Using Forecasting to Detect Corruption in International Football

J. James Reade
University of Birmingham
The Johns Hopkins University, SAIS Bologna Center

Sachiko Akie
Akita International University

May 15, 2013

Abstract

Corruption is hidden action aimed at influencing the outcome of an event away from its competitive outcome. It is likely common in all walks of life yet its hidden nature makes it difficult to detect, while its distortionary influence on resource allocation ensures the importance of trying to detect it both practically and economically. This paper further develops methods to detect corrupt activity contained in Olmo et al. (2011) and Reade (2013) that make use of different forecasting methods and their information sets to detect corruption. We collect data from 63 bookmakers covering over 9,000 international football matches since 2004 and assess a claim made in early 2013 by *Europol* that the outcomes of almost 300 international matches since 2009 were fixed. Our collected data consists of match outcomes and pre-match bookmaker odds, which we use to explore the divergence between two kinds of forecasts of match outcomes: those by bookmakers, and those constructed by econometric models. We argue that in the absence of corrupt activity to fix outcomes these two forecasts should be indistinguishable as they are based on the same information sets, and hence any divergence between the two may be indicative of corrupt activity to fix matches. Such an assertion is conditional on the quality of the econometric model and in this paper we discuss the peculiarities of modelling international football match outcomes. In the absence of corroborating evidence we cannot declare any evidence procured in our manner as conclusive regarding the existence or otherwise of corruption, but nonetheless we argue that is it indicative. We conclude that there is mild evidence regarding potentially corrupt outcomes, and we also point towards yet more advanced strategies for its detection.

JEL Classification: D73, C53, D83.
Keywords: Corruption, Forecasting Models, Information and Knowledge.

1 Introduction

Corruption is hidden action aimed at influencing the outcome of an event away from its competitive outcome. It likely occurs in all walks of life yet its hidden nature makes it difficult to detect, while its distortionary influence on resource allocation ensures the importance of trying to detect it both practically and economically. Practically, resources are diverted from participants in the events to those seeking to influence them. As those seeking to influence them are doing so for financial gain, this falls under the purview of fraud, since those fixing matches gain most through keeping information regarding the fix as private as possible in order to place bets on the fixed outcomes. In the context we consider, namely a sports league, the uncertainty of outcome is a particularly valued aspect of the output being produced; the *uncertainty of outcome* hypothesis of Rottenberg (1956) relates this uncertainty to the revenues generated by a sports league, and hence attempts to reduce this uncertainty must be harmful and thus there is an economic interest to ensuring corruption is detected.

This paper further develops methods to detect corrupt activity contained in Olmo et al. (2011) and Reade (2013) that make use of different forecasting methods with different information sets to detect corruption. We collect data from over 9,000 international football matches since 2004 and consider one specific recent episode of alleged match fixing.[1] We assess a claim made in early 2013 by *Europol* that the outcomes of almost 300 matches between 2009 and 2012 were fixed.[2] Our collected data consists of match outcomes and pre-match bookmaker and betting exchange odds, which we use to explore the divergence between two kinds of forecasts of match outcomes: those by bookmakers, and those constructed by econometric models. These two forecast models differ in their *information sets*; the econometric model, covering a large number of matches, by necessity must consider an information set that is broad but not particularly specific to each individual match, whereas the forecasts of bookmakers for each individual match may reflect local information unobserved by the econometrician (e.g. injuries, weather conditions, mental conditioning of players), and which may be dismissed as statistical noise even were it observed on a broader scale. One particular aspect of that *local information*, however, must be information regarding the potential fixing of outcomes of individual matches. We argue that in the absence of corrupt activity to fix outcomes these two forecasts should be statistically indistinguishable, and hence any systematic divergence between the two *may* be indicative of corrupt activity to fix matches. In the absence of corroborating evidence (which currently is private information in the possession of *Europol*), we cannot declare any evidence procured in this manner as conclusive regarding the existence or otherwise of corruption, but nonetheless we argue that is it indicative.

The focus in this paper is match outcomes, yet we recognise that any soccer match has a number of different sub-outcomes alongside the final match outcome (the final score in terms of goals scored by each team). In recent years as the betting market has expanded, the variety of such sub-outcomes that can be bet on has increased dramatically. For example, the total number of goals scored, the score at half time, the timing of goals (and other incidents like the first throw in) and identity of goalscorers, the margin of victory and many others. Hill (2010) makes a persuasive case that indeed this is most likely the most fertile area for match fixing — such sub-outcomes are less well observed and hence the probability of detection must fall, whilst the obscurity of such sub-outcomes may mean that betting exchange markets for them are illiquid and bookmaker markets inefficiently priced affording larger potential profits from fixing. The website from which we collect our data, $w$ww.OddsPortal.com, contains odds on many of these different outcomes and hence the practicality of investigating many other aspects of match outcomes for corruption exists; simply one must develop a method to statistically model that particular outcome also, if our strategy is to be followed. This we relegate to future research.

We note the practical relevance of our work. In response to widening concerns that corruption is harming the commercial interests of sport, various governing bodies are devoting ever increasing resources towards its detection. For example, Betfair has signed agreements with multiple governing bodies to share information on suspicious trading patterns detected on its markets, and employs a team of analysts to detect such market movements.[3] It is our hope that the method of comparing information sets contained within this paper can be of use in such detection attempts. While we only make use of already public information in this paper, we argue that the method is important as a means of detecting corruption, and comes through applying economic theory to the question of corruption.

In Section 2 we review the existing literature on forensic economics, in Section 3 we introduce our data on bookmaker odds from over 9,000 football matches, and in Section 4 we describe our econometric method, assessing in particular the forecast performance of both the econometric model developed and bookmakers recorded, and carry it out, presenting our results along the way. Section 5 concludes.

[1] By international match we refer to matches involving teams representing countries. This is distinct from matches between club teams from different countries. There are a small number of friendly matches in our dataset involving club teams from countries, and national teams of other countries.

[2] See Harris (2013) and Hill (2013) regarding this. A press conference by *Europol* released a mix of new and old information regarding many matches known to have been fixed in recent years throughout Europe. It was later clarified that of the 700 matches mentioned, 300 were new, and 90% of these new matches were international matches.

[3] See 'Anti-corruption: Technology key to catching fixers' *Financial Times*, 16 June 2011, (last accessed 24 April 2013, http://goo.gl/P0kTc).

2 Using Forecasting to Detect Corruption

The field of forensic economics is expanding rapidly; as Zitzewitz (2012) notes, the aim of this field is "uncovering evidence of hidden behaviour in a variety of domains", and already in a short number of years insights gained from economic theory regarding hidden action have facilitated empirical investigations in a wide range of areas.

Particularly relevant in the case of sports corruption are papers by Price and Wolfers (2010), Wolfers (2006) and Reade (2013). The first two consider hidden action in basketball on the part of referees and teams, and the latter considers Italian soccer and a recent match fixing scandal there. In all cases, economic theory is brought to bear to determine potentially effective channels upon which to test for the presence of corruption. Preston and Szymanski (2000) analyse the economic theory behind cheating in sport, paying particular attention to the subjective decision making process of the sports participants considering corrupt activity, borrowing from Becker (1968). They note that corrupt activity must alter the *objective probability* of any particular outcome of the sporting contest, and that the likelihood of such activity varies depending on the renumeration of participants, the importance of the individual match taking place, the likelihood of punishment and the severity of punishment.

In Reade (2013), the strategy chosen is to make use of public information available via bookmakers on football matches. Specifically, in the absence of any systematic method to influence matches (which must be private information for some subset of agents involved in a match), the forecast of a match outcome (which we can assume has true probability p_t) by bookmakers, $\hat{p}_{B,t}$ ought to be indistinguishable from that of an econometric model, $\hat{p}_t$, suitably specified. This assertion is based on the idea that most relevant information for predicting the outcome of a match is observed: the strengths of teams are observed via previous matches. These information sets are common to both econometric models and bookmakers in forming predictions.

In the presence of corrupt activity to fix the outcome of a football match, the objective probabilities of match outcomes are altered, say to $p_t^* = p_t + q_t$, and it may thus be that information sets differ between bookmakers and econometric methods. The advent, in particular, of betting exchanges like Betfair means that private information regarding events like fixed matches can become public information if those holding private information trade based upon that knowledge. Specifically, such trading drives the implied betting exchange forecast probability much higher for the particular outcome that has been arranged, aiding prediction market accuracy, as Hanson and Oprea (2009) propose. It is generally regarded that information appears first on prediction markets before being absorbed by bookmakers (see, for example Croxson and Reade, 2011), and hence it might be expected that bookmaker prices will follow exchange prices and thus reflect the additional information regarding the fixed match outcome. If the corrupt action to fix the game is efficient, it will have a significant effect on the match outcome and hence represent an outlier in econometric terms. Hence we should expect to observe a significant difference between the bookmaker price and the econometric model price in the presence of corrupt activity due to the *difference in information sets*: the econometric model contains no information on any fix and hence forecasts p_t while bookmaker prices, in principle, do contain this information and forecast p_t^*. This is the premise of our information-based forecasting test for the presence of corrupt activity.

From forecasting theory, basing forecasts based on larger information sets must yield an improvement, although Hendry (2011) notes this is only in the variance rather than bias. We anticipate given the nature of the forecasts we study and the insight of Preston and Szymanski (2000) that corrupt activity significantly changes probabilities of outcomes that forecasts based on subject information will be biased.

3 Data

Our dataset consists of all international football matches listed on the betting odds website *www.OddsPortal.com*.[4] These matches are categorised into various regional competitions for national teams (e.g. European Championships, Asian Cup), and global events such as friendlies and the World Cup. Furthermore, most nations will have both mens and womens' teams and also youth teams (for those under the ages of 17, 19, 20 and

[4] Information from *O*ddsPortal.com was scraped using the Python programming language over 16–17th February 2013.

21 most commonly). In total, since 2004 we have 9,567 matches involving 35 different tournaments plus friendlies.[5] Of those matches, around 32% involve youth teams and 12% involve womens' teams.[6] Overall we have matches involving 915 teams from around 212 national teams from around the world.[7] While it might be *a priori* anticipated that the fixed matches identified by Europol are all mens senior matches, this is not established, and hence it makes sense to consider all international matches rather than simply restrict ourselves to mens senior matches. The salaries paid to youth players are often dramatically less than for senior players, and a significant gender pay gap undoubtedly exists in football, ensuring that considering such matches provides variation along one dimension identified by Preston and Szymanski (2000) as contributing towards the corruption decision by a sports participant.[8]

For each match we have, on average, 24.7 bookmaker prices, with a standard deviation of 15.6 bookmakers, a maximum of 63 and a minimum of 1 bookmaker.[9] Over all our matches around the world, we have bookmaker prices from 63 different bookmakers, all of whom are listed in the Appendix with the relative frequencies with which they appear in our dataset. We simply take the match outcome probabilities from *w*ww.OddsPortal.com, but usually many other types of bets exist. The choice of only match outcome prices is rather arbitrary and it is more than likely that those seeking to fix outcomes attempt to fix particular aspects of a match rather than necessarily its outcome indicating that it may be important to collect prices on other match outcomes in order to further detect corrupt activity.

4 Methodology and Results

We adopt a forecasting test based on a difference in information sets as our methodology for detecting corruption in international football. In this section we describe this method in much more detail. An important part of this procedure is constructing an econometric model for forecasting. While the information set exists upon which to create a forecast, it is important that we construct a model that effectively utilises that information set. We firstly discuss the econometric model we will use to construct forecasts before turning to the nature of the comparison.

All datafiles used in the regression models, and all codes files are available online.[10]

4.1 Modelling International Football Matches

International football consists of matches between national teams, rather than between club teams based on particular countries, and such matches are thus often high profile and prestigious.[11] As such, the allegation that such a considerable number of these matches have been fixed in recent years is important. It seems more than likely, given the insights of Preston and Szymanski (2000), that international matches are a target for corrupt activity. For example, at international level a large number of friendlies are played, upon which little rests for each team involved. Additionally, in many qualification tournaments because only the top one or two teams in a group can qualify, a large number of less meaningful matches occur between teams unlikely to qualify. At major tournaments such as the World Cup, economic theory would dictate that it is less likely that matches are fixed since the prize at stake tends to be large, and the international audience

[5] See Table 6 on page 17 in the Appendix for a breakdown of the tournaments and the number of matches in each tournament.

[6] And around 4% are womens' youth tournaments.

[7] Although there are only 193 members of the United Nations, a number of non-sovereign states have teams that participate in national championships, such as Wales, Scotland and Northern Ireland, as well as particular regions of other countries such as the Basque Country in Spain. See http://en.wikipedia.org/wiki/List_of_FIFA_country_codes for a list of FIFA members and non-member national teams..

[8] See 'England women footballers secure central contract increase', *BBC Sport*, 15 January 2013 (last accessed 24 April 2013, http://goo.gl/wxJDA) on the gender pay gap, and 'Survey reveals footballers' wages', *BBC Sport*, 11 April 2006 (last accessed 24 April 2013, http://goo.gl/99M3S) on youth salaries.

[9] Our dataset does include matches in which bookmakers declined to offer prices, or in which they withdrew prices.

[10] The workpage for our corruption research can be found at http://goo.gl/cNPwt.

[11] In our sample of international friendlies, a handful of club teams do appear as occasionally higher profile club teams will play friendly matches against national teams. We do not omit these matches since they help provide information on the strength of a national team.

vast.[12] However, Hill (2010) documents the fixing of a quarter final match at the 2006 World Cup, a match that *a priori* had a strong favourite (Brazil vs Ghana) but nonetheless in principle could have been a tight match, such is the uncertain nature of football. Brazil were comfortable winners in the match, 3-0, yet the suspicious mind might question some of the decision-making throughout the match by Ghanaian players, particularly relating to the second and third goals scored by Brazil. However, what may appear suspicious can just as easily be explained within the realms of good or bad performance in an activity undertaken under a great deal of pressure, mentally and physically, and scrutiny. It is this which makes detecting corrupt activity difficult.

Turning to the modelling of international football match outcomes using econometric methods, while domestic football is organised into leagues of teams of similar strength, and this league structure dominates, international football has no such league distinction. National teams are composed of players qualified to play for that country (either by birth or by transferring nationality), whereas domestic football teams can be composed, in principle, of players from any country. National teams also play much more infrequently; only thirteen nations play more than 100 matches in our sample covering nine years, showing that on average national teams play at most on average 14 times per year whereas domestic football teams will play in the region of 30–60 matches per calendar year. Domestic leagues enable a simple way of assessing team strength from an econometric point of view: Each team's performance in that league. The closest in international football to a league is the qualification stages of World Cup and regional championships such as the European Championships, however even these stages are seeded such that the better teams have a greater likelihood of qualification for the latter stages meaning that teams of vastly differing qualities can meet in such mini-leagues. Indeed, McHale and Scarf (2006) make particular reference to this phenomena when studying international soccer matches relative to domestic ones. International matches also differ from club matches in the number of friendly matches that occur — something mentioned earlier in the context of match fixing.

The consequence of this lack of a common framework upon which to judge national teams (no single league, and a high variance of opposition quality) is that some other method is required to rank teams in order to construct statistical or econometric predictions. Fifa rankings could be used to attempt to approximate team quality and thus predict outcome, yet Fifa's rankings are but one attempted measure of team quality and hence have their critics, and furthermore would require additional data collection and matching with actual results.[13] An alternative is to make use of the Elo ranking system devised specifically for chess but adapted for numerous other sports. This ranking system updates for each match, affords the ability to relatively weight different types of matches differently, and provides a simple way to generate predicted outcomes for matches. We make use of Elo rankings to create a variable with which we use to help predict match outcomes. In the academic literature Hvattum and Arntzen (2010) test Elo ratings against bookmakers and econometric models as a forecast tool for English Premier League football matches, finding that bookmakers outperform Elo ratings, but Elo ratings are superior to econometric models, while Leitner et al. (2010) use Elo ratings amongst other methods when attempting to forecast outcomes from the 2008 European Championships football tournament. As Elo ratings are but one additional method for measuring team quality, it is helpful to get some idea about how effectively they do this. In Appendix B we carry out a small simulation study to investigate the properties of the measure. We find that while some biases do exist, these are all away from the mean, implying that some Elo predictions may underestimate the true quality differences between teams. Such biases can be corrected by incorporating Elo predictions into a regression method, as we do.

Additionally, we use a variant of Elo rankings that the *World Football Elo Ratings* (WFEL) employ, which give different weights to different matches.[14] The reason for this is to capture the idea that competitive matches reveal more about the actual quality of a team than friendly matches.

Considering econometric modelling, Goddard (2005) considers the two most common econometric methods for modelling and forecasting football match outcomes, notably Poisson methods to predict goal arrival,

[12] Indeed, Fifa estimates that 3.2bn people watched some part of the 2010 World Cup Final between Spain and the Netherlands (Fifa, 2010).

[13] On the criticism of Fifa rankings, see http://en.wikipedia.org/wiki/FIFA_World_Rankings#Criticism.

[14] See http://www.eloratings.net/system.html and http://en.wikipedia.org/wiki/World_Football_Elo_Ratings for more information.

and direct limited-dependent variable models for actual match outcome, finding that the differences between the two methods are marginal. Forrest et al. (2005) carry out a direct comparison of econometric methods and bookmaker forecasts and find that bookmakers tend to forecast better, something they attribute to greater competition in the betting industry in recent years; our dataset exclusively falls in the more recent period of increased competition amongst bookmakers, as evidenced by the number of bookmakers (63 in total, on average more than 20 per match) we have prices from.

We seek to understand the outcome of a match at time t between team i and team j:

$$y_{ijt} = \begin{cases} 0 & \text{if team } j \text{ wins match at time } t, \\ 0.5 & \text{if match drawn,} \\ 1 & \text{if team } i \text{ wins.} \end{cases} . \tag{1}$$

From (1), match outcome is a discrete variable with three possible outcomes. One standard way to model a variable such as y_{ijt} is to assume there exists a continuous latent variable y^*_{ijt} which, if observed in particular regions implies different outcomes for the observed match outcome variable. We thus write:

$$\mathsf{P}\left(y_{ijt} = 0 \left| \mathbf{X} \right.\right) = \Phi(y^*_{ijt} < \mu_1), \tag{2}$$

$$\mathsf{P}\left(y_{ijt} = 0.5 \left| \mathbf{X} \right.\right) = \Phi(\mu_1 < y^*_{ijt} < \mu_2), \tag{3}$$

$$\mathsf{P}\left(y_{ijt} = 1 \left| \mathbf{X} \right.\right) = \Phi(y^*_{ijt} > \mu_2). \tag{4}$$

The parameters μ_1 and μ_2 are described as the cut-off points — the points in the distribution of y^*_{ijt} where the outcome switches from one of the possibilities to others. Hence below μ_1, the observed outcome y_{ijt} is that team j wins the match (it is listed as an 'away' win), while between μ_1 and μ_2, the match ends in a draw ($y_{ijt} = 0.5$), and above μ_2, team i wins.

We estimate this latent variable y^*_{ijt} using an ordered probit regression model:

$$y^*_{ijt} = \beta_0 + \mathbf{X}_{ijt}\beta + e_{ijt}, \quad e_{ijt} \left| X_{\ ijt} \right. \sim \mathsf{N}\left(0, 1\right). \tag{5}$$

In (5) the variable $\mathbf{X}_{ijt}$ contains explanatory variables for match outcome and in our case includes the relative difference in Elo ratings for the two national teams involved in any given match along with other variables that help describe the historical strength of the two teams involved. Such variables can be generated from each team's historical results; information on match outcomes, ability to score goals and to prevent their concession can all be marshalled into explanatory variables for predicting match outcomes. In international matches, the venue in which the match is played can be important also, and hence we control for matches on neutral territory.

We then use our ordered probit model (5) to generate fitted probabilities for each of the possible events:

- Probability of team1 winning: $\widehat{\mathsf{P}}_{1,ijt} = \mathsf{P}(y_{ijt} = 1 \left| \mathbf{X}_{ijt} \right.)$,
- Probability of a draw: $\widehat{\mathsf{P}}_{D,ijt} = \mathsf{P}(y_{ijt} = 0.5 \left| \mathbf{X}_{ijt} \right.)$,
- Probability of an team2 winning: $\widehat{\mathsf{P}}_{2,ijt} = \mathsf{P}(y_{ijt} = 0 \left| \mathbf{X}_{ijt} \right.)$.

We denote these forecasts as $\widehat{p}_{E,ijt} \equiv \widehat{\mathsf{P}}_{ijt}$. We estimate (5) using data up to the end of 2009, and then use forecasts of all matches that take place from 2010 onwards in order to compare these to bookmaker forecasts.

Considering the ordered probit model for predicting match outcomes, the regression output is provided in Table 1. The explanatory variables are the difference between the Elo expected outcomes for the two teams competing, alongside a number of other readily calculable statistics from previous match outcomes. We include variables for recent performance (wins/draws gained, goals scored, goals conceded) and experience (in sample). While the Elo rating difference is significant, and reflects the impact the difference in quality has on outcome (negative Elo means team 2 is stronger and the dependent variable is zero if team 2 wins), it is not the most significant variable. Instead, the notional 'points' gained by each team per game (three points for a win, one for a draw mirroring the almost universal domestic football league scoring system) is much more significant in explaining match outcomes. The difference between the Elo rating difference and the points gained by a team per game is that the latter does not adjust for opposition quality, while the former does. The significance of Elo ratings is consistent with Hvattum and Arntzen (2010) who suggest Elo ratings provide an improvement over econometric methods.

Testing the normality assumption in (5) in order to assess model specificaiton is tricky as we do not observe the latent variable y^*_{ijt}. Nonetheless, as our purpose is to construct a forecasting model, the best metric upon which to judge our model will be its forecast performance; once we have corrected our bookmaker prices in the next section, we will compare the performance of our econometric model with bookmakers.

When using the output of our econometric model to compare to bookmaker prices later we use forecasts from our regression model rather than fitted values. If we compare fitted values from a regression model estimated over our entire sample from 2004 through to 2013, this would yield an inaccurate test since the econometric model would make use of information after the match in question had taken place. In that case, unusual outcomes would be already factored into the fitted values and hence we may be less likely to spot such distinct outcomes. Hence, given that our focus is the most recent three years of international matches, we estimate our model up to the end of 2009 and then forecast all matches in 2010–13 making use of data available before each match. The only notable change from using fitted values over the entire sample is that rather than our parameters be estimated on data up to 2013, they are estimated on data up to 2009; data up until the start of each match (Elo rating, recent form, etc) is still used to construct forecasts. Assuming that the true process determining the outcome of football matches is stationary, as might be expected, then estimation up to 2009 (which still allows us to estimate over 2,662 matches) should not yield significant differences from estimating up until 2013.[15] In order that the Elo rating for each team is better calibrated hence more informative, we only regress on matches for which each team has already played a minimum of four matches.

4.2 Bookmaker Prices

We anticipate that bookmaker prices will reflect the presence of corrupt activity in football matches, and hence it will be important to consider the various dimensions in which this might materialise. Figure 1 gives some idea of the spread of implied probabilities for each of the three events. In our sample of 9,606 international matches since 2004, 48.4% result in wins for team1 (the first team listed, which usually is the home team apart from during tournaments played at neutral venues), 20.1% result in a draw, with 31.75% ending in a win for team2 (the second team listed, usually the away team). The distribution for the draw is much more concentrated on the lower range of the interval, with the 99th percentile falling at 30.3% compared to 89% and 87% for team1 and team2 victories respectively.

A number of observations in our sample, 528, have implied probabilities for the draw of greater than or equal to a half, which is five standard deviations away from the mean implied probability for a draw. These observations correspond to 216 matches, and about 7% (39) are in the African Cup of Nations, 17% (94) are from the AFC Championships Under-16s tournament, and around 14% (81) are from European Championship matches. Only 26 (5%) are from friendly matches. However, it is worth noting that a number of these matches correspond to very one-sided matches, such as matches between Germany, the Netherlands or England and San Marino, Andorra or Luxembourg, and hence likely reflect the overwhelmingly likely outcome in matches such as these (England recently beat San Marino 8-0 in San Marino) rather than anything more sinister. As such it is important to consider odds in the context of an econometric model able to distinguish between such match heterogeneity.

Reade (2013) find distinct behaviour in bookmaker reported odds for draw outcomes at times when corrupt activity might be more likely (the end of the season when neither team has anything to play for). In particular, they find that the odds on the draw, which characteristically never yield implied probabilities above a third, often reach two thirds and higher in suspicious cases. While it seems likely that information spreads amongst bookmakers by the time kick-off occurs (and our odds are observed), it may be that it

[15] This is of course something we can check by comparing the regression model for 2004–2009 to one estimated on 2004–2013. We do this and find minimal differences.

	(1)
	outcome
outcome	
ea_team_diff	-0.970***
	(-8.418)
finals	-0.260*
	(-2.540)
pts_last_31	-0.093***
	(-4.779)
pts_last_32	0.063**
	(3.219)
gdiff_last31	0.042**
	(3.278)
gdiff_last32	-0.020
	(-1.509)
experience1	0.008**
	(2.918)
experience2	-0.010***
	(-3.420)
pts_pg1	1.674***
	(19.511)
pts_pg2	-1.783***
	(-20.702)
days_since_last_match1	-0.001
	(-1.708)
days_since_last_match2	0.000
	(0.504)
cut1	
_cons	-0.999***
	(-6.239)
cut2	
_cons	-0.195
	(-1.221)
N	2662

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 1: Ordered probit regression model output for international match outcome.

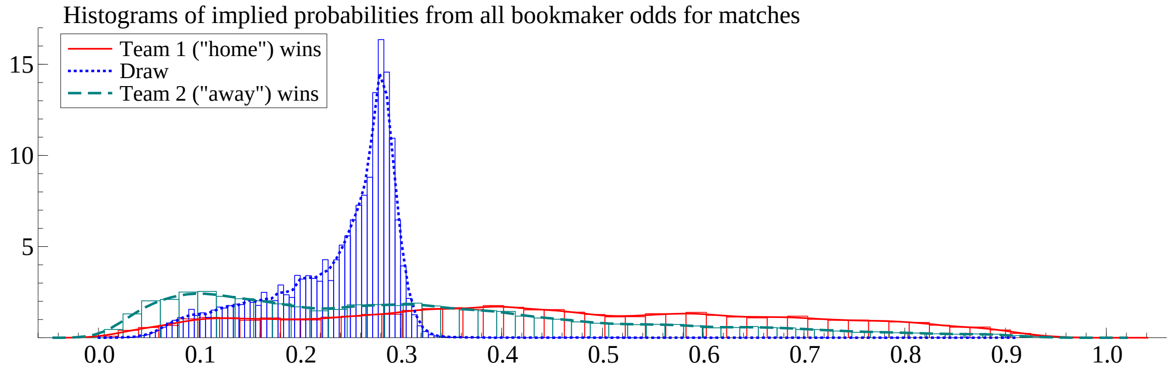

Figure 1: Histograms and estimates of empirical probability distribution of all bookmaker prices for the three events surrounding each football match. Bookmaker prices are corrected for the over-round.

does not do so perfectly, and hence we might observe a greater variation of bookmaker prices when corrupt activity is taking place. However, it is equally likely that variation will increase for non-corruption-induced information changes.

However, it might similarly be that for corrupt matches most (if not all) bookmakers refuse to provide odds and take bets. *OddsPortal* reports a range of bookmakers providing odds, from 63 in the biggest case down to just one bookmaker in a small number of other cases.[16] It could be that the number of bookmakers reporting odds provides some predictive power. It seems likely that the distribution of bookmaker odds as well as simply their mean for any given match may be important, and hence we firstly characterise bookmaker odds over our matches along such potentially interesting dimensions.

Additionally, as mentioned earlier, it might be that particular aspects of matches are fixed meaning that it isn't just odds on a home win, away win or the draw that are affected; more complicated bets like those on spreads and other derivative outcomes may be where the important patterns are that ought to be detected.

Nonetheless, before conducting any characterisation, it is important to correct for the well-known favourite-longshot bias (FLB) in betting markets, whereby favourites win more often than their odds imply, and outsiders (longshots) win less often than their odds imply. Such bias is often corrected for using linear regression methods. Regressing the outcome, say o_t, on the implied probability of bookmaker i's prices for the match at time t, $p_{B,it}$, with a constant:[17]

$$o_{it} = \alpha_o + \beta_o p_{B,it} + u_{it}, \tag{6}$$

yields fitted values $\widehat{p}_{B,it} \equiv \widehat{o}_{it} = \widehat{\alpha} - \widehat{\beta}_o p_{B,t}$ which are the bookmaker odds $p_{B,it}$ corrected for their observed bias.

FLB can be graphically represented by plotting the implied probabilities from bookmaker prices against the frequency with which bets at those odds paid out (i.e. the event in question occurred).[18] We provide such plots in Figure 2; if the data points lie on the 45-degree line (marked with a black dotted line), then there is an absence of any bias and the implied prediction of the bookmaker pays off, on average, as often as the odds imply. For all three markets it is observed that bets at bookmaker odds implying a probability

[16] Our data scraping method collected information from matches where no odds were reported, yet only one match fits such a description, suggesting that although some bookmakers may refuse to set odds for a match they are sceptical about, others still will.

[17] With a slight abuse of notation for o_{it}, since this does not vary over i.

[18] We calculate the implied probabilities by taking the reciprocal of the decimal odds quoted on OddsPortal. This interpretation of odds as probabilities is not without debate; for more on this, see Wolfers and Zitzewitz (2006) and the articles cited therein, in particular Manski (2006)

	(1)	(2)	(3)
	outcomeH	outcomeD	outcomeA
prob1	1.122***		
	(289.214)		
prob2		1.039***	
		(79.103)	
prob3			1.117***
			(283.671)
_cons	-0.030***	-0.038***	-0.032***
	(-15.359)	(-11.943)	(-21.967)
N	240171	240171	240171

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 2: Table containing favourite-longshot bias correction regressions for bookmaker data.

above around 90% or above always pay off, and those events priced implying a probability of less than 3% never pay off.

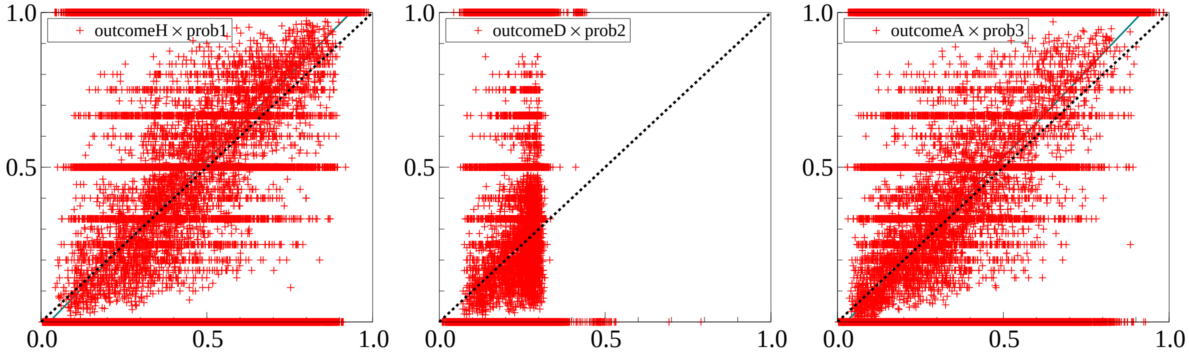

Figure 2: Plots of outcome against implied bookmaker odds for all bookmakers collected in our dataset.

Table 2 provides the output for the regression models used to correct FLB in bookmaker prices. The left hand column is for the team listed first (usually the home team but not always) to win, the middle column for the draw, the final column for the second-team-listed (away) win. All three columns betray a significant favourite longshot bias — the t-statistics in parentheses are sufficiently large that any F test of $\alpha = 0$ and $\beta = 1$ will be emphatically rejected.

4.3 Forecast Comparisons

4.3.1 Brier Score

One measure of the forecast performance of a forecast model in the context of success/failure events like a win, draw or loss is the Brier score, which takes the corrected forecast probability from each forecast model

Outcome	Bookmakers	Model
Team 1	.163	.162
Draw	.159	.153
Team 2	.185	.179

Table 3: Brier scores for econometric model, calculated over the 5,695 matches in our dataset taking place after 2009, compared to bookmaker forecasts.

and compares it to the outcome variable (see (1)):

$$\mathsf{Brier} = \frac{1}{M}\sum_{m=1}^{M}(\widehat{p}_{m,it} - y_{m,it})^2. \tag{7}$$

Table 9 reports Brier scores for both our econometric model and bookmakers, allowing us to compare the performance of the two methods. The scores indicate that on average, both bookmakers and our model were out by about 40 percentage points in their forecasts.

Our econometric model performs indistinguishably differently from the bookmakers when predicting either positive match outcome, and slightly better for the draw.[19]

4.3.2 Comparison of Forecast Differences

While the Brier score yields information on the relative quality of forecasts in general via taking averages, our main focus is on the differences between forecasts in matches we might suspect of corrupt activity, and as such we now consider methods to assess the differences between the forecasts.

Using corrected bookmaker odds $\widehat{p}_{B,it}$, we can compare these to forecasts generated from our econometric model (5). Our interest is in the divergence between the two and hence we run the regression model:

$$\widehat{p}_{B,it} = \alpha_p + \beta_p\widehat{p}_{E,it} + \varepsilon_{it}, \tag{8}$$

and firstly consider the nature of the estimators $\widehat{\alpha}_p$ and $\widehat{\beta}_p$ before investigating the residuals. We theorise that in the absence of corrupt activity, $\widehat{p}_{B,it} = \widehat{p}_{E,it}$ and hence the residuals $\widehat{\varepsilon}_{it}$ ought to be symmetrically distributed around their mean of zero, and we might anticipate $\alpha_p = \beta_p - 1 = 0$. However, the existence of minor biases in bookmaker and econometric forecasts when observed over particular dimensions (such as the draw) is such that we may observe variations in predicted probabilities, despite similar overall predictive performance, and hence to maximise the flexibility of our approach we do not require $\alpha_p = 1 - \beta_p = 0$. While the regression method ensures that $\widehat{\varepsilon}_{it}$ are mean zero, nonetheless the existence of mass in either tail of the distribution may be indicative of suspicious activity, and hence we investigate the residual distribution from our models (for each team's win and the draw markets).

Table 4 presents the results of the regression model (8) for the three possible match outcomes. The regression model is carried out using every single bookmaker probability for a given match, hence we have over 160,000 observations.

We focus on the draw, as Reade (2013) did, as opposed to either positive result. This is because often the draw is considered to be something of a 'residual' event, since most focus is on whether or not each team will win — and naturally, each team in a match sets out to win it in the absence of corrupt activity. As a result, the distribution of bookmaker prices (and as we will see, the predictions of our econometric model), do not venture above about a third, reflecting this residual nature. Furthermore, a draw is something more of a collaborative outcome since each team gains from it, not necessarily only in prestige terms but also in points terms in competitive matches, whereas either positive result is much more non-cooperative. As such,

[19] We conduct matches t-tests for the difference in the two numbers (average squared errors), and find that for the home (team 1) win, the t-statistic is 0.64 ($p = 0.64$) and for the away (team 2) win, the t-statistic is 1.4 ($p = 0.162$), while for the draw the t-statistic is 3.38 ($p = 0.001$). Nonetheless all these differences are all only at the third decimal place; see Table 9.

	(1) Draw	(2) Team 1 wins	(3) Team 2 wins
Model Probability	0.716***	0.694***	0.694***
	(236.673)	(348.189)	(348.825)
_cons	0.080***	0.136***	0.085***
	(122.860)	(125.846)	(108.924)
N	166902	166902	166902

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4: Regressions of bookmaker probabilities on model probabilities in forecast period post 2009.

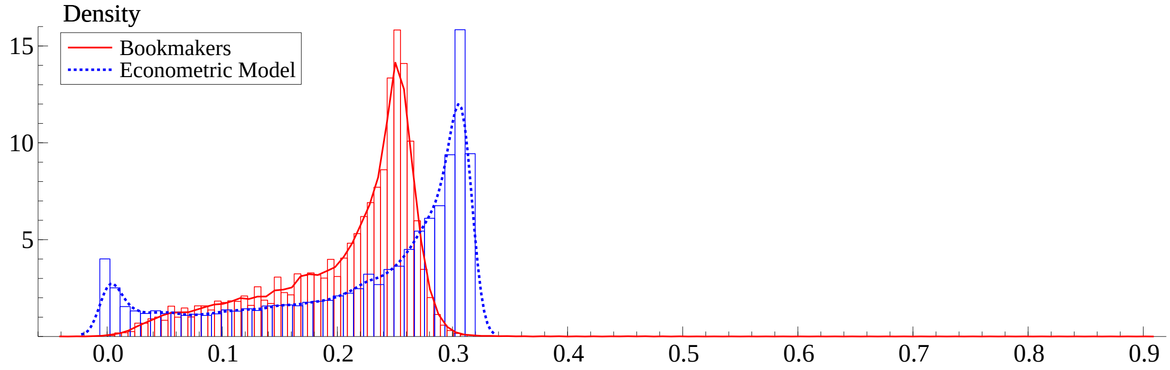

Figure 3: Plot of both bookmaker predictions for match outcomes (corrected), and econometric model predictions.

it seems more likely that distinct patterns in the draw market may be indicative of some kind of collusive activity to ensure such an outcome. Practically, since the implied forecasts extremely infrequently move above a third, whereas the forecasts for either positive result can and do often reach much higher levels, this also enables the spotting of unusual patterns. Thus we focus on the draw outcome.

Figure 4 provides a comparison of the distribution of forecasts for our two models for the draw outcome. The red line is the bookmaker forecasts, and the blue line is our econometric model forecasts. Figure 4 helps explain the regression results for the draw; the coefficient of 0.7 corrects for the wider dispersion of econometric model forecasts relative to bookmakers, and hence any large residual we observe in our model controls for this pattern.

In determining a large outlier indicative of a difference in information sets and hence potential corrupt activity, we use residuals larger than three standard deviations. Assuming a normal distribution, such an observation ought to be observed 0.27% of the time, and hence we might expect to see around 420 in our forecast sample of 167,916 bookmaker prices. For the draw outcome, we actually observe 1,623 such observations from 210 different matches. It should be noted that (8) includes generated regressors on both the left- and right-hand side of the regression equation, creating potential distortions for standard errors and the standard deviation of the residuals used to determine a large outlier. (Wooldridge, 2002, Ch. 6) notes that provided standard OLS assumptions hold, there is no impact on bias and consistency properties for estimators, but for standard errors the sampling variation induced by the first stage regressions can cause problems. Lewis and Linzer (2005) suggest that when a dependent variable is generated that heteroskedasticity robust standard errors can be used to alleviate distortions. To limit the possibility that generated regressors influence the likelihood of a bookmaker price being isolated as an outlier, we report information regarding the nature of relatively large residuals based also simply on residuals, rather than a binary variable representing ‘large residuals’.

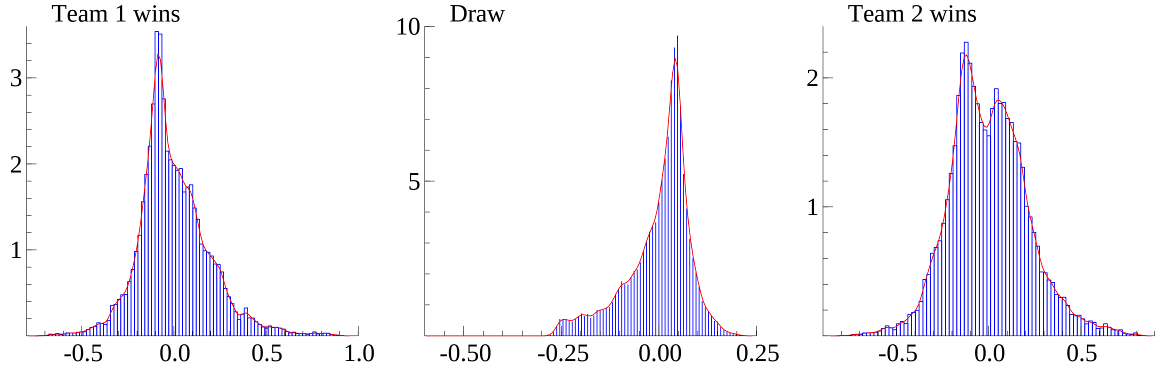

Figure 4: Histogram of residuals from the team 1, draw and team 2 win regressions of model probabilities on bookmaker probabilities.

A graphical inspection of the residuals we seek to investigate will be helpful; Figure 3 provides this for all three match outcomes. Ideally, all three distributions will be unimodal, symmetric and centred on zero — this would reflect that any departures between our econometric model and bookmakers are random. From Figure 3 however it is clear that all three distributions, but most notably the draw distribution, appear to be non-symmetric and centred away from zero. The draw distribution is particularly distinct since the mode of the distribution is positive suggesting that most often bookmakers underprice the draw relative to our econometric model, but nonetheless 40% of the distribution is negative, where bookmakers predict a draw with higher probability than our model. Indeed the largest residuals are negative and are at least three standard deviations from the mean (and four from the mode), and this would be consistent with the observed pattern of corruption in Serie B, where matches are fixed to be draws and the implied draw probability on betting exchanges and at bookmakers reaches disproportionately high levels.

As already mentioned, the largest residuals are five standard deviations away from the mean, and hence we now investigate large residuals. It is of interest to consider the nature of the games with the largest residuals — do they happen to coincide with games that matter the least? It is somewhat more tricky to ascertain the importance of matches in the international sphere relative to those in domestic competition since the qualification process for major continental and global tournaments varies by continent with some areas (e.g. Europe) having small groups of teams competing, whilst others have much larger groups of teams competing against each other. Nonetheless, a fairly clear distinction is between friendlies and competitive matches. Of the 210 matches we observe large negative residuals for, 123 are from friendly matches (this is insignificantly larger than the frequency of friendly matches in our overall dataset). A full list of the large-residual matches is given in Table A.

Considering the teams involved in these friendly matches, many are either youth teams (under 21s, under 19s, under 17s etc) or womens teams. It might thus be hypothesised that teams for which limited data exists are those for which we find large residuals; perhaps as a result of such low numbers of observations, a low draw probability is forecast by our econometric model. However, if the match in question is the first in the sample for both teams in a match, then both teams have equal strength from their Elo score, and the draw will have a forecast of at most 31.22% since this is the largest draw probability recorded by the econometric model over our forecast sample. Hence we cannot conclude that because teams appear infrequently in our sample we might observe strange results.

Considering also the bookmakers involved, we find that a number of bookmakers appear more often in our large-residual matches relative to their occurrence in our overall sample. The table reports on the bottom row (Overall frequency) the percentage of our observations that are from that bookmaker (e.g. 2.8% for 118bet), while the top row reports the increment for how often that bookmaker is observed in our large-residual matches (hence 2.8+3.1=5.9% for 118bet), and the numbers beneath in parentheses are t-tests of the difference in means. Hence for all bookmakers bar bet365 in the table, they are observed significantly more often in large-residual cases than overall in our sample. What is perhaps notable is that none of the major bookmakers appear in this list (see Table 8 for the frequencies with which all bookmakers are observed in our sample).

As a corollary, and also notable, is the average number of bookmakers reporting odds for these matches with large residuals. In our overall sample, on average a match has 25.1 bookmaker prices listed by *OddsPortal.com* (29.9 for matches after 2009), yet for these large-residual matches since 2009, there are on average just 7.7 bookmakers reporting prices. This in itself proves little, yet is again circumstantial as it might be expected that fewer bookmakers report prices on matches they suspect to be dubious in nature.

An additional aspect of our hypothesis is that private information becomes public through the betting markets and hence we might expect that in the cases where we identify large residuals, the majority of bookmakers for that match would report such unusual behaviour. Indeed we find that to be the case, as in 60% of matches with large residuals, more than half of the bookmakers report unusual odds, and around 30% of the time 80% or more bookmakers report large outliers. If we were picking up isolated cases, it would be expected that very infrequently would many bookmakers for the same match report odds inducing an outlier.

A final aspect of our large-residual matches that we consider are the youth and female composition of matches. As mentioned earlier, the range of pay across ages and the sexes in football is considerable, and hence it might be anticipated that youth and female matches are more likely to attract match fixers due to this. Suspect matches are significantly more likely to involve youth or female teams, which it might be argued are easier targets for fixers due to pay disparities. Specifically, almost two thirds of our sample are full international matches, yet only a third of our large outlier matches are full internationals, while only 10% of our sample are womens' matches yet they constitute 40% of our large residual matches. Both of these differences are statistically significant.[20]

It would be expected that the residuals for either positive outcome for the matches we identified using the draw are large, since the probabilities for all three events must sum to unity. Nonetheless, given that probabilities for either positive outcome much more readily span the unit interval (both have standard deviations more than three times that as for the draw), it seems less likely that such outcomes would necessarily attract particularly large residuals. It turns out that the residuals in either positive outcome are at least twice as large when the draw has been identified to have a large residual.

[20] Furthermore, if we run regressions simply on residual size, we get statistically significant coefficients for women and youth matches suggesting that for these matches. However, it might be anticipated that residuals might be larger for such matches due to smaller general information sets for such matches.

	(1) bookie_188bet	(2) bookie_sbobet	(3) bookie_12bet	(4) bookie_dafabet	(5) bookie_betvictor	(6) bookie_10bet	(7) bookie_bet365	(8) bookie_betsson	(9) bookie_unibet	(10) bookie_paddypower
large_outlier	0.031***	0.036***	0.031***	0.039***	0.015***	0.015***	0.012**	0.020***	0.006	0.000
	(8.462)	(10.532)	(9.778)	(15.852)	(3.952)	(4.187)	(2.924)	(6.110)	(1.842)	(0.007)
_cons	0.028***	0.024***	0.020***	0.012***	0.028***	0.027***	0.032***	0.023***	0.025***	0.025***
	(82.094)	(76.273)	(69.551)	(53.497)	(83.274)	(81.611)	(88.712)	(74.105)	(78.464)	(78.545)
N	240687	240687	240687	240687	240687	240687	240687	240687	240687	240687

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 5: T-tests for difference in mean to test whether bookmakers appear statistically significantly more often in large-residual matches than overall in our sample.

5 Conclusions

In this paper we have attempted to investigate the incidence of corrupt football matches using firstly economic reasoning based on information sets and secondly econometric methods, and specifically forecasting methods. We propose a test for suspicious patterns based on two types of forecasts built on different information sets: Those from bookmakers, and those from econometric models. In the absence of corrupt activity aimed at influencing the outcomes of football matches, the predictions of the two methods ought to be identical, conditional on the quality of the econometric model. If a match has been fixed then we might expect to see significant deviations between the two. In response to an allegation raised in early 2013 regarding international football matches over the period 2010–2013, we have investigated international football matches to see whether any suspicious patterns can be uncovered.

We collect a dataset of all international matches since 2004 and create an econometric forecasting model to predict matches after 2009. The forecasts are comparable in quality to bookmaker forecasts also collected for the same matches from over 60 bookmakers when appraised using the Brier score, and as such we compare the two forecasts for matches after 2009.

We find when looking at the probability of the draw in international football matches that 2,677 bookmaker prices across 210 matches have residuals of a sufficient size (more than three standard deviations away from the mean) to attract attention. We then analyse these matches, noting that there is a higher fraction of friendly matches, youth-team matches and womens matches amongst this group than the general population, and also noting that the number of bookmakers reporting prices on these matches is markedly lower than in the rest of our sample; two aspects which might attract additional attention. We also note that these large outlier bookmaker prices are clustered around a small number of matches rather than being spread amongst our sample of matches, and furthermore a statistically significantly smaller number of bookmakers report odds for these particular matches.

The evidence procured in this manner is naturally circumstantial and could be explained by legitimate factors. Nonetheless, an empirical method consistent with economic theory and hence the growing literature on forensic economics has been set out, and its potential power displayed. Practically, an automate-able method of detecting strange betting patterns has that identified in this paper which may well be an important tool as footballing authorities seek to address the problem of match fixing and other corrupt outcomes in the game. Furthermore, the website from which the data for this method is collected provides odds on many other kinds of exotic bets on the kinds of dimensions that might be targets for match fixers.

A Appendix A

B Appendix B: The Elo Score

The Elo scoring mechanism attributes each team a strength which is updated with each match played dependent on the relative strengths of the two teams competing in the match. If team A has true strength at time t of $R_{A,t}$ and team B has true strength at time t of $R_{B,t}$, then the expected score for team A against team B is:

$$E_A = \frac{1}{1 + 10^{(R_{B,t} - R_{A,t})/400}} = \frac{Q_A}{Q_A + Q_B}, \tag{9}$$

and the expected score for team B against team A is:

$$E_B = \frac{1}{1 + 10^{(R_{A,t} - R_{B,t})/400}} = \frac{Q_B}{Q_A + Q_B}, \tag{10}$$

where $E_A + E_B = 1$ and $Q_A = 10^{R_{A,t}/400}$ and $Q_B = 10^{R_{B,t}/400}$. Naturally, the true strengths of teams is unknown, and hence in practice one must choose a starting value for $R_{A,t}$ and allow it to be updated after each match. If the actual score in the match at time t for team A, $S_{A,t}$, differs from the expected score

Competition	Frequency	Percent
AFC Challenge Cup	43	0.45
AFC Championship U16	74	0.77
AFC Championship Women U19	14	0.15
AFF Suzuki Cup	46	0.48
Africa Cup of Nations	300	3.14
African Championship Women	17	0.18
African Nations Championship	27	0.28
Asian Cup	104	1.09
Asian Games	52	0.54
Caribbean Cup	65	0.68
Copa America	52	0.54
East Asian Championship	16	0.17
East Asian Championship Women	8	0.08
European Championships	645	6.74
European Championships U17	569	5.95
European Championships U17 Women	12	0.13
European Championships U19	566	5.92
European Championships U19 Women	181	1.89
European Championships U21	969	10.13
European Championships Women	208	2.17
Fifa Confederations Cup	16	0.17
Friendly International	3,064	32.03
Friendly International Women	402	4.20
Gold Cup	50	0.52
Gulf Cup of Nations	46	0.48
OFC Nations Cup	22	0.23
Olympic Games	32	0.33
Olympic Games Women	60	0.63
Toulon Tournament	62	0.65
UNCAF Nations Cup	28	0.29
West Asian Football Championship	19	0.20
World Cup	1,465	15.31
World Cup U17	103	1.08
World Cup U20	104	1.09
World Cup Women	32	0.33
World Cup Women U17	32	0.33
World Cup Women U20	62	0.65
Total	9,567	100.00

Table 6: Table detailing the different competitions in which the 9,567 matches in our sample are drawn from.

Age Band	Men Women	Total	
Under-15	5	0	5
Under-16	418	16	434
Under-17	1,793	218	2,011
Under-18	162	3	165
Under-19	1,505	547	2,052
Under-20	454	150	604
Under-21	2,459	0	2,459
Under-23	20	16	36
Full	10,262	1,106	11,368
Total	17,078	2,056	19,134

Table 7: Breakdown between different age groups and male/female football matches by team involved (hence there is 19,134 observations, two per match).

then that team's score needs updating; if $S_{A,t} = E_{A,t}$ then the existing strength for each team is accurate. Updating in the event of $S_{A,t} \neq E_{A,t}$ is done according to the formula:

$$R_{A,t+1} = R_{A,t} + K(S_{A,t} - E_{A,t}). \tag{11}$$

The factor K can be varied and is conventionally set at 32 although it is often argued that other values produce more "accurate" rankings. The setting of K affects both the convergence of R_A to its true value and also the variation around that true value. In conventional econometric terms, it is important to think about bias, efficiency and consistency of Elo rankings.

The correcting mechanism in (11) ensures that if team A's ranking is above its true value, then predictions E_A will be upward biased and hence we would expect $S_{A,t} - E_{A,t} < 0$ and hence downward pressure on that team's ranking, and vice versa. Furthermore, if both team A and B rankings are at their true values, then on average there is no movement away from equilibrium since $\mathsf{E}\left(S_A\right) = E_A$ and $\mathsf{E}\left(S_B\right) = E_B$.

The K factor affects the volatility of the ranking R_A as it is a scaling factor on the variance of the prediction error, and hence a higher K induces greater variance in the ranking, potentially yielding erroneous predictions. Equivalently, however, correcting a large disequilibrium will be faster if K is larger, since $\mathsf{E}\left(\Delta R_A\right) = K\Delta U_A$, where U_A is the prediction error for team A, which will be non-zero in expectation if a team has a ranking very distant from its true value.

Hence in a number of variants of the Elo ranking system, a competitor's early matches are weighted more highly in order that their ranking quickly converges on something close to its true value.

The Elo ranking updating function remains a complicated autoregressive time series model — the unit coefficient and prediction error interpretation of $S_A - E_A$ might lead one to suppose this is a unit-root time series model and hence will not converge to the true value of R_A. However, $S_A - E_A$ is a function of $R_{A,t}$ and $R_{B,t}$ and hence in conventional time series interpretation it is correlated with the independent variable $R_{A,t-1}$. Furthermore, $\mathsf{Corr}\left(R_{A,t}, S_{A,t} - E_{A,t}\right) < 0$ hence implying stability for the process: As the ranking increases from its equilibrium value, the prediction error decreases and hence the process corrects. Nonetheless, the speed of this correction, particularly allied with the role played by team B ensures that analytical expressions to describe the path back to equilibrium as time increases are very difficult. This is somewhat unfortunate since it is important for the applicability of Elo rankings to know about such properties.

In order to investigate we conduct a small-scale simulation exercise. We specify a true strength for N teams in Elo format and generate match outcomes according to a multinomial event where the probability of success for each team derived from the same calculations using the strength of the two teams used to calculate expected scores for the Elo score (so equations (9) and (10)). We then update Elo rankings using results from a number of full rounds of matches, where a round means all N teams play each other once. We use an initial value for the strength of each team of 1000.

Bookmaker	Number of observations	Percent
10Bet	6,428	2.77
12Bet	4,725	2.04
188Bet	6,542	2.82
32Red Bet	1,164	0.50
5Dimes	4,935	2.13
888Sport	3,487	1.50
Bestbet	3,306	1.43
Bet at Home	6,986	3.01
Bet365	7,574	3.27
Betboo	3,354	1.45
Betcris	4,313	1.86
Betfred	4,210	1.82
Betgun	4,270	1.84
Betinternet	4,013	1.73
Betonline	1,596	0.69
Betredkings	2,569	1.11
Betsafe	6,269	2.70
Betsson	5,315	2.29
Betvictor	6,678	2.88
Betway	4,811	2.07
Blue Square	4,444	1.92
Bookmaker	2,192	0.95
Boylesports	4,265	1.84
Bwin	7,369	3.18
Canbet	3,396	1.46
Coral	3,547	1.53
Dafabet	2,828	1.22
Doxxbet	6,219	2.68
Expekt	6,515	2.81
Fortunawin	972	0.42
Instant Action Sports	813	0.35
Intertops	3,605	1.55
Interwetten	5,043	2.17
Island Casino	3,156	1.36
Jetbull	4,801	2.07
Justbet	411	0.18
Ladbrokes	4,610	1.99
Legends	768	0.33
Leon Bets	4,034	1.74
Luxbet	1,295	0.56
Marathonbet	863	0.37
Mybet	5,099	2.20
Nordicbet	6,152	2.65
Noxwin	3,001	1.29
Offsidebet	875	0.38
Paddy Power	5,891	2.54
Paf	2,367	1.02
Pinnacle Sports	3,686	1.59
Redbet	2,145	0.92
SBG Global	589	0.25
Sbobet	5,686	2.45
Sportingbet	6,609	2.85
Sportsbetting	56	0.02
Stan James	3,462	1.49
The Greek	187	0.08
Tipico	5,058	2.18
Titanbet	3,147	1.36
Tonybet	533	0.23
Totesport	3,000	1.29
Ucabet	58	0.03
Unibet	5,897	2.54
Wagerweb	718	0.31
William Hill	3,993	1.72
Total	231,900	100.00

Table 8: Bookmakers from whom we draw our prices in our dataset. Source: www.OddsPortal.com.

	(1) brier_diff1	(2) brier_diff2	(3) brier_diff3
_cons	0.002	0.003^{***}	0.004
	(0.658)	(3.377)	(1.426)
N	5661	5661	5661

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 9: Test for significance of difference between Brier scores for bookmaker prices and econometric model predictions. Number in parentheses is the t-test for the difference between the two forecast models.

Considering firstly convergence, we consider a league with 10 teams and vary K in a number of dimensions. We vary K from 10 to 80 by doubling it each time. The results are plotted graphical from 1000 replications in Figure 5. The particular K for a given plot of the Elo ratings for each of the team teams is plotted. This plot would appear to suggest that higher K factors are more appropriate — the bias converges towards zero at a much faster rate, and even once there the variation around equilibrium is not particularly large. With $K = 80$, apart from for very large departures from equilibrium, convergence is very quick, taking place within around 10 matches. A number of Elo schemes use a high K for a participant's first few matches, before this reverts back down to a lower value, the reasoning being that the Elo ranking has settled around its true value by this point.

On average in our dataset, we observe a team 28.2 times, (with a maximum of 106 observations and a minimum of a single match), hence we thus assert that we have sufficient data on which to calibrate our ranking.

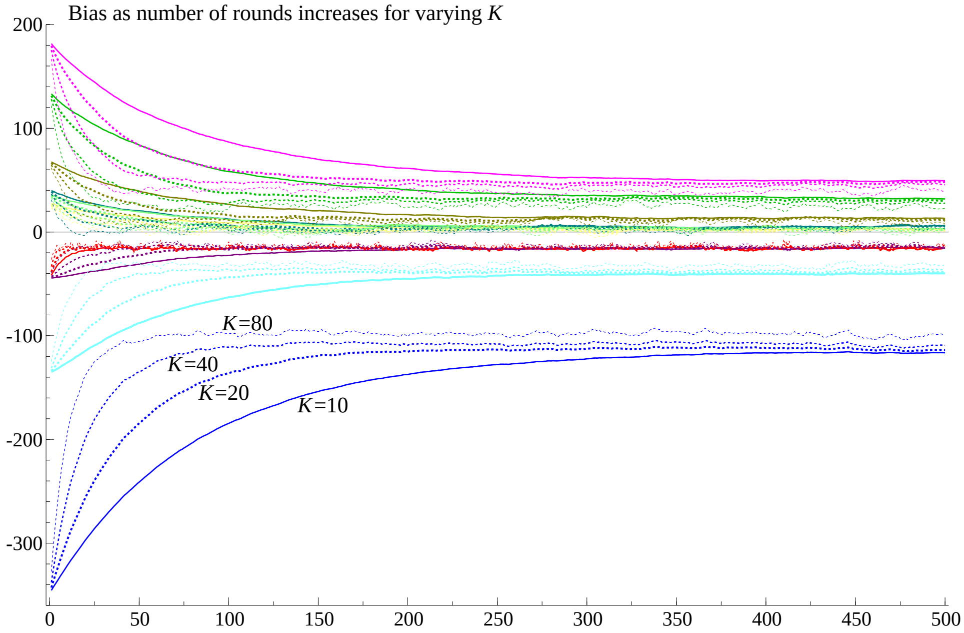

Figure 5: Plot of simulation output for Elo rankings. Bias is plotted, difference between actual Elo rating and true Elo rating. Horizontal axis is number of match rounds completed, vertical axis is Elo rating, R. The correction parameter K is varied in the different plots, and the pattern can be observed for the blue series where the respective K is marked on.

Date	Team 1	Score		Team 2	Competition
06feb2010	Japan W	2	0	China W	east-asian-championship-women-2010
19mar2010	Croatia U17	2	1	Bosnia and Herzegovina U17	euro-u17-2010
19mar2010	Northern Ireland U17	1	0	Poland U17	euro-u17-2010
18may2010	France U21	2	0	Colombia U21	toulon-tournament-2010
19may2010	Croatia U19	2	1	Belgium U19	euro-u19-2010
20may2010	Colombia U21	0	2	Ivory Coast U21	toulon-tournament-2010
21may2010	Chile U21	5	0	Russia U21	toulon-tournament-2010
22may2010	Ivory Coast U21	1	2	France U21	toulon-tournament-2010
22may2010	USA W	4	0	Germany W	friendly-international-women-2010
23may2010	Denmark U21	1	1	Chile U21	toulon-tournament-2010
16jul2010	Brazil U20 W	1	1	Sweden U20 W	world-cup-women-u20-2010
17jul2010	England U20 W	0	1	Mexico U20 W	world-cup-women-u20-2010
17jul2010	Ghana U20 W	2	4	South Korea U20 W	world-cup-women-u20-2010
17jul2010	Nigeria U20 W	2	1	Japan U20 W	world-cup-women-u20-2010
20jul2010	Costa Rica U20 W	0	3	Colombia U20 W	world-cup-women-u20-2010
21jul2010	Japan U20 W	3	1	England U20 W	world-cup-women-u20-2010
30sep2010	Scotland U19	2	1	Estonia U19	euro-u19-2011
07oct2010	Russia U19	4	2	Sweden U19	euro-u19-2011
07oct2010	Serbia U19	3	0	Bulgaria U19	euro-u19-2011
09oct2010	Ireland U19	2	1	Bulgaria U19	euro-u19-2011
10oct2010	Madagascar	0	1	Ethiopia	africa-cup-of-nations-2012
22oct2010	Northern Ireland U17	0	0	Montenegro U17	euro-u17-2011
07nov2010	Thailand	6	0	Pakistan	asian-games
16nov2010	Bulgaria U21	2	1	Ghana U21	friendly-international-2010
01dec2010	Guadeloupe	1	0	Antigua and Barbuda	caribbean-cup-2010
05mar2011	Armenia W	0	0	Georgia W	euro-women
05mar2011	Faroe Islands W	2	0	Malta W	euro-women
08mar2011	Georgia W	1	0	Faroe Islands W	euro-women
08mar2011	Malta W	1	1	Armenia W	euro-women
18may2011	Netherlands W	1	1	North Korea W	friendly-international-women-2011
02jun2011	France U21	4	1	Mexico U21	toulon-tournament-2011
02jun2011	Italy U19 W	1	0	Switzerland U19 W	euro-u19-women-2011
02jun2011	Norway U19 W	3	0	Netherlands U19 W	euro-u19-women-2011
02jun2011	Russia U19 W	3	1	Belgium U19 W	euro-u19-women-2011
04jun2011	Hungary U21	0	2	Mexico U21	toulon-tournament-2011
05jun2011	Norway U19 W	5	1	Spain U19 W	euro-u19-women-2011
15jun2011	Russia U16	3	2	Turkey U16	friendly-international-2011
16jun2011	Azerbaijan U16	2	1	Georgia U16	friendly-international-2011
16jun2011	Belarus U16	0	2	Czech Republic U16	friendly-international-2011
17jun2011	Italy U16	0	2	Turkey U16	friendly-international-2011
16jun2011	Serbia U16	3	4	Ukraine U16	friendly-international-2011
16jun2011	Turkey U16	1	0	Spain U16	friendly-international-2011
18jun2011	New Zealand W	1	0	Colombia W	friendly-international-women-2011
18jun2011	Russia U16	1	0	Italy U16	friendly-international-2011
19jun2011	Ukraine U16	3	0	Russia U16	friendly-international-2011
20jun2011	Colombia W	3	1	Wales W	friendly-international-women-2011
23jun2011	Australia W	2	0	England W	friendly-international-women-2011
24jun2011	Ivory Coast U17	4	2	Denmark U17	world-cup-u17
24jun2011	Uzbekistan U19	4	2	Azerbaijan U19	friendly-international-2011
26jun2011	Burkina Faso U17	0	2	Ecuador U17	world-cup-u17
26jun2011	Singapore U16	0	1	Flamengo RJ U15	friendly-international-2011
29jun2011	Afghanistan	0	2	Palestine	world-cup-2014
29jun2011	Uzbekistan U17	4	0	Australia U17	world-cup-u17
01jul2011	Japan W	4	0	Mexico W	world-cup-women
05jul2011	Latvia U19 W	0	0	Estonia U19 W	friendly-international-women-2011
06jul2011	Din Bucuresti	1	1	United Arab Emirates	friendly-international-2011

continued on next page

continued from previous page

A notable feature of Figure 5 is that some ratings appear biased. Exploring this further, it appears that bias is a function of both the number of teams competing in a competition and their relative deviation from 1000. From Figure 6, where the number of teams competing is increased from 3 to 20, when only three teams compete all rankings are biased down, while as the number of teams is increased, the biases appear to be similarly upward as downward biased. Figure 7 sheds light on this phenomena by plotting the levels of the rankings. The further is a team's true ranking from 1000, the larger is its bias. This means that relative differences in team strength will be distorted as team quality departs from the mean. For example, the quality difference between a very good team and a very bad team will be attenuated using the Elo rankings. Nonetheless, if this effect is consistent throughout our sample, our regression technique for predicting match outcomes using the Elo scores should correct for this.

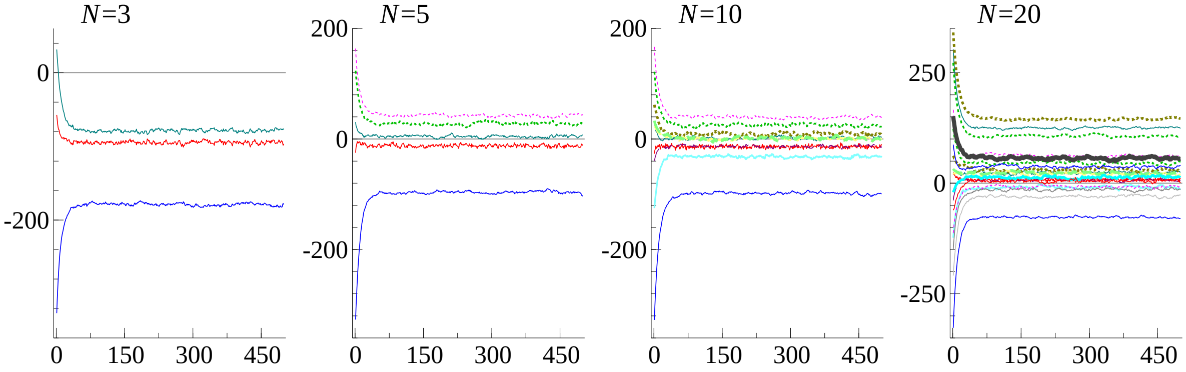

Figure 6: Plot of simulation output for Elo rankings. Difference between actual Elo rating and measured Elo rating plotted for each round of games played in round-robin tournament system. Number of teams in league varied between plots.

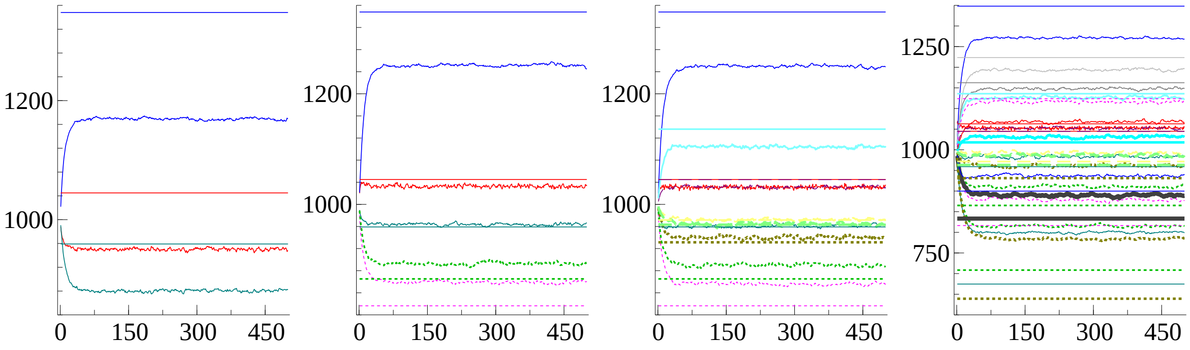

Figure 7: Plot of simulation output for Elo rankings. Actual Elo and measured Elo ratings are plotted for each round of games played in round-robin tournament system. Number of teams in league varied between plots.

A final aspect of the weighting factor K worth considering is that one particular application of Elo ratings to international football, the *World Football Elo Ratings* (WFEL) weights different matches.[21] The

[21] See http://www.eloratings.net/system.html and http://en.wikipedia.org/wiki/World_Football_Elo_Ratings for more information.

reasoning behind this is that competitive football reveals much more about the true strengths of a nation's team. Compared to sub-national football, friendly matches constitute a significant proportion of matches; about a third of our dataset consists of friendly matches as opposed to competitive matches. Friendlies are often used by teams to experiment, and it may thus be that even if players do try their hardest the outcome is not reflective of the true quality difference between the teams. Given this, WFEL sets $K = 20$ for friendly matches, $K = 60$ for World Cup finals matches, $K = 50$ for continental championship finals matches (e.g. African Cup of Nations, European Championships), $K = 40$ for all qualifying matches for such competitions, and $K = 30$ for all other international tournaments. Given our analysis of K via simulation, such a weighting of different types of matches would appear sensible; *some* information will be provided by friendly matches about the relative quality of teams, but not as much as in competitive matches, and hence the results of the latter kinds of matches are weighted more heavily in order to ensure that the Elo ranking better reflects actual quality differences between teams.

Another augmentation that WFEL use and is also mentioned by Hvattum and Arntzen (2010) is to use the goal difference in a match to update rankings. Hvattum and Arntzen suggest a formula of $K = K_0(1 + \delta)^\lambda$, where K_0 and λ are parameters to be set, and δ is the absolute goal difference in the match. The goal difference in a match can be viewed as additional information on the realised quality of the teams involved and hence could be used to further update each team's rank. This may also be helpful in allowing adjustments when team rankings are substantially away from their true values (such as the blue lines in Figures 5–7). Because of this latter argument we implement the goal-different augmentation for our Elo ranking.

Next we present our Elo rankings for international teams. We initialise each team with a strength of 1000 and update this for each match a team participates in. By the end of the sample, the top international teams with their Elo ranking, score and Fifa ranking are listed in Table 11. A dynamic plot of a selection of countries through the sample is provided in Figure 8. The horizontal axis in the plot relates to each team in match in our sample (hence there is around 18000 observations), and shows how the ratings evolve through time, expanding away from 1000 very quickly. The vertical dotted line represents the end of our estimation period and the beginning of the forecast period, notably 2010 onwards. The Spearman's Rank correlation coefficient for the 183 teams matched between Fifa and our Elo rankings is 0.796. The top two teams, Spain and Germany, are the same in both ranking systems, whilst the third Elo ranked team, Brazil, intuitively ought to be ranked more like third than the 18th it currently ranks in the Fifa ratings. Nonetheless, a ranking system is simply a rankings system, what is more important is the predictive power it may have, which we consider in the next section.

References

Becker, G. (1968), 'Crime and Punishment: An Economic Approach', *The Journal of Political Economy* **76**, 169–217.

Croxson, K. and J.J. Reade (2011), Exchange vs. Dealers: A High-Frequency Analysis of In-Play Betting Prices, Discussion Papers 11-19, Department of Economics, University of Birmingham.

Fifa (2010), Almost half the world tuned in at home to watch 2010 FIFA World Cup South Africa, Media release, Fifa.
URL: *http://www.fifa.com/worldcup/archive/southafrica2010/organisation/media/newsid=1473143/index.html*

Forrest, D.K., J. Goddard and R. Simmons (2005), 'Odds-Setters As Forecasters: The Case of English Football', *International Journal of Forecasting* **21**, 551–564.

Goddard, J. (2005), 'Regression Models for Forecasting Goals and Match Results in Association Football', *International Journal of Forecasting* **21**, 331–340.

Hanson, R. and R. Oprea (2009), 'A Manipulator can Aid Prediction Market Accuracy', *Economica* **76**(302), 304–314.

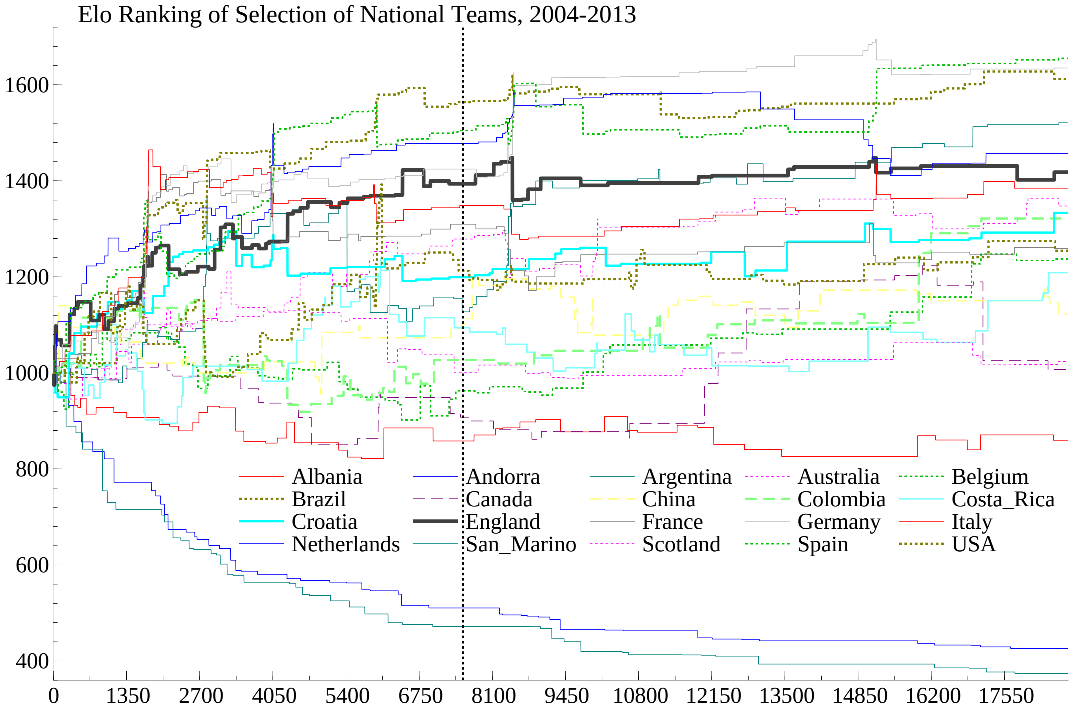

Figure 8: Plot of evolution of Elo rankings through our sample for a selection of national teams.

Harris, N. (2013), 'EXCLUSIVE: Top match-fix investigator reveals 'real story' about new cases', *sporting-intelligence.com* .
URL: *http://www.sportingintelligence.com/2013/02/07/exclusive-top-match-fix-investigator-reveals-real-story-070201/*

Hendry, David (2011), Unpredictability in Economic Analyis, Econometric Modelling and Forecasting, Economics Series Working Papers 551, University of Oxford, Department of Economics.
URL: *http://ideas.repec.org/p/oxf/wpaper/551.html*

Hill, D. (2010), *The Fix: Soccer and Organized Crime*, McClelland Stewart.

Hill, D. (2013), 'Another 'I Told You So' Moment', *Declan Hill's Blog* .
URL: *http://www.howtofixasoccergame.com/blog/?p=319*

Hvattum, Lars Magnus and Halvard Arntzen (2010), 'Using elo ratings for match result prediction in association football', *International Journal of forecasting* **26**(3), 460–470.

Leitner, Christoph, Achim Zeileis and Kurt Hornik (2010), 'Forecasting sports tournaments by ratings of (prob) abilities: A comparison for the euro 2008', *International Journal of Forecasting* **26**(3), 471–481.

Lewis, J.B. and D.A. Linzer (2005), 'Estimating Regression Models in Which the Dependent Variable Is Based on Estimates', *Political Analysis* **13**(4), 345–364.

Manski, C. (2006), 'Interpreting the predictions of prediction markets', *Economic Letters* **91**(3), 425–429.

McHale, I. and P. Scarf (2006), 'Forecasting International Soccer Match Results Using Bivariate Discrete Distributions', *Working Paper, Management and Management Sciences Research Institute, University of Salford* .

team	elo rank	fifa rank	elo score	Fifa points
Spain	1	1	1656.275	1610
Germany	2	2	1636.555	1473
Brazil	3	18	1614.205	908
Argentina	4	3	1525.154	1309
Japan	5	26	1460.978	811
Netherlands	6	8	1457.427	1106
Portugal	7	7	1424.283	1133
England	8	4	1418.652	1174
Mexico	9	15	1394.525	995
Ivory Coast	10	13	1389.442	1022
Italy	11	5	1385.759	1173
Iran	12	57	1354.075	540
Russia	13	10	1353.237	1064
Australia	14	39	1348.685	634
Nigeria	15	30	1347.425	775
Colombia	16	6	1335.625	1159
Croatia	17	9	1334.745	1080
Panama	18	42	1298.108	624
Honduras	19	49	1293.302	592
Sweden	20	21	1289.373	849

Table 11: Comparison of top 20 Elo ranked teams and their corresponding Fifa rank, as of 21 March 2013.

Olmo, J., K. Pilbeam and Pouliot W. (2011), ‘Detecting the Presence of Insider Trading via Structural Break Tests’, *Journal of Banking and Finance* **35**, 2820–2828.

Preston, I. and S. Szymanski (2000), ‘Racial Discrimination in English Football’, *Scottish Journal of Political Economy* **47**(4), 342–363.

Price, J. and J. Wolfers (2010), ‘Racial Discrimination Among NBA Referees’, *Quarterly Journal of Economics* **4**, 1859–1887.

Reade, J.J. (2013), Detecting corruption in football, *in* J.Goddard and P.Sloane, eds, ‘Handbook on the Economics of Professional Football’, Edward Elgar.

Rottenberg, S. (1956), ‘The Baseball Players’ Labor Market’, *The Journal of Political Economy* **64**(3), 242–258.

Wolfers, J. (2006), ‘Point Shaving: Corruption in NCAA Basketball’, *The American Economic Review* **96**(2), 279–283.

Wolfers, J. and E. Zitzewitz (2006), ‘Interpreting prediction market prices as probabilities’, *CEPR Discussion Paper* (5676).

Wooldridge, J.M. (2002), *Econometric Analysis of Cross Section and Panel Data*, The MIT Press, Cambridge, Mass.

Zitzewitz, Eric (2012), ‘Forensic economics’, *Journal of Economic Literature* **50**(3), 731–69.