group stage are fairly well tracked by our data. The team with the fewest matches played over the 16 year period is Iceland (with 31), but most countries have between 40 and 60 total games played. For each team, the majority of matches played come from their qualifying rounds. A cursory examination of the rosters used for qualifying matches show that World Cup quality nations are fairly consistent about fielding their best roster, even in qualifying rounds. In fact, there are a variety of restrictions on clubs that prevent them from having priority over those who could play for a national team, so qualifying matches tend to field representative rosters, despite long travel times [5]. This means that qualifying matches are fairly indicative of a team's strength, although we will later evaluate whether our model improves by weighting main event matches higher than qualifying matches (Section 3.3). The main concern about qualifying matches is that they tend to involve repreated matches against other teams from the same federation. Thus, if the initial ratings for those federations are off, it can take a long time for an Elo rating system to properly balance them out, increasing the need for a reasonable initialization of Elo ratings..

2.1 Base ELO rating

We next consider how to initialize the Elo rating of each team (beginning in 2002) with predetermined values. Given a sufficiently large sample of matches, this would eventually have little effect, as the Elo rating system would settle to the proper values. However, for each team, we do not have a nearly sufficiently large sample of games for this to be true. In particular, the confederations system of qualification means that weak regions are fairly incestuous, and rarely play the stronger regions. This means that if we initialize each team to the same starting rating (for this model, we choose 1000 as the average rating), the ratings may not properly calibrate for the entirety of our dataset (see Section 4.2 for an example of this).

FIFA maintains a rating of each team in the world, which they use for a variety of official purposes (for example, their group stage seeding). We can use the FIFA ranking from October 2001 [6] as the basis for our initial ELO values. The FIFA ratings range from a score of 800 on the high end (this is not the maximum possible, but the maximum in October 2001), to 10 on the low end. We rescale the scores so that they have a mean of 1000 and standard deviation of 200. This provides only a rough approximation, as we have no proof of the true distribution of team skill. However, without this initialization, we see that our Elo system is wildly inaccurate for the early time period of our dataset. Further, it will include some bizarre outliers in Elo rating up through 2018 (usually small island nations that only play a few matches amongst themselves which retain much of their initial average ratings). We will see that initializing the scores in this fashion vastly increases the accuracy of our ratings.

3 Constructing the Model

We have chosen to model team strength using an Elo rating system over the course of this dataset. However, this only provides a framework for our analysis, and there are numerous choices that we have to make regarding its implementation. In broad terms, this section outlines our options for how to tune the model to better match the reality of international soccer play. Obviously, a primary difficulty of this approach is that we do not have an extensive ability to test our model. In Section 4, we will consider various model specifications and consider the resulting predictions for the 2014 group stage (treating it as an informal version of a "training set", to guide our search). Obviously, we should not treat the 2014 results as gospel (as the results of individual games are high variance), so we evaluate the resulting models using a combination of the results of 2014, the betting markets prior to 2014 (Section 4.1), and most importantly, our intuition about what sensible model output would look like (Section 4.2). We then use the model selected in this process to evaluate the upcoming 2018 World Cup group stage.

3.1 Handling Draws

Once we have determined an Elo rating for each team, the Elo system can calculate an Expected Score for a given matchup. In the case of a binary outcome (win or loss), the Expected Score for a given match competitor is simply their probability of winning the match. However, when draws are possible, Expected Score essentially estimates the expected value of the result of the match for the competitor, where wins are worth a point and draws are worth half a point. That is, Expected Score $= \mathbb{P}[\text{Win}] + \mathbb{P}[\text{Draw}]/2$. Clearly, a computed Expected Score does not uniquely define a winning and drawing probability. For two evenly matched teams, the expected score is 1/2, and that could arise from a each team having a 50% chance to win (with a 0% chance to draw), each team having 100% chance to draw (and 0% chance to win), or any combination in between. Thus, we must turn to historical soccer data to estimate how to convert from Expected Score to a direct probability of a draw.

We first note that any unique correspondence between Expected Score and drawing probability is an approximationt. It is very likely that two matchups with the same Expected Score might have two different drawing probabilities. For instance, two evenly matched defensively minded teams are more likely to draw than two evenly matched aggressive teams, because fewer goals in a game increases the chances of draws. However, a team's playstyle is difficult to estimate over a long period of time with our small sample of matches, so we will approximate this by following a one to one conversion from Expected Score to the probability of a draw (which in turn defines the probability of a direct win).

We can use this conversion as one of the parameters which we adjust so that it best fits our observed data, but for an initial value we turn to historical data. If we average the draw percentage of the top four leagues[1], we get 26% of matches ending in draws. In terms of expectations, one betting site for the Premiere league on average offers a draw percentage of 26.1%, with a standard deviation of 4.1% (however, the odds of a draw almost never rose above 31%, so it has a heavy left tail) [8]. In terms of our dataset, only 21.7% of recorded matches in the entire dataset are draws, while 25.1% of matches in the group stages were a draw (this disparity is unsurprising, as the qualifying rounds tend to involve much less balanced competition). We do not want to overfit to our data by guaranteeing that the past group stage draw trend will hold, which is why we consider the broader proportions of draws in professional soccer. For an initial estimate, we choose that for two exactly evenly matched teams, there is a 29% chance of a draw. This is a bit less than one standard deviation above what see in the top leagues, to compensate for the fact that group stages seem to draw slightly less often. Then, as we account for the average disparity between teams, we wish to see this draw probability diminish so that the average is roughly 25%. We create a starting formula to achieve this, which is based on an Expected Score (ES) and a tuning parameter (k_d, where a larger k_d means that more unequal matchups have a smaller chance of a draw and a larger chance of outright victory).

$$\mathbb{P}\{\text{Draw}\} = .29(1 - 2|ES - 1/2|)^{k_d}. \tag{3}$$

This is our initial estimate, and we can use the value of k_d to tweak the relationship between Expected Score and probability of a draw until our model produces results consistent with what we observe.

3.2 Margin of Victory

Throughout this report, when we discuss the “score” (or “Expected Score” or “Observed Score”) in the context of our Elo ratings, we have been referring to the match result (with 1 denoting a win, 1/2 a tie, and 0 a loss). This approach is sensible, but it is possible that we are missing out on additional valuable information by ignoring the margin of victory (which is the difference between the scores of the two respective teams). We have to be careful not to overweight margin of victory, as it is match result that ultimately determines advancement (in qualifying and group stage matches, margin of victory is only ever a tiebreaker), so teams do not have as much incentive to widen a blowout lead as they do to gain the lead. Further, we have to be concerned with the concept of “garbage time”, which is a widely believed phenomena where teams will not necessarily play their hardest once flipping the game result is essentially out of reach (the idea is that if a team is down multiple goals without much time left, they may “give up” and be more likely to concede additional goals, leading to a wider blowout). Thus, we wish to test margin of victory as an alternative approach to our regular score format.

We still need to have the score range from 0 to 1, with 1/2 denoting a draw. However, when using margin of victory, we can now treat narrow victories as not being worth a full point, and the same for narrow defeats being worth more than 0 points. As mentioned before, we want to minimize the impact of “garbage time” goals on our rating, thus we want to make sure that the difference between a 1 goal win and a draw is substantially more than

[1] La Liga: 24.1%, Premiere League: 28.2%, Ligue 1: 28.4%, Bundesliga: 23.2%[7]

the difference between a 5 goal win and a 4 goal win. As before, let S_A be the observed score of the match, and let M_A be team A's margin of victory (which is negative in a loss). Then, we choose our score function so that each additional goal brings the score half the remaining distance to 1 (or 0), as shown in Equation 4 (for $M_A = 0$, we define this sum to be 0, so $S_A = 1/2$). Thus, a draw is worth 1/2 score, winning by one goal is worth 3/4 score, and losing by 2 goals is worth 1/8 score, and so on. Thus, the most significant goal is still the difference between a win and a draw, and each successive goal adds more (but decreasing) weight to the significance of the result. In Section 4, we will examine whether incorporating margin of victory improves the predictive efficacy of our model.

$$S_A = 1/2 + \text{Sign}\{M_A\} \sum_{i=1}^{|M_A|} 1/2(1/2)^i \tag{4}$$

3.3 Weighting Historical Results

The principal challenge of national team predictions is that for any given team, the matches played represent a small sample over a long period of time. The 2018 group stage team with the largest number of total matches in our dataset is Mexico (with 74), but most play far fewer (with the average being 48 matches played in our dataset). We can see that Mexico's roster has shifted dramatically during this span (indeed, no players from the 2002 squad are still on the roster), which calls into question the predictive efficacy of looking at results from the distant past. However, the best world cup nations tend to show consistent success even as rosters shift (Germany, Italy, and Brazil combine for 13 World Cup victories, while the rest of the world only has a combined 7), likely due to their infrastructure for the sport. On the other hand, if a collection of quality players mature at the same time, a roster can achieve success that is at odds with the nation's historical strength,. For instance, in the five World Cups from 1990 to 2010, Belgium did not qualify twice, had one group stage exit, and two round of 16 exits. However, entering the 2014 World Cup, bookmakers gave them the fifth best odds to win the tournament (behind perennial powerhouses Brazil, Argentina, Germany, and Spain). This example demonstrates a major challenge that we face. Historical success is a reliable initial predictor, but it does not capture the rise of a particular roster. Indeed, en route to qualifying for the 2014 World Cup, Belgium only played 6 matches, mostly against opponents who were not strong enough to qualify for the World Cup themselves. It is hard to predict the Belgium squad's strength based on their historical success, but it is dangerous to weight their small sample size of qualifying victories against lesser opponents too highly. Thus, we want to find a delicate balance between weighting a team's full set of games and their smaller sample of recent games that better matches their current roster.

The main way that we account for this balance is in the K weights of the Elo system (Equation 2). The shift in Elo after each match result is tuned by the K parameter, with larger values placing higher weight on the match. We can adjust the K values to be small for games far in the past, and large for recent games. However, our weighting need not be limited to simply the date of the match. We will also experiment with giving higher weight to main event (group stage or knockout) matches over qualification matches. This makes intuitive sense, as qualification matches are more likely to be disrupted by rushed travel, and might not be as predictive. There is also a higher likelihood of the match being comparatively unimportant to the team (this year, Brazil secured qualification with several qualification matches remaining on their schedule).

Let c_t be our tuning parameter to weight match recency, and c_q be our tuning parameter to weight matches which are played in the main event. In each case, a value of 1 denotes that there is no additional weighting, and a value greater than 1 increases the weighting. Let k_b be the base K value (which will generally be initialized to around 10, leading to intuitive shifts in Elo, but it will vary based on our other tuning parameters), let Y_m be the year of the match we are considering, and Y_c the year we are projecting for (it will ultimately be 2018, but we will test our models on earlier years). Then our weighting formula is given by

$$K(k_b, c_t, c_q) = k_b + \frac{5}{8}(c_t - 1)(16 - (Y_c - Y_m)) + k_b(c_q - 1)\mathbb{1}\{\text{Main Event Match}\}. \tag{5}$$

We have a broad choice in defining this formula, but this leads to fairly intuitive results. We note that the additional weighting (beyond the k_b base value) is only present if we increase c_q and c_t beyond 1, and when they are present they are additive to the base value. We construct the time modification so that at the tail end of our dataset (16 years prior), the time bonus is 0, and it scales up to $10(c_t - 1)$ in the present day. The qualifying tuning parameter simply adds a constant based on whether the match was part of the main event of the World Cup. If we include these additive values (and set $k_t, k_q > 1$), we can simply reduce k_b so that the total Elo shifts do not become too dramatic. These tuning parameters will be a core part of our search for optimal parameters (Section 4), and we will select values for which they give us intuitive results.

3.4 Incorporating Betting Odds

We have outlined a variety of choices that we can make in specifying our Elo rating system, but ultimately we will need a way to evaluate competing models. Our first option is to use previous World Cup group stages (in particular, 2014, as we have the most data leading up to it) as a rough equivalent of a "training set" for our model (although only in an informal sense). While previous group stages are likely to be the best approximation of the 2018 group stage, the obvious weakness of that approach is that we do not want to overly emphasize the specific results of those games in particular. Tweaking the parameters of our Elo rating model is somewhat robust against overfitting (because our parameters simply adjust how we interpret the games leading up to the group stage, rather than directly reacting to the group stage results themselves), but the results of individual soccer matches are high variance. We want to use the past group stage results in our predictions, but we do not want to treat any upset as definitive, as any given group stage is a relatively small sample of 48 games.

Thus, one way to strike a balance between the observed results and an estimate of the spread of likely results is to use previous pre-tournament group stage betting odds. We obviously avoid betting odds related to the 2018 group stage, because basing our model on the betting odds of the games which we are trying to predict will simply fit our model to other people's predictions. However, we can use betting odds for the 2014 group stage as a way of hedging uncertainty for surprise upsets as we search for the correct set of tuning parameters (in Section 4.1).

We can find find the historical "moneyline" for the 2014 group stage matches online [9]. The lines are given in the form $\pm X$. If $X > 0$, then this implies that a correct bet of \$100 would earn you winnings of $\$X$, and if $X < 0$, it implies that you would have to bet $\$|X|$ to win \$100 upon a correct guess. We can convert a moneyline betting spread into its implied probability of each outcome. For instance, Switzerland vs France is listed as $+371$ for Switzerland win, $+254$ for a draw, and -119 for a France win. This implies that for each bet to break even in expected value, one needs the probability of a Switzerland win to be .212, the probability of a draw to be .282, and the probability of a France win to be .543. However, we note that these probabilities add up to more than 1, because bookmakers need to take a "vig" (usually around 5%) in order to make a profit. Thus, to get the true match probabilities estimated by these odds, we normalize them so that they add to 1, which means this line corresponds with probabilities of $0.204, 0.272$, and 0.523 for a Switzerland win, draw, and France win, respectively. In Section 4.1, we will show how we can can use the pre-tournament betting odds for the 2014 group stage games as an alternative metric for model evaluation that we balance with the observed results of the matches themselves.

3.5 Home Field Advantage: Qualifying and Hosting

Home field advantage is omnipresent among sports, and is purported to arise from a number of factors. Home teams avoid the discomfort of travel and staying in a foreign country, and many fans believe that players perform better when the crowd is cheering for them. However, the most widely documented aspect of home field advantage is the effect of the crowd on the referees. In soccer, referees are shown to generally provide more stoppage time at the end of a match when the home team is behind a goal, than they do when the home team is ahead by a goal [10]. It is also widely thought that referees are more generous on awarding penalties to the home team (one recent notable example of this was the public outcry over the officiating between Brazil and Croatia in the 2014 World Cup), although it is harder to demonstrate this quantitatively [11]. This raises two concerns about our model. First, qualification matches do consistently have a "home team", and we may not be properly accounting for that intrinsic advantage in our Elo rating. Second, we need to estimate the possible advantage that Russia will receive in each of their matches during the 2018 World Cup (which will not be included by our base model), by considering historical host advantage.

The first issue is relatively straightforward to test. In the case of qualification matches, we can pretend that the "home team" has a fixed value (c_h) added to their Elo for the purposes of calculating the Expected Score of the match. Thus, if A is the home team, the new Expected Score formula is given by

$$E_A = \frac{1}{1 + 10^{-(c_h + R_A - R_B)/400}}. \tag{6}$$

The concern is that without providing this home field advantage to our update formula, we will not be properly assessing the expected outcome of the match given the two initial Elo ratings. We run simulations of our Elo model on the dataset to determine the correct level of home field advantage to assign to the qualifying matches. It is important to note this differs in structure from the simulations in Section 4.1, which runs our Elo rating model on the dataset and analyzes its predictive results on a later World Cup group stage. Here, we are analyzing the results of our Elo rating model on the full dataset as it is updated along the way. With standard base tuning parameters, we note that with no home field advantage provided, our model calculates a mean Expected Score for the home team of 0.501, so on average our Elo rating system estimates that the home team is roughly equally likely to be the stronger or weaker side. However, when we consider the mean *Observed Score* for those matches, the home teams actually wins a 0.598 proportion of the total points (again, with 1 denoting a win, 0.5 a draw, and 0 a loss). This shows that our base Elo rating model consistently underestimates the performance of the home team, which will systematically skew its ratings update procedure. However, we can run the Elo rating system on our dataset using differing levels of home field advantage (c_h) as shown in Equation 6, and the results are shown in Table 3. We wish to select the c_h constant that best matches the average Expected Score with the average Observed Score (which is 0.598 in every case), so we select a c_h of 90 for our home field advantage constant. This corrects for the systematic bias in our Elo rating system for qualification games, and in Section 4 we show that this does improve our predictive accuracy.

Table 3: Average Expected Score under each level of home field advantage (c_h), compared to Average Observed Score

c_h	0	10	20	30	40	50
Avg. Exp. Score	0.501	0.512	0.523	0.534	0.545	0.556

c_h	60	70	80	90	100	110
Avg. Exp. Score	0.567	0.577	0.588	0.599	0.609	0.620

Avg. Obs. Score
0.598

Our second concern with home field advantage specifically involves the edge that we expect Russia to gain by hosting the 2018 World Cup. Some of the perks provided to hosts (such as automatic qualification, and preferable seeding into the group draw [12]) are not likely to directly impact our estimation of each match outcome. However, the aforementioned advantage of a favorable cheering crowd, or the increased emphasis that host nations place on their performance in that year, might provide a similar (or different) boost to that of the home field advantage enjoyed in qualifying matches. Indeed, this fits with widespread belief, as South Korea was notable for going on a surprise deep tournament run well beyond their usual strength (placing fourth), and the host advantage was often credited. For each host nation, we consider their mean Expected Score throughout the tournament they host as predicted by our base Elo rating model, and compare it to their mean Observed Score (based on wins, draws, and losses) under the same model, shown in Figure 1. The result is clear: a conventional Elo rating model that does not account for host advantage will tend to underestimate the performance of the host nation. Each of these tournament runs only represent a sample of 3 to 7 games, however the trend they demonstrate is consistent and significant, and fits with conventional wisdom of the host advantage. It's worth noting that the Elo rating system is self correcting, so as a team over performs, it naturally shifts its expectations throughout the course of the tournament itself (rapidly, in the case of large upsets). So over performance over the course of a longer tournament run like that of Germany and South Korea is particularly impressive, because their initial Elo had risen notably since the start to account for their success.

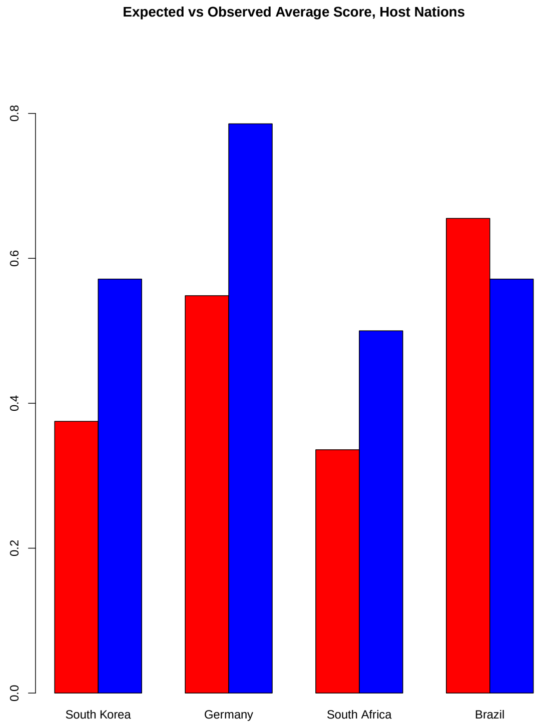


Figure 1: Mean Expected Score: 🟥
Mean Observed Score: 🟦

We consider potential tweaks to our approach to incorporate the strong performance of host nations. The most obvious approach is to simply use the additive constant (applied to Elo rating before the match, c_h) used to account for home field advantage in the qualification rounds (estimated above to be at $c_h = 90$ Elo rating). We run our simulation again, now giving hosts this additive constant when evaluating each of their games, and see the results in Figure 2. We see that this adjustment immediately balances out the bulk of the inequity. Three of the host nations performed slightly better than expected, but Brazil performed notably worse than expected. Each host only represents a small sample of games, and we have not found any significant evidence that causes us to assume that the advantage of being tournament host differs from usual home field advantage. Thus, we select the most intuitive option, and simply apply the same home field advantage constant used for qualification stages to account for the advantage of being a host nation.

4 Model Analysis

In Section 3 we outlined the myriad specifications we can use to tune the parameters of our Elo rating system. The primary challenge is that we have limited tools to evaluate the

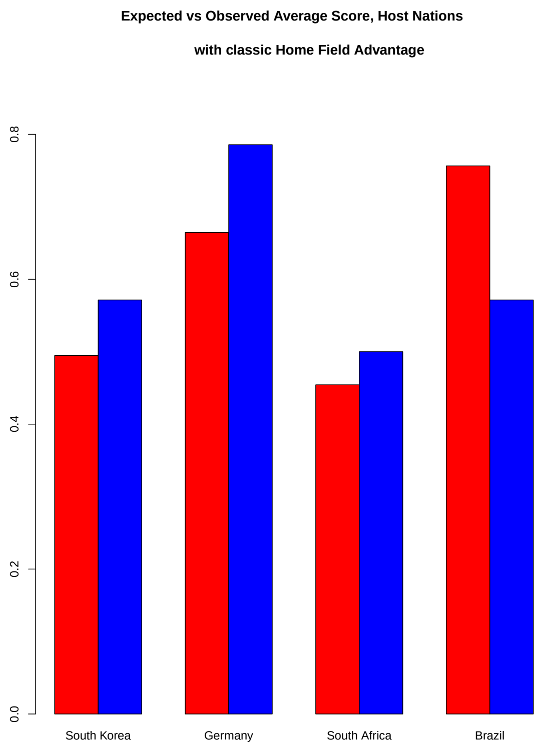


Figure 2: Mean Expected Score: 🟥
Mean Observed Score: 🟦
$c_h = 90$.

effectiveness of the resulting model. In Section 4.1, we will select a collection of tuning parameters that provide reasonably predictive results for the 2014 group stage, by balancing how well they predict the observed results of the group stage and how well they mirror pre-tournament betting odds. In Section 4.2, we evaluate these choices of tuning parameters with a broader approach of determining which Elo rating system is behaving appropriately (not just for the 2014 group stage, but for all recent games). This allows us to select the final model that we use to construct our predictions for the 2018 group stage.

4.1 Tuning Parameters Search

We can make our initial parameter selection by evaluating the predictive efficacy of our model on the 2014 group stage data, running our our Elo ratings on all matches played before it begins. We evaluate the results by comparing them to both the observed results and the pre-tournament betting predictions (the rationale behind this is explained in Section 3.4). For observed results, we calculate the mean squared error from the Expected Score (as calculated by Equation 1 using our final Elo ratings) and the Observed Score (a 1 for a win, a 1/2 for a draw, and 0 for a loss). We call this “Results Error”. Next, we compare our predictions to the pre-tournament betting odds by calculating the mean squared error among each of the three outcomes as predicted by our conversion from Expected Score to outcome probabilities (we call this “Betting Error”). We note that Results Error does not depend on how we determine draws (as it only focuses on Expected Score), while Betting Error is dependent on our choice of draw probability function. Again, it is crucial to note that these are simply methods of estimating our model validity, and neither of them represents “error on a test set” in the classical sense. However, they can still provide us with valuable intuition about our choice of model parameters (and we analyze the models on the entire dataset in the following section).

We run our Elo rating model on the dataset of all matches prior to the 2014 group stages, with a large combination of parameter inputs. We select a sequence of reasonable values for each tuning parameter and consider every combinatorial combination of these parameters, and this defines the parameter space that we search through. Specifically, the parameters that we vary are as follows.

1. $k_b \in \{5, 10, 20, 40\}$, the base constant for our K.

2. $c_h \in \{0, 40, 90, 150\}$, the added Elo rating for home field advantage.

3. $c_q \in \{1, 1.5, 2, 4\}$, the tuning parameter for increased weight to main event matches.

4. $c_t \in \{1, 1.5, 2, 4\}$, the tuning parameter for increased weight to recent matches.

5. Using margin of victory as $\{\text{True}, \text{False}\}$, whether we adjust the Observed Score using Equation 4 to incorporate margin of victory.

In total there are 512 combinations in this parameter space which we test. We analyze the results in two ways: we consider the mean Betting or Results error for certain parameter choices (where we average over all other combinations tried, with one parameter choice fixed), and we consider the parameter choices that lead to the lowest Betting and Results Error. In Table 4, we consider the average errors when using or ignoring the margin of victory (MOV), averaging over the results of all other parameters in our space. We can see that there is a noticeable improvement in the predictive accuracy by both metrics when incorporating margin of victory into our Elo model. Our model tends to both be closer to the observed results, as well as a better match for pre-tournament betting odds, when incorporating this extra information. In Table 5, we consider our prediction errors for a variety of levels of home field advantage. Here, we see that not incorporating any home field advantage leads to slightly worse predictions on both fronts, and incorporating a 150 point home field advantage is similarly too extreme. However, between two medium approaches (of 40 points and 90 points), we see equivalent predictions from both. In this case, we make this decision based on how the Elo rating system performs broadly on the dataset, rather than its 2014 group stage predictions. The results may be similarly predictive for the 2014 group stage with a home field advantage of 40 rating points, but in Section 3.5 we see that the resulting model has a home team mean Expected Score of about 0.54 and a mean Observed Score of about 0.59. Thus, we select the home field advantage of 90 which creates a more consistent Elo rating model.

We do not perform the same analysis on c_h, c_t, and k_b, because they are each quite related (as inputs to our K match weighting function), thus averaging over the results of all

Table 4: Mean Errors when using or ignoring margin of victory (MOV), averaged over all other parameters

	Mean Results Error	Mean Betting Error
Use MOV = True	0.169	0.0297
Use MOV = False	0.173	0.0321

Table 5: Mean Errors with different levels of home field advantage (c_h), averaged over all other parameters

	Mean Results Error	Mean Betting Error
$c_h = 0$	0.1714	0.0309
$c_h = 40$	0.1710	0.0306
$c_h = 90$	0.1710	0.0306
$c_h = 150$	0.1718	0.0313

other parameters would lead to muddled results. Instead, we focus on which selections of values lead to the smallest Betting and Results Error. In each case, we use the margin of victory correction (we saw that it made a substantial improvement). The parameter selection P_1 achieved the lowest Results Error, and the parameter selection P_2 achieved the lowest Betting Error (after we adjusted for the fact that we believe $c_h = 90$ is the best option for home field advantage).

1. $P_1 = \{k_b = 5, c_q = 4, c_t = 4, c_h = 90\}$.

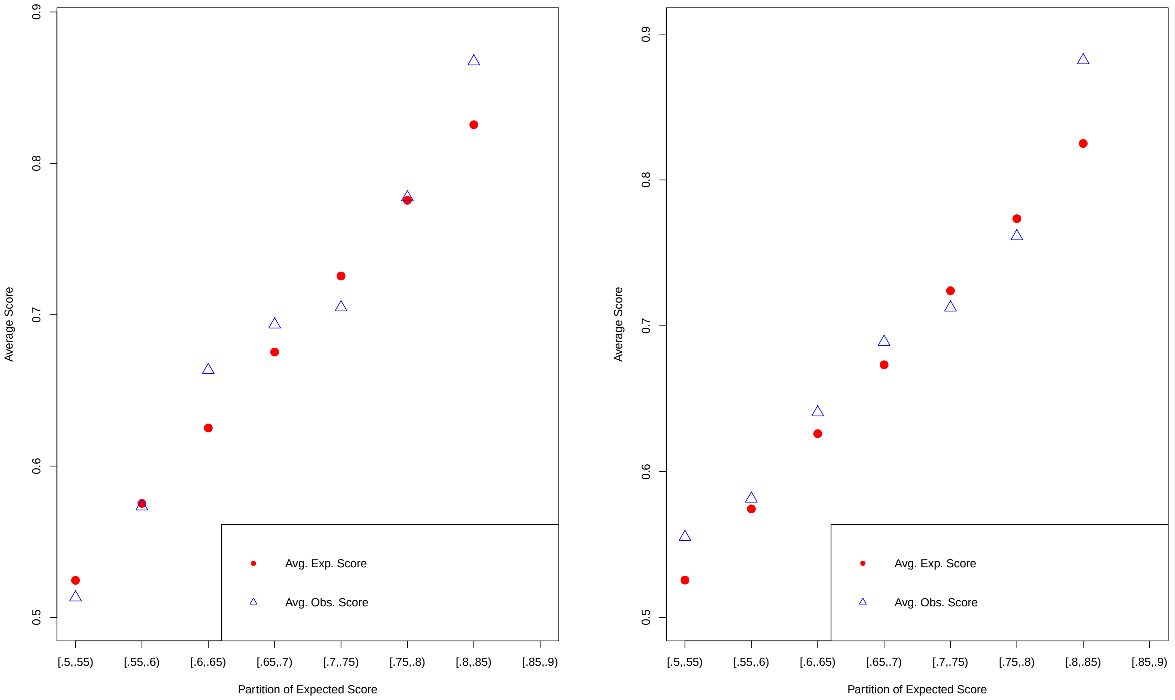
2. $P_2 = \{k_b = 10, c_q = 2, c_t = 2, c_h = 90\}$.

4.2 Elo Ratings Analysis

Our next step is to determine which of these selected model parameters (each of which proved reasonably predictive for the 2014 group stage) lead to a sensible progression of Elo ratings when applied to our full dataset. We need to examine more closely whether our Elo rating predictions are working as intended on the dataset as a whole. In this case, we focus our analysis on all matches from 2010 and onward. This is still a fairly large sample of matches (1651), but gives sufficient time for our Elo rating system to initialize the rating of each team. We do not solely consider the 32 teams that are in the 2018 World Cup group stage, because the Elo rating system will only function if it is reasonably consistent for all included teams (as their results implicitly effect the results of our teams of interest). Our parameter selection P_2 places a bit of extra weight on main event matches and more recent matches, and our parameter selection P_1 places greater weight on these matches, while reducing the base weighting of matches to compensate.

There is no single metric to evaluate the compatibility of an Elo rating system, but we can delve into some choice details to see that it is working as it should. We partition our set of matches based on the Expected Score produced by the given model (we consider Expected Score from the perspective of the pre-match favorite, so Expected Score $\in [1/2, 1)$), Using

Expected Score partition boundaries of $[0.5, 0.55), [0.55, 0.60), \ldots, [0.85, 0.9)$ (after this point, there is minimal data), we take the mean Observed Score among all matches in each partition and compare it to the mean Expected Score in that partition, plotting the results in Figure 3. We consider two competing models, using the parameter choices P_1 and P_2 respectively.



(a) $P_1 = \{k_b = 5, c_q = 4, c_t = 4, c_h = 90\}$.

(b) $P_2 = \{k_b = 10, c_q = 2, c_t = 2, c_h = 90\}$.

Figure 3: Plotting mean Observed Score of matches partitioned by Expected Score.

We hope to see that among a partition of roughly similar matches (determined by our model's confidence in the victor), that the mean Observed and Expected Scores are roughly the same, showing that for this category of matches our model avoided systemic bias. We can see that the P_1 model follows the right trend, but seems to underrate the chances of moderate underdogs (with an expected score of $[.6, 0.65)$), which is a very common type of match. In comparison, the P_2 model has a more consistent performance, as the discrepancies are small and without obvious trend. The notable weakness shared by both models is that they tend to underrate the chances of extreme underdogs (the $[.85, .90)$ partition). This is an unfortunate trait for our model, but it is not terribly surprising. Extremely lopsided matches often involve a smaller island nation, or a new addition to FIFA. These teams tend to be far less consistent than top nations with established infrastructure, making prediction difficult. Further, we note that most of the extreme underdog matches came from the 2010 World Cup, and very few from 2018, showing that our model was likely still adjusting itself at the time, and that this should be not as pressing a concern in our 2018 group stage predictions.

It is informative to compare our successful P_2 model to what a poor fit of the data looks like. In Section 2.1, we describe how we initialize our Elo ratings using the October 2001 FIFA rankings. It is unfortunate to rely on an outside source with questionable predictive power, but we stated that the dataset was not sufficiently large and the matches not sufficiently mixed to initialize all ratings to begin at 1000. In Figure 4, we use the same parameters as P_2, but we initialize all teams to begin at 1000 Elo rating in 2002 instead. We immediately see that even after giving our model the 2002 and 2006 World Cups to adjust, it is an extremely poor predictive fit. We need this rating initialization or else our dataset is insufficient for our model to be a reasonable approximation.

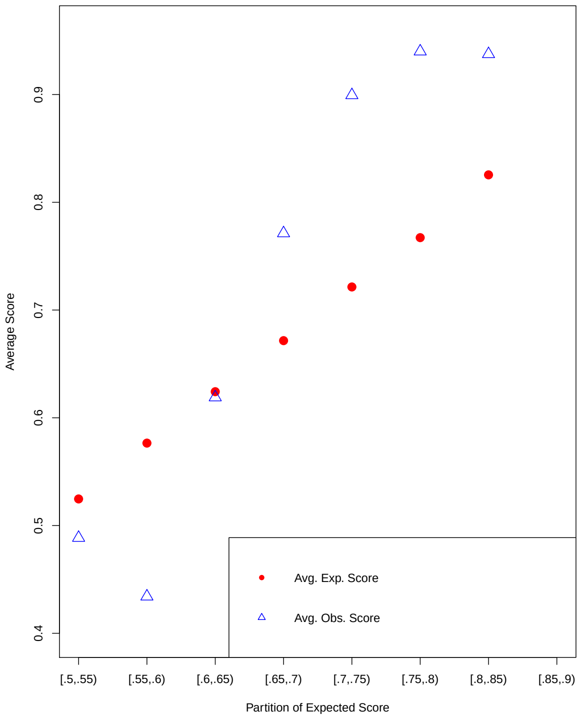


Figure 4: Plotting mean Observed Score of matches partitioned by Expected Score, when we initialize Elo ratings at 1000.

Our analysis has focused on the comparison of Expected Score to Observed Score, which we note does not directly depend on our consideration for the likelihood of draws, discussed in Section 3.1 (as Expected Score is implicitly comprised of the likelihood of a win and a likelihood of a draw). In Equation 3, we can choose a final tuning parameter k_d to adjust the correspondence from Expected Score to a likelihood of a draw. We try a variety of tuning parameter values, and compare the resulting draw percentages (first among all matches in the dataset, and then among just the group stage matches) to the observed draw percentages. As previously noted, the group stage matches tend to be more evenly matched than qualification matches, so there are more draws. We are particularly interested in predicting group stage data, but that is a quite small sample so we also hope to match the overall draw percentage. In Table 6, we decide that the best estimate is $k_d = 0.5$. This slightly underestimates the overall draw percentage, and slightly overestimates the draw percentage in group stage matches. We can use this parameter in the calculation of 2018 group stage probabilities (indeed, we see the mean probability of a draw in 2018 is projected to be about 0.249, which is right about where we want it to be).

Table 6: Comparison of % of Draws, among group stage matches and overall, for different k_d parameter specifications

k_d Parameter	Draw % : Overall	Draw % : Groups
0.3	0.242	0.263
0.5	0.218	0.247
0.7	0.198	0.233
0.9	0.181	0.220

Obs. Draw % : Overall
0.226
Obs. Draw % : Groups
0.240

5 Conclusion & 2018 Group Stage Projections

We use these final selection parameters to predict the probabilities of match outcomes for the 2018 group stage, using the Elo rating system described here for all World Cup matches beginning in 2002. We initialize all Elo ratings by scaling the FIFA rankings in October 2001. For our scaling K factor, we use tuning parameters of $c_q = 2, c_t = 2$, and $k_b = 10$ (see Equation 3.3). We incorporate margin of victory into our scoring system (see Equation 4). We select the tuning parameter for the likelihood of draws $k_d = .5$ (see Equation 3). We assume that home teams have a rating advantage of 90 points, and that this applies for World Cup hosts as well, and the resulting outcome probabilities are shown in Table 7. For the table of final Elo ratings prior to the 2018 World Cup group stage used to calculate these probabilities, see Appendix A (Table 8).

Team 1	Team 2	$\mathbb{P}\{\text{Team 1 Win}\}$	$\mathbb{P}\{\text{Draw}\}$	$\mathbb{P}\{\text{Team 2 Win}\}$
Russia	Saudi Arabia	0.562	0.233	0.205
Russia	Egypt	0.566	0.231	0.203
Russia	Uruguay	0.462	0.262	0.276
Saudi Arabia	Egypt	0.360	0.289	0.351
Saudi Arabia	Uruguay	0.278	0.262	0.460
Egypt	Uruguay	0.274	0.261	0.465
Portugal	Spain	0.306	0.273	0.420
Portugal	Morocco	0.586	0.225	0.189
Portugal	Iran	0.527	0.243	0.230
Spain	Morocco	0.641	0.207	0.152
Spain	Iran	0.587	0.225	0.188
Morocco	Iran	0.307	0.273	0.420
France	Australia	0.536	0.241	0.224
France	Peru	0.583	0.226	0.191
France	Denmark	0.447	0.266	0.287
Australia	Peru	0.407	0.277	0.316
Australia	Denmark	0.287	0.266	0.448
Peru	Denmark	0.250	0.251	0.498

Argentina	Iceland	0.635	0.209	0.156
Argentina	Croatia	0.516	0.246	0.237
Argentina	Nigeria	0.583	0.226	0.191
Iceland	Croatia	0.257	0.254	0.488
Iceland	Nigeria	0.309	0.274	0.417
Croatia	Nigeria	0.428	0.271	0.301
Brazil	Switzerland	0.560	0.233	0.206
Brazil	Costa Rica	0.556	0.234	0.209
Brazil	Serbia	0.571	0.230	0.199
Switzerland	Costa Rica	0.352	0.289	0.360
Switzerland	Serbia	0.367	0.287	0.346
Costa Rica	Serbia	0.371	0.286	0.343
Germany	Mexico	0.536	0.240	0.224
Germany	Sweden	0.560	0.233	0.207
Germany	South Korea	0.698	0.187	0.115
Mexico	Sweden	0.381	0.283	0.335
Mexico	South Korea	0.547	0.237	0.216
Sweden	South Korea	0.523	0.244	0.233
Belgium	Panama	0.608	0.218	0.174
Belgium	Tunisia	0.488	0.255	0.258
Belgium	England	0.308	0.274	0.418
Panama	Tunisia	0.259	0.255	0.486
Panama	England	0.140	0.201	0.659
Tunisia	England	0.216	0.237	0.547
Poland	Senegal	0.397	0.279	0.324
Poland	Colombia	0.204	0.232	0.563
Poland	Japan	0.314	0.276	0.411
Senegal	Colombia	0.179	0.221	0.600
Senegal	Japan	0.283	0.264	0.452
Colombia	Japan	0.512	0.247	0.240

Table 7: Final projected probability outcomes for each game in the 2018 group stage.

One possible addition to our approach would be to incorporate national team data from other competitions, such as the Euros or Copa America. This would not necessarily be a simple addition, because the matches are not played in as standard a fashion as the World Cup (for instance, many teams have no such major tournaments, and national teams may approach them differently). However, if done with care, this could help treat the issue of the large time gap between the end of one World Cup and the start of qualifying for the next (which might mask large shifts in team strength).

Ultimately, a pure statistical approach to predicting match results is unlikely to compete with one that successfully blends quantitative models with deep hands on knowledge of the minutiae of sport. In this report, we temper our analysis when possible with intuition and other checks, but there are limits to what can be determined through match results alone.

Successful analysis will incorporate the impact of injuries and roster changes. Obviously, this is no simple task, given that there is no surefire way to understand the impact of a roster change on a national team (indeed, questions like these are worth millions of dollars to soccer clubs, and have no certain answer). When betting markets open for the 2018 group stage matches, we will see analysts with deep knowledge of the sport hold strong and directly conflicting opinions on the likelihood of the results. The best way to improve this approach would be to analyze the actual recent play of each team to give the match results further context. However, analysis of soccer play is an imprecise art, and our approach provides a reasonable baseline prediction that fits with observed past results.

References

[1] www.fifa.com/worldcup/news/y=2017/m=11/news=the-final-draw-how-it-works-2921565.html

[2] Mark Glickman and Albyn C. Jones. “Rating the chess rating system.” Chance-Berlin then New York – 12 (1999): 21-28.

[3] The Rec.Sport.Soccer Statistics Foundation. www.rsssf.com/

[4] Elo package in R, created by Ethan Heinzen. https://cran.r-project.org/web/packages/elo/

[5] Hughes, Rob. “Players get a Call they Can’t Ignore from National Teams”. New York Times. 2015. www.nytimes.com/2015/01/07/sports/\soccer/players-get-a-call-they-cant-ignore-from-national-teams.html.

[6] www.fifa.com/fifa-world-ranking/ranking-table/men/index.html

[7] www.progressivebetting.co.uk/statistics/football_statistics/leagues_by_draws/

[8] pena.lt/y/2015/12/12/frequency-of-draws-in-football/

[9] www.oddsportal.com/soccer/world/world-cup-2014/results/

[10] “The 12th Man”. The Economist. 2014. www.economist.com/blogs/gametheory/2014/06/home-advantage-football..

[11] Steve Denning. “The Home-Team Advantage In The World Cup – And Beyond”. Forbes Magazine. 2014. www.forbes.com/sites/stevedenning/2014/06/13/the-home-team-advantage-in-the-world-cup-and-beyond/#34be50223ff1

[12] www.fifa.com/worldcup/news/y=2017/m=11/news=the-final-draw-how-it-works-2921565.html

A Appendix: Final Elo Table

Table 8: Final Elo rating table prior to 2018 World Cup for qualified teams

Team	Elo Rating
Germany	1413.74
Brazil	1375.23
Spain	1351.30
Argentina	1338.77
France	1321.57
Portugal	1311.52
Colombia	1310.96
England	1303.28
Mexico	1301.45
Sweden	1285.50
Denmark	1265.82
Belgium	1265.15
Uruguay	1257.39
Costa Rica	1249.41
Switzerland	1246.65
Serbia	1239.61
Croatia	1239.21
Russia	1232.75
Japan	1214.17
Australia	1209.51
Iran	1204.75
Nigeria	1194.97
Saudi Arabia	1193.30
Egypt	1190.45
Tunisia	1183.82
South Korea	1181.74
Poland	1180.45
Peru	1177.72
Morocco	1165.22
Iceland	1157.48
Senegal	1154.91
Panama	1103.44