Journal of Sports Analytics 8 (2022) 231–250
DOI 10.3233/JSA-200619
IOS Press

FIFA ranking: Evaluation and path forward

Leszek Szczecinski[a,*] and Iris-Ioana Roatis[b]
[a]*Institut National de la Recherche Scientifique, Montreal, Canada*
[b]*Imperial College, London, UK*

Received 5 January 2022
Accepted 9 June 2022
Pre-press 6 July 2022
Published 30 December 2022

Abstract. In this work, we study the ranking algorithm used by Fédération Internationale de Football Association (FIFA); we analyze the parameters that it currently uses, show the formal probabilistic model from which it can be derived, and optimize the latter. In particular, analyzing games since the introduction of the algorithm in 2018, we conclude that game's "importance" (defined by FIFA and used by the algorithm) is counterproductive from the point of view of the predictive capacity of the algorithm. We also postulate that the algorithm should be rooted in the formal modeling principle, where the Davidson model proposed in 1970 seems to be an excellent candidate, preserving the form of the algorithm currently used. The results indicate that the predictive capacity of the algorithm is considerably improved by using the home-field advantage (HFA), as well as the explicit model for the draws in the game. Moderate but notable improvement may be achieved by introducing the weighting of the results with the goal differential, which, although not rooted in a formal modeling principle, is compatible with the current algorithm and can be tuned to the characteristics of the football competition.

Keywords: Association football, ranking, rating, FIFA, Elo algorithm, margin of victory

1. Introduction

In this work we evaluate the algorithm used by Fédération Internationale de Football Association (FIFA) to rank the international Men teams, as well as we propose and study simple modifications to improve the prediction capacity of the algorithm.

Rating and ranking are important elements of sport competitions and the surrounding entertainment environments. In general, the rating has an informative function that provides fans and profane observers with quick insight into the relative strength of the teams. For example, the press is often interested in the "best" teams or the national team reaching some record position in the ranking.

More importantly, ranking leads to consequential decisions such as (i) seeding, i.e., defining which teams play against each other in competitions (e.g., used to establish the composition of the groups in the qualification rounds of the FIFA World Cup), (ii) promotion / relegation (e.g., determining which teams move between the English Premier League (EPL) and the English Football League Championship, or teams that move between the groups in Nations Leagues), or (iii) defining the participants in prestigious (and lucrative) end-of-season competitions (such as Champions League in European football, Stanley Cup series in National Hockey League (NHL)).

Most of the currently used ratings simply count wins/losses (and draws, when applicable), but some of the sport governing bodies have gone beyond these simple methods and implemented more sophisticated rating algorithms where the rating levels attributed to the teams are meant to represent the "skills" or "strengths"; the ranking is obtained by sorting these numbers and is also known as a "power ranking".

In particular, FIFA started a new ranking/rating algorithm in 2018, where the rating levels (skills) assigned to the teams are calculated from the game outcome, of course, but also from the skills of the teams before the game. The resulting rating algorithm has the virtue of being simple and defined in a (mostly) transparent manner.

Considering that the football association is, by any measure, the most popular sport in the world, and given the importance of the rating/ranking, the main objective of this work is to analyze the FIFA rank-

*Corresponding author: Leszek Szczecinski, Institut National de la Recherche Scientifique, Montreal, Canada. E-mail: Leszek.Szczecinski@inrs.ca.

ing using statistical methods. This evaluation has a value of its own and follows the line of many works that analyzed the past ranking strategies used by FIFA, e.g., (Lasek, Szlávik, & Bhulai, 2013; Ley, Van de Wiele, & Van Eetvelde, 2019). Furthermore, the approach we propose can also be applied to evaluate other rating algorithms, e.g., such as the one used by Fédération Internationale de Volleyball (FIVB), (FIVB, 2020).

In this work we will:

- Derive the FIFA algorithm from the first principles. In particular, we will define the probabilistic model underlying the algorithm and identify the estimation method used to estimate the skills.
- Assess the relevance of the parameters used in the current algorithms. In particular, we will evaluate the role played by the change of the adaptation step according to game-importance (as defined by FIFA).
- Optimize the parameters of the proposed model. As a result, we derive an algorithm which is equally as simple as FIFA's one, but allows us to improve the prediction of the game results.
- Propose modifications of the algorithm that take into account the goal differential, also known as margin of victory (MOV). We consider legacy-compliant algorithms and a new version of the rating.

Our work is organized as follows. In Section 2 we describe the FIFA algorithm in the framework that simplifies the manipulation of models and the evaluation of the results. This is also where we clarify the origin of the data, make a preliminary evaluation of the relevance of the game-importance parameters currently used to control the size of the adaptation step, and assess the impact of the shootout/knockout rules present in the FIFA algorithm.

The algorithm is then formally derived in Section 3 where we also discuss the evaluation of the results and the batch estimation approach we use. Incorporation of the MOV in the rating is evaluated in Section 4 using two different strategies. In Section 5 we return to the on-line rating, evaluating and re-optimizing the proposed algorithms, and discussing the practical role played by the scale. We conclude the work in Section 6 summarizing our findings and in Section 6.1 we make an explicit list of recommendations which may be introduced to improve the current version of the FIFA algorithm.

2. FIFA ranking algorithm

We consider the scenario in which there are a total of M teams playing against each other in the games indexed with $t \in \mathcal{T} = \{1, \ldots, T\}$, where T is the number of games in the observed period. FIFA ranks 211 international teams and, between June 4, 2018 and March 31, 2022, there were 3446 FIFA-recognized games. However, for the purpose of this study, we removed two games and one team;[1] thus, we use $M = 210$ and $T = 3444$.

Let $\theta_{t,m}$ denote the skill of the teams $m \in \{1, \ldots, M\}$ before the game t. The skills of all teams are gathered in a vector $\boldsymbol{\theta}_t = [\theta_{t,1}, \ldots, \theta_{t,M}]^\top$, where $(\cdot)^\top$ denotes the transpose. The home and away teams are denoted by i_t and j_t, respectively.

Game results $y_t \in \mathcal{Y}$ are ordinal variables, where the elements of $\mathcal{Y} = \{\mathsf{H}, \mathsf{D}, \mathsf{A}\}$ represent the win of the home team ($y_t = \mathsf{H}$), the draw ($y_t = \mathsf{D}$), and the win of the away team ($y_t = \mathsf{A}$). These ordinal variables are often transformed into numerical *scores* $\check{y}_t = \check{y}(y_t)$: $\check{y}(\mathsf{A}) = 0$, $\check{y}(\mathsf{D}) = 0.5$, and $\check{y}(\mathsf{H}) = 1$.

The basic rules of the FIFA rating for a team $m \in \{i_t, j_t\}$ are defined as follows:

$$\theta_{t+1,m} \leftarrow \theta_{t,m} + I_{c_t}\delta_{t,m} \tag{1}$$

$$\delta_{t,m} = \check{y}_{t,m} - F\left(\tfrac{z_{t,m}}{s}\right) \tag{2}$$

$$F(z) = \frac{1}{1 + 10^{-z}} \tag{3}$$

$$z_{t,m} = \theta_{t,m} - \theta_{t,m'}, \tag{4}$$

[1] We had to deal with minor exceptions:

- We recognized the victory of Guyana (GUY) over Barbados (BRB) in the game played on Sept. 6, 2019 already on the date of the game, while in the FIFA rating, the draw was originally registered and GUY's victory was recognized only later, when BRB was disqualified for having fielded an ineligible player.
- We remove the Côte d'Ivoire (CIV) vs. Zambia (ZAM) game, played on June 19, 2019, where CIV (the winner) and ZAM exchanged 2.21 points. The removal of this game from the FIFA-recognized list seems to be the reason why FIFA changed the ratings of both teams between two official publications on Dec. 19, 2019 and on Feb. 20, 2020. Namely, CIV's rating was changed from 1380 to 1378 and ZAM's from 1277 to 1279. This was done despite both teams not playing at all in this period of time.
- The game of Cook Islands (COK) played against Solomon Islands (SOL) on March 17, 2022 was removed because this is the only COK's game in the entire period and thus its rating, which was assumed by FIFA to be equal to 908 before the game, is not based on any recent results. The disadvantage of this removal is that we affect the rating of SOL which played three more games before March 31, 2022. We recognize that the introduction of a new team to the system is indeed a challenging issue from a rating perspective.

Table 1
Game-categories c and the corresponding update steps $I_c = K\xi_c$, (FIFA, 2018), where $K = 5$ and $\xi_c = I_c/K$. The number of games T_c and their frequency, $f_c = T_c/T$ in the observed categories between June 4, 2018 and March 31, 2022 is also given (total number of games is $T = 3444$)

c	I_c	ξ_c	Description	T_c	f_c[%]
0	5	1	Friendlies outside International Match Calendar windows	518	15
1	10	2	Friendlies during International Match Calendar windows	701	20
2	15	3	Group phase of Nations League competitions	351	10
3	25	5	Play-offs and finals of Nations League competitions	84	2.4
4	25	5	Qualifications for Confederations/World Cup finals	1413	41
5	35	7	Confederation finals up until the QF stage	253	7.3
6	40	8	Confederation finals from the QF stage onwards	60	1.7
7	50	10	World Cup finals up until QF stage	56	1.6
8	60	12	World Cup finals from QF stage onwards	8	0.2

where $s = 600$ is the scale,[2] m' is the index of the team opposing the team m in the t-th game (i.e., if $m = i_t$ then $m' = j_t$ and if $m = j_t$ then $m' = i_t$), $\check{y}_{t,m}$ is the "subjective" score of the team m (if $m = i_t$, then $\check{y}_{t,m} = \check{y}_t$, and if $m = j_t$, then $\check{y}_{t,m} = 1 - \check{y}_t$). The result produced by the logistic function, $F(z_{t,m}/s)$ in (2) is referred to as *expected score* and I_{c_t} is the update step defined by FIFA that depends on the game category (or game "importance"), c_t shown in Table 1.

When the team m does not play, its skills do not change, i.e., $\theta_{t+1,m} \leftarrow \theta_{t,m}$.

The steps I_c are defined by FIFA, and we divide them into two components:

$$I_c = K\xi_c, \quad c = 0, \ldots, 8, \tag{5}$$

where ξ_c is a category-dependent adjustment shown in Table 1. Since the split (5) is not unique, we remove any ambiguity by setting $\xi_0 = 1$, i.e., $K = I_0 = 5$ (from Table 1).

The basic equation governing the change in skills in (2) is next supplemented with the following rules:

- *Knockout rule*: in the knockout stage of any competition (which follows the group stage), instead of (2) we use

$$\delta_{t,m} \leftarrow \max\{0, \delta_{t,m}\} \tag{6}$$

  which guarantees that no points are lost by teams moving out of the group stage.
- *Shootout rule*: If the team m wins the game in the shootouts, we use

$$\check{y}_{t,m} \leftarrow 0.75, \quad \check{y}_{t,m'} \leftarrow 0.5, \tag{7}$$

  where m' is the index of the team that lost.

  This rule, however, does not apply in two-legged qualification games if the shootout is required to break the tie.

The rating we describe has been published by FIFA since August 2018, roughly once a month. The algorithm was initialized on June 4, 2018, with the initialization values $\boldsymbol{\theta}_0$ based on the previous ranking system.[3]

To run the algorithm, we need to know the initialization $\boldsymbol{\theta}_0$, the presence of conditions that trigger the use of the knockout/shootout rules, and most importantly the category/importance of each game c_t. These elements are not officially published, so here we use the unofficial data shown in Football Rankings (2021) which keeps track of the FIFA rating since June 2018. Using it, we were able to reproduce the ratings $\boldsymbol{\theta}_t$ with a precision of fractions of rating points, which gives us confidence that the game-categories are assigned according to the FIFA rules.[4]

In our discussion of models and algorithms, we want (i) to understand the rationale behind the current FIFA rating algorithm and (ii) to propose new and simple rating algorithms.

We start by asking simple questions: Are the parameters I_c defining the "importance" of the game

[2]The role of the scale is to ensure that the values of the skills $\theta_{t,m}$ are situated in a visually comfortable range; the interplay between the scale and the initialization $\boldsymbol{\theta}_0$ is discussed in Section 5.1.

[3]The team ranked r was assigned the rating $\theta_{m,0} = 1600 - 4(r-1)$, so Germany (ranked first, $r = 1$) was assigned $\theta_{m,0} = 1600$ and the rating of the other teams was then decreased by four points with each position, so Brazil was assigned $\theta_{m,0} = 1596$, Belgium $\theta_{m,0} = 1592$, etc. The tied positions in the previous ranking were dealt with by removing the lowest of the ranking positions, e.g., two teams ranked $r = 10$, meant that the next available ranking position was $r = 12$.

[4]Information provided by Football Rankings (2021) is highly valuable because it is far from straightforward to verify which games are included in the rating and what their importance I_c is. In particular, games in the same tournament can be included or excluded from the rating, and in some cases the changes can be made retroactively, further complicating the understanding of the rating results.

suitably set? If not, how should we define them to improve the results?

These questions are interesting in their own right because the concept of game-importance is not unique to the FIFA rating: it also appears in the FIVB rating, (FIVB, 2020) and in the statistical literature, e.g., (Ley et al., 2019, Sec. 2.1.2).

2.1. *Effect of weighting using game-importance: preliminary evaluation*

In statistics, a conventional approach to performance evaluation is to rely on a metric, called a scoring function, which relates the result y_t to its prediction obtained from the estimates at hand (here, $\boldsymbol{\theta}_t$), (Gelman, Hwang, & Vehtari, 2014).

At this point we want to use only the elements that are clearly defined in the FIFA ranking and since the only explicit predictive element defined in the FIFA algorithm is the expected score (3), $F(z_t/s) = \mathbb{E}[\check{y}_t|z_t]$, we will base the evaluation on the metric affected by the mean. Later we will abandon this simplistic approach.

Using the squared prediction error,

$$\mathsf{m}(z_t, y_t) = \big(\check{y}_t - F(z_t/s)\big)^2, \tag{8}$$

averaged over the large number of games, we obtain the Mean Squared Error (MSE) estimate

$$\mathsf{MSE} = \frac{1}{T - T'} \sum_{t=T'+1}^{T} \mathsf{m}(z_t, y_t), \tag{9}$$

where $T' = \lceil T/2 \rceil$ is the time-index separating the observations into two approximately equal parts. So the games in the first part are used to initialize the algorithm and the second part is used to calculate the metrics. This separation, which aims to attenuate the initialization effects, is somewhat arbitrary, of course, but should not significantly affect the results for large T.

The MSE in (9) may be treated as an estimate of the expectation,

$$\begin{aligned} \mathsf{MSE} &\approx \mathbb{E}_{z_t}\Big[\mathbb{E}_{y_t|z_t}[\big(\check{y}_t - F(z_t/s)\big)^2]\Big] \\ &= \mathbb{V}[\check{y}_t] + \mathbb{E}_{z_t}\Big[\big(B(z_t, y_t)\big)^2\Big], \end{aligned} \tag{10}$$

which highlights the bias-variance decomposition, (Duda, Hart, & Stork, 2001, Ch. 9.3.2) and where $\mathbb{V}[\check{y}_t] = \mathbb{E}_{z_t}\big[\mathbb{V}[\check{y}_t|z_t]\big]$ is the average conditional variance of $\check{y}_t$, and $B(z_t, y_t) = F(z_t/s) - \mathbb{E}[\check{y}_t|z_t]$ is the bias in the estimation of the mean.[5]

Therefore, by reducing the (absolute value of the) bias $B(z_t, y_t)$, that is, by improving the calculation of the expected score $F(z_t/s)$, should manifest itself in a lower value of the MSE, which is calculated as in (9).[6]

Using the MSE, we are now able to assess how the values of the importance parameters I_c (or alternatively, K and ξ_c) affect the expected value of the score (i.e., the estimate of the mean).

We find the coefficients K and/or ξ_c by minimizing the MSE (9) using the following alternate optimization which turned out to converge quickly (and be independent of the initialization)

$$\xi_c \leftarrow \arg\min_{\xi_c} \mathsf{MSE}, \quad c = 1, \ldots, 8 \tag{11}$$

$$K \leftarrow \arg\min_{K} \mathsf{MSE}; \tag{12}$$

MSE was thus optimized with respect to only one variable at a time, while all the others were kept fixed. We perform optimization (11) on eight weights $\xi_1, \ldots, \xi_8$ (recall that we set $\xi_0 = 1$), one optimization (12) on K, and repeat these steps until convergence, defined as an insignificant change in MSE (less than 0.01%). The arrow $\leftarrow$ means that, once optimization over a variable ξ_c is finished, the optimal result is used in the following optimizations; the same applies to K, of course.[7]

The one-dimensional optimizations (11) and (12) only require one-dimensional line search (e.g., over a grid) and we preferred to avoid derivative-based methods which are not well suited to deal with the complicated functional relationship resulting from the recursive rating algorithm.

The results are shown in Table 2 and we observe the following:

[5] We emphasize that $\mathbb{E}[\check{y}_t|z_t]$ is the true but unknown mean of the score, while $F(z_t/s)$ is the *estimate* of the mean.

[6] We note that, in (9) we assume that all scoring functions $\mathsf{m}(z_t, y_t)$ are equally important, which seems to be in contradiction with the idea of using variable weights ξ_{c_t} attributed to different games. But, while trying to weight the scoring functions is a theoretically interesting issue, not only would it add another layer of complexity to the problem but, in hindsight (i.e., after obtaining the optimization results where the optimized weights ξ_c are very similar) it seems to be not very useful.

[7] Alternatively, we may remove (12), e.g., set $K = 5$, and carry out only the optimization (11) for $c = 0, 1, \ldots, 8$. The optimal solution will then be obtained by exploiting (5) as $K \leftarrow K\xi_0$ and $\xi_c \leftarrow \xi_c/\xi_0$, $c = 1, \ldots, 8$.

Table 2

Parameters K and ξ_c, in (5), are either fixed (shadowed cells), or obtained by minimizing the MSE (9). The last three column show the value of the MSE obtained after the FIFA algorithm is modified by removing the shootout rule ($\mathsf{MSE}_{\setminus\text{so}}$), by removing the knockout rule ($\mathsf{MSE}_{\setminus\text{ko}}$), and by removing both the shootout and the knockout rules ($\mathsf{MSE}_{\setminus\text{so\&ko}}$). The first row corresponds to the values obtained with the parameters fixed in the FIFA algorithm, i.e., K and ξ_c are taken from Table 1.

$\mathsf{MSE}_{\text{opt}}$	K	ξ_0	ξ_1	ξ_2	ξ_3	ξ_4	ξ_5	ξ_6	ξ_7	ξ_8	$\mathsf{MSE}_{\setminus\text{so}}$	$\mathsf{MSE}_{\setminus\text{ko}}$	$\mathsf{MSE}_{\setminus\text{so\&ko}}$
0.1340	5	1	2	3	5	5	7	8	10	12	0.1340	0.1348	0.1349
0.1299	15	1	2	3	5	5	7	8	10	12	0.1299	0.1308	0.1308
0.1295	55	1	1	1	1	1	1	1	1	1	0.1295	0.1299	0.1299
0.1287	55	1	0.8	1.5	0.5	1.2	0.8	2.4	0.06	7.8	0.1287	0.1298	0.1299

- The common update step K increases ten-fold in the optimized solution and it seems that it is the most important contributor to the improvement of the MSE (which changes from $\mathsf{MSE}_{\text{opt}} = 0.1340$ in the original algorithm to $\mathsf{MSE}_{\text{opt}} = 0.1295$ in the algorithm with fixed weights ξ_c but larger common adaptation step).
- For the games in the categories well represented in the data, i.e., $c \in \{0, 1, 2, 4, 5\}$, the relative importance of the games ξ_c does not seem to be critically different and, for sure, does not match the values used in the FIFA algorithm. Overall, the optimized weights ξ_c yield a very small improvement in MSE, when compared to the use of constant weights, $\xi_c \equiv 1$.

  In fact, the Friendlies played in the International Match Calendar window are weighted down ($\xi_1 = 0.8$) compared to the Friendlies played outside the window, and this is contrary to what the FIFA algorithm does.
- Estimates of ξ_c for categories $c \in \{3, 6, 7, 8\}$ should not be considered very reliable because the number of games in each of these categories is rather small (less than 6% of the total). Furthermore, games in the categories $c = 7$ and $c = 8$ were observed only in June 2018, during the 2018 World Cup; therefore, their effect is most likely very weak in the games from the second half of the observed batch, see (9), which starts in Sept. 2020.

Using a very simple MSE criterion derived from the definitions used by the FIFA algorithm, we obtain results that cast doubt on the optimality/utility of the game-importance parameters, I_c proposed by FIFA.

However, drawing conclusions at this point may be premature. For example, regarding K (which, after optimization should be much larger than 5), it is possible that the relatively short period of observation time (34 months) is not sufficient for small K to guarantee the sufficient convergence but may pay off in a long run, when smaller values of K will improve the performance after the convergence is reached. We will return to this issue in Section 5.1 when analyzing the interaction between the initialization of the algorithm and the scale s.

On the other hand, to address concerns about the weights ξ_c, the situation is quite different. Even after the convergence, the weights associated with different categories should affect the results in a meaningful way. To elucidate this point, we will take a more formal approach and, in Section 3, go back to the "drawing board" to derive the rating algorithm from the first principles.

Before that, however, we will evaluate the impact of the knockout/shootout rules.

2.2. Effect of knockout/shootout rules

The basic algorithmic Equations (1)–(2) guarantee that the teams "exchange" the rating points so that their total stays constant, i.e., $\sum_{m=1}^{M} \theta_{t,m} = \sum_{m=1}^{M} \theta_{t+1,m}$: this is a well-known property of the Elo rating algorithm, (Langville & Meyer, 2012, Ch. 5). On the other hand, by applying the knockout/shootout rules, we always obtain a new $\delta_{t,m}$ that is equal to or greater than the original one in (2). Thus, the shootout/knockout rules are a source of "inflation" in the rating. In fact, in the analyzed period, the total number of points increased by 2099 (with the initial total being 254680) and this was due to 24 games with the shootout rule (but not the knockout), 90 games with the knockout rule (but not the shootout), and 30 games where both rules were applied.

In the absence of known mathematical principles from which the knockout/shootout rules are derived, our initial hypothesis is that the knockout rule is a heuristics introduced to compensate for the increased value of I_c in the advanced stages of competitions. To test this hypothesis, we will proceed by removing the shootout or/and knockout rules from the algorithm and observe the impact of such removals on the MSE.

Table 3
Ranking of the top teams: Brazil (BRA), Belgium (BEL), France (FRA), Argentine (ARG), and England (ENG). The first row shows the Spearman correlation coefficient, ρ, calculated between the modified rankings and the ranking obtained by the original FIFA algorithm, shown in the first columns (as of March 31, 2022); the three next columns display the results when the shootout or/and knockout rules are removed

	Original algorithm	No shootouts rules	No knockouts rules	No shootouts /knockouts rules
ρ	1.0	0.80	0.66	0.58
	BRA (1832.7)	BRA (1829.6)	BRA (1790.5)	BRA (1782.5)
	BEL (1827.0)	BEL (1825.7)	FRA (1780.9)	FRA (1779.3)
	FRA (1789.9)	FRA (1788.8)	BEL (1756.3)	BEL (1754.4)
	ARG (1765.1)	ARG (1758.3)	ARG (1737.1)	ARG (1726.8)
	ENG (1761.7)	ENG (1752.0)	ENG (1724.5)	ENG (1714.1)

The results are shown in the last three columns of Table 2 and we conclude that the prediction capacity of the algorithm is negligibly affected by the shootout rule and it slightly but still notably deteriorates if the knockout rule is removed.[8] Thus, our hypothesis is not supported by the results, even if the argument in favor of using the knockout rule is very weak, as we will also see later in Section 5.2.

To obtain an intuitive understanding of how the removal of shootout/knockout rules affects the results, Table 3 compares the ranking obtained using the FIFA algorithm (first column) with the rating resulting from the modified algorithm in which we (i) eliminate the shootout rule (second column), (ii) eliminate the knockout rule (third column), as well as (iii) eliminate both rules (fourth column).

The Spearman correlation coefficient, ρ (Myers & Well, 2003, Ch. 18.5.3), which quantifies the difference in the rankings of all teams (perfect agreement yields $\rho = 1.0$) indicates that the changes in the ranking are similar to what was obtained by observing the MSE: comparing the shootout and the knockout rules, the latter affect the results more significantly.

We also show the rankings of the top teams where the changes are not major and the most notable is the switch of ranks between Belgium (BEL) and France (FRA), which can be attributed to a different number of times the teams benefitted from the knockout rules. Indeed, by analyzing the results of the games, we observed that in the original ranking, BEL benefited four times from the knockout rule for a total of 85 points (which would be lost without the rule (6)), while FRA benefited only once, gaining 14 points.[9]

Although the knockout rule provides a slight but notable improvement from the prediction point of view, we may still debate whether this heuristics is fair and desirable.

In particular, we note that the points-preserving knockout rule partially ignores the direct comparison between the teams. For example, games in which BEL was not penalized (for losing in the knockout stages) were played against FRA (twice). Thus, despite direct evidence indicating that FRA was able to beat BEL, the knockout rule preserved the points earned by BEL in other games.

In fact, such situations are not surprising and, indeed, the top teams are likely to make it to the final stages of the important competitions and then play against each other in the games where knockout rules are applicable (in case of BEL's games: World Cup 2018, Euro 2020, and UEFA Nations League 2021). Although these games will provide direct comparison results, the current knockout rule will preserve the points of the losing team.

3. Derivation of the algorithm and batch-rating

To understand and eventually modify the rating algorithm used by FIFA we propose to cast it in the well-defined probabilistic framework. To this end we define explicitly a model relating the game outcome y_t to the skills of the home-team (θ_{i_t}) and the away-team (θ_{j_t}), where the most common assumption is that the probability that a random variable Y_t takes the value y_t, depends on the skill difference

[8] The immediate question is whether the optimization of the $\mathsf{MSE}_{\backslash \mathrm{so}}$ or $\mathsf{MSE}_{\backslash \mathrm{ko}}$ would change the conclusion? The answer is negative, as we also verified: for example, by minimizing $\mathsf{MSE}_{\backslash \mathrm{ko}}$ with respect to K, and then adding the knockout rule, the results are improved. This is not particularly interesting or surprising, which is why we do not show these results.

[9] Of course, due to recursive calculations in the FIFA algorithm, eliminating the knockout/shootout rules is not the same as evaluating the points (not lost in the original algorithm) and discarding them from the final results.

$z_t = \theta_{t,i_t} - \theta_{t,j_t}$, i.e.,

$$\Pr\{Y_t = y_t|\boldsymbol{\theta}_t\} = L(z_t/s; y_t) \tag{13}$$

$$z_t = \boldsymbol{x}_t^\top \boldsymbol{\theta}_t, \tag{14}$$

where $L(z_t/s; y_t)$ is the *likelihood* of $\boldsymbol{\theta}_t$ (for a given outcome y_t) and we define a *scheduling* vector $\boldsymbol{x}_t = [x_{t,0}, \ldots, x_{t,N-1}]^\top$ for the game t, as

$$x_{t,m} = \mathbb{I}[i_t = m] - \mathbb{I}[j_t = m], \tag{15}$$

with $\mathbb{I}[a] = 1$ when a is true and $\mathbb{I}[a] = 0$, otherwise. Thus, $x_{t,m} = 1$ if the team m is playing at home, $x_{t,m} = -1$ if the team m is visiting, and $x_{t,m} = 0$ for all teams m which do not play. This compact notation deals with all the skills $\boldsymbol{\theta}_t$ for each t. As before, s is the scale.

We are interested in the on-line rating algorithms, in which the skills of the participating teams are changed immediately after the results of the game are known. Nevertheless, we will start the analysis with batch processing i.e., assuming that the skills $\boldsymbol{\theta}_t$ do not vary in time, $\boldsymbol{\theta}_t = \boldsymbol{\theta}$. This is a reasonable approach if the time window defined by T is not too large, so that the skills of the teams may, indeed, be considered approximately constant. On-line rating algorithms will then be derived as approximate solutions to the batch optimization problem. The purpose of such an approach is to (i) connect the algorithm used by FIFA to the theoretical assumptions, which are not spelled out when the algorithm is presented, (ii) remove the dependence on the initialization and/or on the scale, and (iii) treat the past and present data in the same manner, e.g., avoiding the partial elimination in the performance metrics, see (9).

Assuming that the observations are independent when conditioned on skills, the rating may be based on the *weighted* maximum likelihood (ML) estimation principle

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{t\in\mathcal{T}} \xi_{c_t} \ell(z_t/s; y_t), \tag{16}$$

where

$$\ell(z_t/s; y_t) = -\log L(z_t/s; y_t) \tag{17}$$

is a (negated)[10] log-likelihood.

The weighting of the log-likelihoods (by ξ_{c_t} in (16) is used in the estimation literature to take care of the model mismatch, (Hu & Zidek, 2001; Amiguet, 2010): less confidence we have that the observations are generated according to the assumed model, smaller weights should be applied.

In our problem, the confidence is associated with the game category c, so smaller ξ_c means that we have less confidence that the games outcomes in the category c are well described by the model (13).

Since multiplication of all ξ_c by a common factor is irrelevant in minimization, we remove any ambiguity by setting again $\xi_0 = 1$.

We may solve (16) using the steepest descent

$$\hat{\boldsymbol{\theta}} \leftarrow \hat{\boldsymbol{\theta}} - \mu/s \sum_t \boldsymbol{x}_t \xi_{c_t} g(z_t/s; y_t), \tag{18}$$

where μ is the adaptation step and

$$g(z; y) = \frac{\mathrm{d}}{\mathrm{d}z}\ell(z; y). \tag{19}$$

The on-line version of (18) is obtained replacing batch-optimization with the stochastic gradient (SG) which updates the solution each time a new observation becomes available, i.e.,

$$\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t - K\xi_{c_t}\boldsymbol{x}_t g(z_t/s; y_t), \tag{20}$$

where the update amplitude is controlled by the weight ξ_{c_t} and the step K which absorbs the scale s.

3.1. Davidson model

The rating now depends on the choice of the likelihood function $L(z; y)$ and here we opt for the Davidson model, (Davidson, 1970), being a particular case of the multinomial model used also in Egidi and Torelli (2021).

$$L(z_t; \mathsf{H}) = \frac{10^{0.5(z_t+\eta b_t)}}{10^{0.5(z_t+\eta b_t)} + \kappa + 10^{-0.5(z_t+\eta b_t)}}, \tag{21}$$

$$L(z_t; \mathsf{A}) = \frac{10^{-0.5(z_t+\eta b_t)}}{10^{0.5(z_t+\eta b_t)} + \kappa + 10^{-0.5(z_t+\eta b_t)}}, \tag{22}$$

$$L(z_t; \mathsf{D}) = \kappa\sqrt{L(z_t; \mathsf{H})L(z_t; \mathsf{A})}, \tag{23}$$

where the home-field advantage (HFA) is modeled as an apparent increase of the skills of the local team by the value η, the indicator $b_t = \mathbb{I}[$ game t is played in the home-team country] allows us to distinguish between the games played on the home or the neutral venues,[11] and κ is used to adjust to the presence of the draws.

[10] The negation in (17) allows us to use a minimization in (16) which is a very common formulation.

[11] Out of $T = 3444$ games we considered, 948 were played on neutral venues. To automatically verify the venues, we used (The Roon Ba, 2022; SoccerWay, 2022); only for the game Djibuti vs. Mauritius played on Nov. 23, 2019, the venue (home) was not registered in the data bases and we found it manually.

Using (21)–(23) in (19), with straightforward algebra we obtain (see Appendix A and (Szczecinski & Djebbi, 2020, Sec. 3.1))

$$g(z;y_t) = \frac{\mathrm{d}}{\mathrm{d}z}\ell(z;y_t) \tag{24}$$

$$= -\ln 10(\check{y}_t - F_\kappa(z)), \tag{25}$$

where $\check{y}_t$ is the "score" of the game which we already defined, and

$$F_\kappa(z_t) = \frac{\frac{1}{2}\kappa + 10^{0.5(z_t+\eta b_t)}}{10^{0.5(z_t+\eta b_t)} + \kappa + 10^{-0.5(z_t+\eta b_t)}}; \tag{26}$$

by using (21)–(23) we immediately see that $F_\kappa(z_t)$ has the meaning of the conditional expected score, $F_\kappa(z_t) = \mathbb{E}[\check{y}_t|z_t] = \sum_{y\in\mathcal{Y}} \check{y} L(z_t;y)$.

Therefore, the SG algorithm (20) becomes

$$\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t + K\xi_{c_t}\boldsymbol{x}_t\big(\check{y}_t - F_\kappa(z_t/s)\big). \tag{27}$$

This *Davidson* algorithm obviously has the form similar to the Elo/FIFA rating algorithm, see (1)–(3), except that we use $F_\kappa(z)$ while the former use $F(z)$. Note that the step K in (27) absorbs the term $\ln 10$ from (25).

It is easy to see that for $\eta = 0$ and $\kappa = 0$ (i.e., when $L(z;\mathsf{D}) = 0$ and the draws are ignored) we have $F_0(z) = F(z)$ which is simply a logistic function as in the Elo algorithm. Furthermore, for $\eta = 0$ and $\kappa = 2$ we obtain $F_2(z) = F(z/2)$ and thus (27) is again equivalent to the Elo rating algorithm, but with the doubled scale value. These observations explain our preference for the model: it leads to a simple algorithm which generalizes the Elo/FIFA rating algorithm, (Szczecinski & Djebbi, 2020).

Although we conclude that the FIFA rating algorithm may be seen as the instance of the maximum weighted likelihood estimation, this is, of course, a "reverse-engineered" hypothesis because the FIFA document, (FIFA, 2018), does not mention any remotely similar concept.

3.2. Regularized batch rating

While our goal is to obtain the on-line rating algorithms where the skills at time $t+1$ are calculated from the observations up to time t, we will, for a moment, ignore this on-line rating aspect and rather focus on the evaluation of the model and the optimization criterion that underlie the algorithms.

We thus concentrate on the original problem defined in (16) for the entire set of data, and, in this way, we (i) will not need to remove a significant portion of the data (meant to eliminate the initialization effects during evaluation, see (9)) and (ii) eliminate the limitation of the SG optimization where, by fine-tuning the adaptation step K, the estimation error is traded off against the convergence speed.

We start by noting that the problem (16) is, in general, ill-posed: since the solution depends only on the differences between the skills, z_t, all solutions $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}} + \theta_o\mathbf{1}$ are equivalent because the differences z_t are independent from "origin" value θ_o. To remove this ambiguity, we may *regularize* the problem as

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \tag{28}$$

$$J(\boldsymbol{\theta}) = \sum_{t\in\mathcal{T}} \xi_{c_t}\ell(z_t/s;y_t) + \frac{\alpha}{2s^2}\|\boldsymbol{\theta}\|^2, \tag{29}$$

where α is the regularization parameter and we have opted for a so-called ridge regularization (Hastie, Tibshirani, & Friedman, 2009, [Ch. 3.4.1]).

Under the model (21)–(23), the regularized batch-optimization problem (28) is useful to resolve another difficulty. Namely, if there is a team m having registered only wins, i.e., when $\forall i_t = m,\ \ y_t = \mathsf{H}$ and $\forall j_t = m,\ \ y_t = \mathsf{A}$, then (16) cannot be solved (or rather, $\hat{\theta}_m \to \infty$) because $J(\boldsymbol{\theta})$ does not limit the value of θ_m. Such a solution not only is unattainable numerically but is, in fact, meaningless, and the regularization (28) settles this issue.[12]

The estimated skills $\hat{\boldsymbol{\theta}}$ now depend on the weights ξ_c, on the regularization parameter α, and on the model parameters η and κ. If not known, all of these parameters must be optimized.

Regarding the optimization criterion, we recall that the FIFA algorithm only specified the expected score, so the quadratic error (8) allowed us to evaluate the algorithm and stay within the boundaries of its definitions. Now, however, with the explicit skills-outcome model, we may go beyond this limitation and will use the prediction metrics known in machine learning such as the (negated) log-score, (Gelman et al., 2014)

$$\mathsf{m}^{\mathrm{ls}}(z_t;y_t) = \ell(z_t/s;y_t), \tag{30}$$

[12] The same problem arises, of course, when a team registers a sequence of pure losses. This is not a hypothetical issue, and in the official FIFA games, three teams registered streaks of unique losses: Tonga (three), Eritrea (two), and American Samoa (four). Thus, the attempt to solve the batch-optimization problem without regularization (i.e., with $\alpha = 0$) would yield $\hat{\theta}_m = -\infty$, with m being the index of any of these teams. Cook Islands played only one game, which we removed from the considerations; see footnote 1.

often preferred due to its compatibility with the log-likelihood used as the optimization criterion, or the accuracy score, (Lasek & Gagolewski, 2020)

$$\mathsf{m}^{\text{acc}}(z_t; y_t) = \mathbb{I}[y_t = \arg\max_y L(z_t/s; y)], \quad (31)$$

which equals one if the event with the largest predicted probability was actually observed; otherwise it is zero.

Furthermore, thanks to the batch-rating, we are able to consider the entire data set in the performance evaluation by averaging the scoring functions (30) or (31) over all games

$$\mathsf{LS} = \frac{1}{T}\sum_{t\in\mathcal{T}} \mathsf{m}^{\text{ls}}\big(\boldsymbol{x}_t^\top \hat{\boldsymbol{\theta}}_{\setminus t}, y_t\big), \quad (32)$$

$$\mathsf{ACC} = \frac{1}{T}\sum_{t\in\mathcal{T}} \mathsf{m}^{\text{acc}}\big(\boldsymbol{x}_t^\top \hat{\boldsymbol{\theta}}_{\setminus t}, y_t\big), \quad (33)$$

where

$$\hat{\boldsymbol{\theta}}_{\setminus t} = \arg\min_{\boldsymbol{\theta}} J_{\setminus t}(\boldsymbol{\theta}), \quad (34)$$

$$J_{\setminus t}(\boldsymbol{\theta}) = \sum_{\substack{l\in\mathcal{T}\\ l\neq t}} \xi_{c_l}\ell(\boldsymbol{x}_l^\top\boldsymbol{\theta}/s; y_l) + \frac{\alpha}{2s^2}\|\boldsymbol{\theta}\|^2. \quad (35)$$

In simple words, for given parameters (α, κ, η, ξ_c), we find the skills $\hat{\boldsymbol{\theta}}_{\setminus t}$ from all, but the t-th game [this is (34)–(35)], and then use them to predict the results y_t; we repeat it for all $t \in \mathcal{T}$, summing the scores obtained. This is the well-known leave-one-out (LOO) cross-validation strategy (Hastie et al., 2009, Sec. 2.9), (Duda et al., 2001, Ch. 9.6.2): no data is discarded when calculating metrics (32)–(33) and this comes with the price of having to find $\hat{\boldsymbol{\theta}}_{\setminus t}$ for all $t \in \mathcal{T}$. To reduce the computational load, we opt here for the approximate leave-one-out (ALO) cross-validation (Rad & Maleki, 2020) based on the local quadratic approximation of the optimization function defined for all data. Details are given in Appendix B.

Although both the average log-score (32) and the accuracy (33) can now be optimized with respect to α, κ, η, and/or ξ_c, we only optimize the log-score whose optimal value is denoted as LS_{opt}; the resulting accuracy, ACC will also be shown.[13] It is, of course, possible to optimize the log-score with respect to any subset of parameters while keeping the rest constant; we will do to this to determine how useful it is to optimize only some parameters.

Again, we use alternate minimization similar to the one shown in (11)–(12): LS was minimized with respect to one parameter at a time: α, κ, η, or ξ_c, until no improvement was observed. This simple strategy led to the minimum LS_{opt} which turned out to be independent of various starting points we used and, although we cannot prove the solution to be global, in all our observations the log-score functions seemed to be unimodal.

Optimized α, κ, η, ξ_c are shown in Table 4 and indicate that

- The data does not provide evidence for using category-dependent weights ξ_c. In fact, the results obtained using the FIFA weights ξ_c are *worse* than those obtained using constant weights $\xi_c = 1$ (i.e., essentially ignoring the possibility of weighting). Although it may be argued that the results are affected by a small number of games in some categories (such as a World Cup), it is very unlikely that observing more games will speak in favor of variable weights and almost surely not in favor of the highly disproportionate weights used in the FIFA algorithm.

  Note that the optimal weights ξ_1 (Friendlies within the IMC) and the weights ξ_2 (Group phase of Nations Leagues) are *smaller* than those of the regular Friendlies. This result stands in contrast with the FIFA algorithm which doubles the weight ξ_1 of the Friendlies played in the IMC and triples the weight of ξ_2. But, of course, we should note a shallow minimum of the objective function which attains the same values $\mathsf{LS}_{\text{opt}} = 0.942$ (for $\eta = 0$ and $\kappa = 2$) and $\mathsf{LS}_{\text{opt}} = 0.856$ ($\eta = 0.3$ and $\kappa = 0.9$) whether

[13] A quick comment may be useful regarding the interpretation of the performance metrics. The accuracy (33) is easily understandable: it is an average number of events that were predicted correctly (as those y which yield the largest likelihood $L(z_t/s; y)$). On the other hand, the metric (32) may be represented as $\exp(-\mathsf{LS}) = [\prod_{t=1}^{T} L(z_t/s; y_t)]^{1/T}$ which is a geometric mean of the predicted probabilities assigned to the events that were actually observed. While the accuracy metric penalizes the wrong guesses with zero (so $\mathsf{ACC} \in [0, 1]$), the log-score penalizes them via the logarithmic function, which may be arbitrarily large (so $\mathsf{LS} \in (0, \infty)$).

However, the fundamental difference between the two metrics is that we can use the accuracy without specifying the distribution for all possible outcomes, but we cannot calculate the log-score in such a case.

We also note that the common confusion is to interpret the function $F(z_t/s)$ in the Elo/FIFA algorithm as the probability of the home win, and the value $1 - F(z_t/s)$, as the probability of an away win. This, of course, implies that the draw probability is equal to zero. With this interpretation, we can still calculate the accuracy metric even if we never predict the draw. On the other hand, we cannot calculate the log score because, when the draw occurs, we have an undefined metric $\mathsf{m}^{\text{ls}}(z_t/s; \mathsf{D}) \to \infty$.

Table 4
Batch-rating parameters obtained via minimization of the log-score (32). The parameters (α, κ, η, ξ_c) are either fixed (shadowed cells), or obtained via optimization. The upper-part results correspond to the conventional FIFA algorithm: using $\kappa = 2$ and $\eta = 0$, the expected score is calculated using a logistic function. The last line corresponds to the parameters η and κ obtained via (36)–(37).

LS_{opt}	α	η	κ	ξ_0	ξ_1	ξ_2	ξ_3	ξ_4	ξ_5	ξ_6	ξ_7	ξ_8	ACC[%]
0.954	0.9	0	2.0	1	2.0	3.0	5.0	5.0	7.0	8.0	10.0	12.0	56
0.942	0.2	0	2.0	1	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	55
0.942	0.2	0	2.0	1	0.9	0.7	0.8	0.9	1.1	0.8	0.8	1.1	55
0.912	0.2	0.4	2.0	1	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	57
0.856	0.3	0.3	0.9	1	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	61
0.856	0.4	0.3	0.9	1	0.9	0.7	0.8	1.1	1.1	0.9	1.0	1.1	61
0.864	0.3	0.3	0.6	1	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	61

we use the equal $\xi_c = 1$ or the optimized ξ_c.

- A notable improvement in the prediction capacity measured by the log-score is obtained by considering the HFA. The value $\eta \in \{0.3, 0.4\}$ emerges from the optimization fit and we note that $\eta = 0.25$ was used in eloratings.net (2020).[14]
- A more important improvement is obtained by optimizing the parameter κ which takes into account the draws and their frequency as discussed in Szczecinski and Djebbi (2020).

Here, it is interesting to compare the parameters we found by optimization with the simplified formulas proposed in Szczecinski and Djebbi (2020, [Sec. 3.2])

$$\eta = \log_{10} \frac{f_{\mathsf{H}}}{f_{\mathsf{A}}} \tag{36}$$

$$\kappa = \frac{f_{\mathsf{D}}}{\sqrt{f_{\mathsf{H}} f_{\mathsf{A}}}} \approx \frac{2 f_{\mathsf{D}}}{1 - f_{\mathsf{D}}}. \tag{37}$$

where f_y, $y \in \mathcal{Y}$ are empirical frequencies of outcomes. We can consider separately the games played on the neutral venues; the frequencies are obtained from the data, $f_{\mathsf{D}}^{\text{neut.}} = 0.24$, and those played on home venues as $f_{\mathsf{A}}^{\text{hfa}} = 0.26$, $f_{\mathsf{D}}^{\text{hfa}} = 0.23$, $f_{\mathsf{H}}^{\text{hfa}} = 0.51$, which yields

$$\kappa^{\text{hfa}} = 0.61 \qquad \eta^{\text{hfa}} = 0.29 \tag{38}$$

$$\kappa^{\text{neut.}} = 0.63 \tag{39}$$

The parameter η^{hfa} predicted by (36) is practically equal to the one obtained by optimization. And while the parameters κ^{hfa} and $\kappa^{\text{neut.}}$ are slightly different from the one predicted by (37), using them in the rating, we obtained $\mathsf{LS}_{\text{opt}} = 0.864$ (see last line in Table 4), which is still notably better than using the conventional FIFA rating. This is interesting because finding the parameters η and κ from the frequencies of the games not only avoids optimization, but also provides a simple-to-verify origin of the values used in the algorithm.

[14] Therein, the unnormalized value $\eta s = 100$ is reported and since $s = 400$, we obtain $\eta = 0.25$.

4. Margin of victory

In the search for a possible improvement of the rating, we now want to consider the use of the MOV variable, defined by the difference of the goals scored by each team, denoted by d_t. In this regard, the most recent works adopt two conceptually different approaches.

The first keeps the structure of the known rating algorithm (such as the FIFA algorithm) and modifies it by changing the adaptation step size as a function of d_t. This was already done in eloratings.net (2020), Hvattum and Arntzen (2010), Silver (2014), Ley et al. (2019), and Kovalchik (2020), and is conceptually similar to the weighting according to the game-category we consider in the previous section.

The second approach, studied before in Maher (1982), Ley et al. (2019), Lasek and Gagolewski (2020), and Szczecinski (2022), changes the model between the skills and the MOV variable d_t. We will focus on the simple proposition from Lasek and Gagolewski (2020) based on the formulation of of Karlis and Ntzoufras (2008).

4.1. MOV via weighting

For context, in Table 5 we show the number of games depending on the value of the MOV variable d. While, in principle, it is possible to use directly d,

it is customary to consider their absolute value, $|d|$.

The Elo/FIFA algorithms (27) can be easily modified as follows, to take into account the MOV variable:

$$K_{c,d} = K\xi_c\zeta_d, \tag{40}$$

where, as before, K is the common step, ξ_c is the weight associated with the game-category c, and ζ_d is the function of the MOV-variable d.

Integrating the MOV-weight into the online rating defined in (27) (by replacing, therein $K\xi_{c_t}$ with $K\xi_{c_t}\zeta_{v_t}$) yields the *Davidson-MOV* algorithm.

For example, (eloratings.net, 2020) uses

$$\zeta_d = \begin{cases} 1 & |d| \le 1 \\ 1.5 & |d| = 2 \\ 1.75 + 0.125(|d| - 3) & |d| \ge 3 \end{cases}. \tag{41}$$

Similar propositions can be found in Hvattum and Arntzen (2010) (in the context of association football), in Kovalchik (2020) (to rate tennis players), or in Silver (2014) (for the rating of teams in American football).

To elucidate how useful such heuristics are, we note that the problem is very similar to the importance weighting we analyzed before; the difference lies in the fact that the weighting now depends on the product $\xi_c\zeta_d$. Therefore, we may reuse our optimization strategy to find the optimal weights for games with different values of $|d|$.

To this end, we discretize $|d|$ into $V+1$ MOV-categories, $v = 0, \ldots, V$ and we use a very simple mapping $v = |d|$ for $v < V$ and $v = V \iff |d| \ge V$. For example, with $V = 2$, ζ_0 weights the draws ($|d| = 0$), ζ_1 weights the games with one goal difference ($|d| = 1$) and ζ_2 weights the games with more than one goal difference ($|d| \ge 2$).

Breaking with the predefined functional relationship as shown in (41) we are more general than the latter, e.g., treating the cases $|d| = 0$ and $|d| = 1$ separately. This makes sense since these events are not only the most frequent ones (covering, respectively, 23% and 36% of the total; see Table 5), but also correspond to the events of draw and win/loss treated differently by the algorithm.

On the other hand, we are also less general due to the merging of events $|d| \ge V$, although this effect will decrease with V, simply because there will be very few observations, as may be understood from Table 5. For example, with $V = 4$, the weighting ζ_4 will be the same for the events with $|d| = 4$ and $|d| > 4$ but the latter make only about 5% of the total.

We again consider the game categories defined in Table 1 and thus we now solve the following problem:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{t\in\mathcal{T}} \xi_{c_t}\zeta_{v_t}\ell(z_t/s; y_t) + \frac{\alpha}{2s^2}\|\boldsymbol{\theta}\|^2, \tag{42}$$

where v_t is the index of the MOV variable d_t. Again, to remove ambiguity of the solution, we set $\xi_0 = 1$ and $\zeta_0 = 1$.

Parameters ξ_c, ζ_v, η, κ, and α will be optimized again using the ALO approach we described in Section 3.2, that is, by minimizing the log-score criterion (32) using an alternate optimization similar to that defined in (11)–(12). The results shown in Table 6 allow us to conclude that:

- The weighting of the MOV-categories is more beneficial than the weighting of the game-categories: by optimizing the MOV-weights ζ_v (and keeping $\xi_c = 1$) yields $\mathsf{LS}_{\text{opt}} = 0.937$ while the optimization of ξ_c (keeping $\zeta_v = 1$) yields $\mathsf{LS}_{\text{opt}} = 0.948$ (see Table 4).
- Optimization indicates that ζ_v defined by (41) are suboptimal. In particular, the optimal MOV weights, ζ_v are monotonically growing (as foreseen by the heuristic (41)) only for $|d| \ge 1$ while the draws (i.e., $|d_t| = 0$) have a weight that is more important than the weights of the events $|d_t| = 1$; thus, these two events ($d_t = 0$ and $|d_t| = 1$) should not be combined, nor should we impose a particular functional form on the weights ζ_v.
- The best prediction improvement is obtained again by optimizing the parameters η and κ of the Davidson model together with the MOV weights ζ_v.

Table 5
Number of games $T_{|d|}$ which finished with the goal difference $|d|$ and their relative frequency $f_{|d|} = T_{|d|}/T$

	$\lvert d\rvert = 0$	$\lvert d\rvert = 1$	$\lvert d\rvert = 2$	$\lvert d\rvert = 3$	$\lvert d\rvert = 4$	$\lvert d\rvert = 5$	$\lvert d\rvert > 6$
$T_{\lvert d\rvert}$	791	1248	640	379	190	87	109
$f_{\lvert d\rvert}$	23%	36%	19%	11%	5.5%	2.5%	3%

Table 6

Batch-rating parameters obtained via minimization of the log-score (32) with weighting of the MOV-variables. The parameters (α, κ, η, $\boldsymbol{\xi}$, ζ_v) are either fixed (shadowed cells), or obtained through optimization; for $d > V$, the parameters ζ_d are not defined and the corresponding cells are indicated with "×". To save space, in the sole case when the parameters ξ_c are optimized, their optimal values are gathered in the vector $\hat{\boldsymbol{\xi}} = [1.0, 1.2, 1.0, 1.2, 1.2, 1.7, 2.4, 1.5, 5.2]$. The first line corresponds to the weights defined in (41) except for ζ_6 which cannot be fixed as it corresponds not only to the event $|d_t| = 6$ but also weights all the events $|d_t| > 6$

LS_{opt}	V	α	η	κ	$\boldsymbol{\xi}$	ζ_0	ζ_1	ζ_2	ζ_3	ζ_4	ζ_5	ζ_6	ACC[%]
0.944	6	0.7	0	2.0	1	1	1	1.5	1.75	1.875	2.0	4.2	56
0.933	6	0.2	0	2.0	1	1	0.3	0.5	0.7	1.1	1.5	2.6	55
0.931	6	0.2	0	2.0	$\hat{\boldsymbol{\xi}}$	1	0.3	0.5	0.7	1.0	1.5	2.4	55
0.903	6	0.2	0.4	2.0	1	1	0.4	0.6	0.8	1.1	1.7	3.2	57
0.849	6	0.3	0.3	0.9	1	1	0.4	0.6	0.8	1.1	1.7	3.0	62
0.850	4	0.3	0.3	0.9	1	1	0.4	0.6	0.8	1.5	×	×	61
0.852	2	0.3	0.3	0.9	1	1	0.4	0.8	×	×	×	×	61
0.854	1	0.2	0.3	0.9	1	1	0.7	×	×	×	×	×	61

4.2. MOV via modeling

The MOV modelling consists in defining a formal relationship between the skills $\boldsymbol{\theta}_t$ and the observed MOV variable d_t, where a simple approach proposed in Karlis and Ntzoufras (2008) relies on a modeling of the goal difference using the Skellam distribution

$$\Pr\{d_t = d|\boldsymbol{\theta}_t\} = L(z_t; d) \tag{43}$$

$$= \mathrm{e}^{-(\mu_{\mathrm{h},t}+\mu_{\mathrm{a},t})} \left(\frac{\mu_{\mathrm{h},t}}{\mu_{\mathrm{a},t}}\right)^{d/2} I_{|d|}(2\sqrt{\mu_{\mathrm{h},t}\mu_{\mathrm{a},t}}), \tag{44}$$

where $I_v(\cdot)$ is the modified Bessel function of order v and $\mu_{\mathrm{h},t}$ and $\mu_{\mathrm{a},t}$ are means of the Poisson variables modeling the home- and away- goals. The latter are functions of the skills difference z_t, (Karlis & Ntzoufras, 2008, [Sec. 2.2])

$$\mu_{\mathrm{h},t} = \mathrm{e}^{c+z_t+b_t\eta}, \quad \mu_{\mathrm{a},t} = \mathrm{e}^{c-z_t-b_t\eta}, \tag{45}$$

where c is a constant and, as before, η is the HFA coefficient.[15]

The model (43) is a particular case of a more general form shown in Karlis and Ntzoufras (2008), which models the offensive and the defensive skills. Here, however, we are interested in rating and thus one skill per team should be used. As noted in Ley et al. (2019) and in Lasek and Gagolewski (2020) this offers a sufficient prediction capacity avoiding the problem of overparameterization due to the doubled number of skills.

Using (45) in (43), the following log-likelihood is obtained:

$$\ell(z_t; d_t) = -\log L(z_t; d_t) \tag{46}$$

$$= (\mu_{\mathrm{h},t} + \mu_{\mathrm{a},t}) - d_t(z_t + b_t\eta) - 2\mathrm{e}^c - \log \tilde{I}_{|d_t|}(2\mathrm{e}^c) \tag{47}$$

where, for numerical stability, it is convenient to use an exponentially modified form of the Bessel function, $\tilde{I}_v(u) = I_v(u)\mathrm{e}^{-u}$, available in many computation packages.

The derivative of (46) is given by

$$g(z; d) = \frac{\mathrm{d}}{\mathrm{d}z}\ell(z; d) = -(d - \overline{F}(z)), \tag{48}$$

$$\overline{F}(z) = \mu_{\mathrm{h}} - \mu_{\mathrm{a}} = \mathrm{e}^c(\mathrm{e}^{z+b\eta} - \mathrm{e}^{-z-b\eta}). \tag{49}$$

The batch rating then consists in solving the following problem:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{t\in\mathcal{T}} \ell(z_t/s; d_t) + \frac{\alpha}{2s^2}\|\boldsymbol{\theta}\|^2 \tag{50}$$

and the SG implementation of the ML principle will produce the *Skellam* algorithm

$$\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t + K\boldsymbol{x}_t\big(d_t - \overline{F}(z_t/s)\big), \tag{51}$$

which is again written in a form similar to the FIFA rating algorithm, where the goal difference d_t plays the role of the "score", and $\overline{F}(z_t/s) = \mathbb{E}[d_t|z_t]$ is the expected score. The Skellam algorithm (51) can also be obtained by applying the Poisson model to the goals scored by each of the teams (Lasek & Gagolewski, 2020).

[15] We can rewrite (45) as $\mu_{\mathrm{h},t} = \mathrm{e}^{c'_t+z_t+b_t\eta'}$, $\mu_{\mathrm{a},t} = \mathrm{e}^{c'_t+z_t}$ with $c'_t = c - b_t\eta$ and $\eta' = 2\eta$. Furthermore, if we have $b_t \equiv 1$ (as in league games that are not played on neutral venues) then $c'_t \equiv c' = c - \eta$ is independent of the game index t and our notation is identical with the one proposed in (Karlis & Ntzoufras, 2008, [Eq. (2.2)-(2.3)] and used also in (Lasek & Gagolewski, 2020). However, international FIFA games can be played on neutral venues and we need c'_t to depend on the game index. Since this is not possible with the notation of Karlis and Ntzoufras (2008), our model is preferred.

Table 7
Batch-rating parameters obtained via minimization of the log-score (32) using the Skellam model (46).

LS_{opt}	α	η	c	ACC[%]
0.843	0.05	0.2	−0.02	62

To calculate the log-score, we have to calculate the probabilities $\mathsf{m}^{\text{ls}}(z_t;\mathsf{A}) = -\log \Pr\{d_t < 0\}$ (away-win) and $\mathsf{m}^{\text{ls}}(z_t;\mathsf{H}) = \Pr\{d_t > 0\}$ (home-win). Since the closed-form formulas do not exist, we use truncated sums

$$\mathsf{m}^{\text{ls}}(z;\mathsf{A}) = -\log \sum_{d=-D}^{-1} L(z;d), \tag{52}$$

$$\mathsf{m}^{\text{ls}}(z;\mathsf{D}) = -\log L(z;0), \tag{53}$$

$$\mathsf{m}^{\text{ls}}(z;\mathsf{H}) = -\log \sum_{d=1}^{D} L(z;d), \tag{54}$$

where we applied $D = 50$ which we empirically verified to satisfy $1 - \sum_{d=-D}^{D} L(z;d) < 10^{-4}$.

The results shown in Table 7 indicate that, with this very simple approach (only two parameters of the model which must be optimized), we are able to improve over the MOV-weighting strategy and this should be attributed to the use of a formal skills-outcome model. The price to pay for the improvement lies in abandoning the legacy of the Elo algorithm.

Moreover, possible implementation issues may arise since the expected score (49) is theoretically unbounded. Thus, whether the improvement of the log-score from $\mathsf{LS} = 0.849$ (in the MOV-weighting, see Table 6) to $\mathsf{LS} = 0.843$ in the Skellam MOV model is worth the change and the implementation risks is at least debatable.

5. On-line rating

Before starting a metrics-based comparison of the on-line algorithms, in Section 5.1 we will address the practical issue of setting the scale.

5.1. Scale adjustment

The scale is obviously irrelevant in batch optimization, and the on-line update can also be written in a scale-invariant manner by dividing (20) by s:

$$\boldsymbol{\theta}'_{t+1} \leftarrow \boldsymbol{\theta}'_t - K'\xi_{c_t}\boldsymbol{x}_t g(z'_t; y_t) \tag{55}$$

$$z'_t = z_t/s \tag{56}$$

$$\boldsymbol{\theta}'_t = \boldsymbol{\theta}_t/s \tag{57}$$

$$K' = K/s; \tag{58}$$

in other words, we will obtain the same results $\boldsymbol{\theta}'_t$ as long as we use the same initial $\boldsymbol{\theta}'_0$ and the same step K'. In particular, with an all-zero initialization of the skills, i.e., $\boldsymbol{\theta}_t = \mathbf{0}$ and using $K = sK'$ we will obtain the same scaled results $\boldsymbol{\theta}_t = s\boldsymbol{\theta}'_t$.

However, in the FIFA ranking, a non-zero initialization $\boldsymbol{\theta}_0$ was determined in advance (see footnote 3) so $\boldsymbol{\theta}'_0$ is not scale-invariant. Thus, given the initialization at hand, the question is how to determine the scale. We do not know any clear answer, but an insight into finding the useful scale value may be gained assuming that the initialization corresponds to the "optimal" solution, e.g., $\hat{\boldsymbol{\theta}}$ obtained in batch optimization with a given scale s_0.

It is easy to see that using $s > s_0$ will force the algorithm to change significantly $\boldsymbol{\theta}_t$ (attainable with large values of the adaptation step, K); the same will happen for $s < s_0$ because the optimal estimates $\boldsymbol{\theta}_t$ will have to be scaled down.

Since scaling the skills up/down changes their empirical moments, we suggest choosing the scale s in a moment-preserving manner. To this end, we define the empirical standard deviation of the skills

$$\sigma_t = \sqrt{\|\hat{\boldsymbol{\theta}}_t - \bar{\hat{\boldsymbol{\theta}}}_t\|^2/M} \tag{59}$$

where $\bar{\hat{\boldsymbol{\theta}}}_t = (\sum_{m=1}^{M} \hat{\theta}_{t,m})/M$ is the empirical mean and postulate that, at the initialization and at the final step, we have $\sigma_0 \approx \sigma_T$.

In fact, the initialization used by FIFA yields $\sigma_0 = 220$ and, after running the original FIFA algorithm, we obtain $\sigma_T = 252$.

Changing the scale s, we will obtain different σ_T so the idea is to run the algorithms for different values of the scale s (e.g., for multiples of 50) and to choose the one that produces a standard deviation $\sigma_T \approx \sigma_0$. In practice we might do it using historical data *before* the new rating is deployed.

In this manner we found $s = 150$ to be suitable for the Davidson algorithm: we obtained $\sigma_T \approx 220$ when $\kappa = 2$ and $\sigma_T \approx 210$ for $\kappa = 1$.

This indicates that the scale $s = 600$ was too large for the FIFA rating. This can be seen by comparing the result of the FIFA rating with $\xi_c = 1$ (in Table 8a) to the results of the Davidson algorithm (with $\eta = 0$

Table 8
Parameters and performance of the on-line rating algorithms obtained by minimizing the log-score (60) for a) FIFA algorithm, b) Davidson algorithms, c) Davidson-MOV algorithm from Section 4.1, and d) Skellam algorithm from Section 4.2

weights	LS_{opt}	K	η	κ	ACC [%]	$\mathsf{LS}_{\setminus\text{so\&ko}}$
ξ_c from Table 1	0.975	5	0	2	48	0.980
$\xi_c \equiv 1$	0.952	55	0	2	50	0.955

a) FIFA algorithm, $s = 600$. The log-score obtained by removing the knockout/shootout rules is indicated by $\mathsf{LS}_{\setminus\text{so\&ko}}$.

LS_{opt}	K	η	κ	ACC[%]
0.939	35	0	2	54
0.915	35	0.4	2	57
0.875	35	0.3	1.0	60

b) Davidson algorithm, $s = 150$.

LS_{opt}	V	K	η	κ	ζ_0	ζ_1	ζ_2	ζ_3	ACC[%]
0.864	1	40	0.3	0.9	1.0	0.9	×	×	60
0.863	2	45	0.3	0.9	1.0	0.6	1.1	×	60
0.862	3	40	0.3	0.9	1.0	0.7	0.9	1.5	60

c) Davidson-MOV algorithm, $s = 200$.

LS_{opt}	K	η	c	ACC[%]
0.851	7.5	0.2	−0.07	60

d) Skellam algorithm, $s = 300$.

and $\kappa = 2$). Both algorithms are essentially the same (although FIFA uses the shootout/knockout rules, which have rather small impact on performance) and the main difference resides in the scale. Since the Davidson algorithm ($\eta = 0$, $\kappa = 2$) with the scale $s = 150$ is equivalent to the FIFA algorithm with the scale $s = 300$, this latter scale value would ensure a better performance of the FIFA rating. However, this effect appears only due to the limited observation time we have at our disposal and will vanish after a sufficiently large number of games.

Similarly, for the Davidson-MOV algorithm, using $s = 200$, and for different values of V we obtained $\sigma_T \approx 220$, while using the scale $s = 300$ in the Skellam algorithms yields $\sigma_T = 225$.

5.2. *Evaluation of the algorithms*

To evaluate the SG algorithms, we used the same methodology we applied to make a preliminary evaluation of the FIFA rating in Section 2.1. That is, we used the first (approximate) half of the observation period for initialization and the second half is used to calculate the performance metrics. The difference from Section 2.1 is that now we use the log-score and the accuracy metrics

$$\mathsf{LS}' = \frac{1}{T - T'} \sum_{t=T'+1}^{T} \mathsf{m}^{\text{ls}}(z_t, y_t) \tag{60}$$

$$\mathsf{ACC}' = \frac{1}{T - T'} \sum_{t=T'+1}^{T} \mathsf{m}^{\text{acc}}(z_t, y_t), \tag{61}$$

where T' is defined after (9).

We consider the original and modified FIFA algorithm (Table 8a), the Davidson algorithm (Table 8b), the Davidson-MOV algorithm (Table 8c), and the Skellam algorithm (Table 8d).

In all cases, but in the original FIFA algorithm, we ignore the game-category weighting (i.e., we use $\xi_c \equiv 1$) because, as we have already shown, its effect is negligible. This is clearly shown in the first row of Table 8a where we see that, using the FIFA weighting, we obtain worse results than when the weighting is ignored. This is essentially the same result as the one we have shown in Table 2 but we repeat it here to show the log-score metric which we could not calculate without first introducing the Davidson model underlying the FIFA algorithm.

In Table 8a we also show the log-score $\mathsf{LS}_{\setminus\text{so\&ko}}$ obtained applying the FIFA algorithm but removing the shootout and knockout rules. We observe that the knockout rule improves the prediction in terms of the log-score which confirms our preliminary evaluation in Section 2.2 based on the MSE. Nevertheless, the improvement is minor, and a much more important decrease of the log-score may be obtained by changing the model as indicated by the remaining results. In particular

- The most notable improvements are due to, in similar measures, two elements: the introduction of the HFA coefficient η and the explicit use of the Davidson model (and thus the optimization of the coefficient κ).
- Additional small but still perceivable gains are obtained by introducing the MOV-weighting, where from the lesson learned in Section 4.1 we independently weight the draws and the home/away wins. It is sufficient to use only two weights ($V = 1$), i.e., the very concept of the MOV is, de facto reduced to a distinction between the draws and the howe/away wins.
- The MOV-modeling using the Skellam distribution again brings a small benefit.

Table 9
Ranking of the top teams using the algorithms compared in Table 8: FIFA with $\xi_c \equiv 1$, $K = 55$, Davidson with $K = 35$, $\eta = 0.3$, $\kappa = 1.0$, Davidson-MOV with $V = 1$, $K = 40$, $\eta = 0.3$, $\kappa = 0.9$, and Skellam with $K = 7.5$, $\eta = 0.2$, $c = -0.07$

FIFA	Davidson	Davidson-MOV	Skellam
BRA (1922.5)	FRA (1658.0)	FRA (1702.4)	BRA (1722.8)
BEL (1919.0)	ARG (1650.2)	BRA (1693.8)	ARG (1660.1)
ENG (1904.2)	BRA (1649.5)	ARG (1693.5)	ENG (1653.8)
FRA (1903.9)	ENG (1628.8)	ENG (1667.9)	FRA (1638.7)
ARG (1879.8)	BEL (1609.9)	BEL (1650.9)	ESP (1622.6)
ITA (1837.8)	ESP (1608.0)	ESP (1643.6)	BEL (1615.7)

We present in Table 9 the rating obtained for the top teams through new rating algorithms. Of course, due to the different scales that we used, the skills obtained with different algorithms cannot be compared directly.

We emphasize that the quality of these rankings (i.e., ordered skills) cannot be assessed because there is no reference order of the teams to which the shown rankings can be compared. The only tool we have to assess their validity is to calculate the performance criteria as we did in Table 8 using the log-score.

Noting that even rather mild differences between the Davidson and the Davidson-MOV algorithms alter the final order/ranking, Table 9 should be treated as a cautionary illustration that the ranking/order of the team can be very easily changed by relatively benign modifications of the rating algorithm. With that caveat, the different algorithms based on different models consistently put the same group at the top of the list. In fact, the algorithms are rather consensual regarding the current (as of March 31, 2002) official top team, BRA which, or remains on the top of the list, or has skills within fraction of percentage of top team's rating. On the other hand, BEL's second position in the official ranking (see Table 3) is much more questionable. While the second spot is preserved with the optimized FIFA-like algorithm (the first column of Table 9 where we use $K = 55$ and $\xi_c \equiv 1$, but the knockout and shootout rules are kept), the new algorithms consistently demote BEL to the fifth and lower position.

6. Conclusions

In this work, we analyze the FIFA ranking using the methodology conventionally used in probabilistic modeling and statistical inference. In the first step, we made a preliminary evaluation of the algorithm using the probabilistic concepts explicitly used in the FIFA description. In this way, we were already able to question the need for the weighting of the outcomes which depends on the FIFA-defined game category.

We also evaluate the heuristic shootout/knockout rules that are used in the FIFA rating. We concluded that since their impact on overall performance is small and they may distort the relationship between the ratings of the strong teams, which often face each other in the final stages of the competitions, their usefulness is questionable.

To go beyond the limitation of the rudimentary probabilistic concepts of the FIFA algorithm, we identified the model that relates the game results to the parameters that must be optimized (skills). More precisely, we have shown that the FIFA algorithm can be formally derived as the stochastic gradient (SG) optimization of the weighted maximum likelihood (ML) criterion in the Davidson model (Davidson, 1970).

This step allows us to define the performance metrics related to the predictive performance of the algorithms we study. This is particularly important in the case of the FIFA ranking algorithm, which does not model the outcomes of the game but only explicitly specifies the expected score; and this is not sufficient to accurately assess the rating results. It also allows us to apply the batch approach to rating and skills estimation. This conventional machine learning strategy frees us from considerations related to scale, initialization, or modeling of skills dynamics.

Using the batch rating, we have shown that the weighting dependent on the game-category is negligible at best, and counterproductive at worst, which is the case of the weighting used by the FIFA rating. This observation is interesting in its own right because, while on the one hand the concept of weighting is used in the rating literature, e.g., (Ley et al., 2019), on the other hand, the literature does not show any evidence that it is in any way beneficial and our findings consistently indicate the contrary.

Next, we consider extensions of the algorithm by including the home-field advantage (HFA) and optimizing the parameter responsible for the draws. These two elements seem to be particularly important from the point of view of the performance of the rating algorithm. While the HFA is well-known and is part of FIFA Womens' rating (FIFA, 2007), the possibility of generalizing the Elo algorithm by using the Davidson model has only recently been shown in in Szczecinski and Djebbi (2020).

We also evaluated the possibility of using the margin of victory (MOV) given by the goal differential: we analyzed the inclusion of the MOV

through weighting, as well as the explicit modeling of the MOV variables using the Skellam distribution. These two methods further improve the results at the cost of greater complexity. Here, optimization of the weights also yields interesting and somewhat counterintuitive results. That is, we have shown that games won with a small margin should have smaller weights than the draws. This stands in stark contrast with the weighting strategies proposed before, e.g., by Hvattum and Arntzen (2010), Silver (2014), or by Kovalchik (2020) which, using monotonically increasing functions of the MOV variable, do not allow for a separate treatment of the draws.

6.1. Recommendations

Given the analysis and the observations we made, if the FIFA rating is to be changed, the following steps are recommended:

(1) Add the home-field advantage (HFA) parameter to the model because playing at the home venue is a strong predictor of victory. This well-known fact is already exploited in Women's FIFA ranking, and such a modification is most likely the simplest and the least debatable element. In our opinion, it is surprising that the current rating adopted in 2018 does not include the HFA. The HFA can be obtained through optimization, or it can be calculated using the simple formula (36).
(2) Use an explicit model to relate skills to outcomes. Not only is the expressiveness increased by providing the explicit probability for the draws, but also the prediction results are improved. Note that the rating algorithm introduced recently by FIVB adopts this approach and specifies the probability for each of the game outcomes. In the context of the FIFA ranking, the Davidson model we used in this work is an excellent candidate for that purpose: it relies on a natural generalization of the Elo algorithm, preserving the legacy of the current algorithm. Again, to find the parameter of the model, we may use optimization or the simple formula (37).
(3) Remove the weighting of the games according to their assumed importance because the data does not provide any evidence for their utility, or rather provides the indication that the weighting in its current form is counterproductive. If the concept of the game-importance is of extra-statistical nature (such as entertainment), it is preferable to diminish its role, e.g., by shrinking the gap between the largest and the smaller values of ξ_c used.
(4) Remove the shootout and knockout rules that are not rooted in any solid statistical principle.

    As far as the knockout rule is concerned, its beneficial effect on the prediction quality is negligible comparing to the advantages of using the HFA and the draw model. Regarding the shootout rule, from a rating perspective, we recommend that shootouts be treated as draws.

    Overall, a small frequency of events when the shootout/knockout rules can be applied, and a marginal change in the obtained score, make their impact negligible and their fairness is very debatable although the heuristics behind the knockout deserves more study.
(5) If the rating was to consider the MOV, the simplest solution would be to weight the update step using the goal differential. On the other hand, modification of the model to the Skellam distribution may cause numerical problems, and relatively small performance gains hardly justify the added complexity.

6.2. Further work

The analysis we carried out in this work should not be considered exhaustive by any means and was meant (i) to provide an understanding of the current FIFA rating and (ii) to propose the simplest and yet meaningful modifications of the current algorithm.

Our recommendations regarding further work on the improvement of the rating are the following:

(1) Beside the simple weighting strategies we analyzed here, to deal with the MOV we should consider alternative solutions similar to those already considered in Women's teams FIFA ranking. Again, the latter should be studied, e.g., using the methodology we used in this work and basing the results on a formal probabilistic model.
(2) To improve the tracking capabilities of the algorithm and to reduce its sensitivity to the randomness of the game outcomes, we should consider the Bayesian estimation methods proposed previously in Glickman (1999) (Glicko algorithm) and in Herbrich and Graepel (2006) (True Skill algorithm). These algorithms explicitly estimate the reliability

of skills estimates, which improves the predictive capacity and provides a more nuanced interpretation of the rating.

However, it should be noted that the Glicko and True Skill algorithms are not model-agnostic, and using them with the Davidson model, as we postulated here, is not straightforward.[16] In this regard, the recent formulation in Szczecinski and Tihon (2021) which is applicable in any model, may streamline the development.

(3) We should design a uniform treatment to deal with teams that play very infrequently, which may happen naturally (e.g., due to geographic isolation) or be done intentionally (to preserve the ranking position). Among the possible venues to deal with this issue is (i) an automatic rating-point penalty similar to what is done in the FIVB ranking and/or (ii) a decrease of estimation reliability similar to what is done in Glicko and True Skill algorithms.

(4) Adding new teams to the rating should be handled with more care. This issue was highlighted by the recent (March 2022) reintroduction of the Cook Islands to the rating after many years without playing any FIFA-recognized game. In fact, many national teams, e.g., among those already recognized by CONCACAF may, at some point, be also recognized by FIFA which will have to decide on their initial rating. In this case, taking into account games that are not recognized by FIFA is likely the most efficient approach.

References

Amiguet, M., 2010. Adaptively weighted maximum likelihood estimation of discrete distributions (Unpublished doctoral dissertation). Universtité de Lausanne.

Barber, D., 2012. *Bayesian reasoning and machine learning*. Cambridge University Press.

Burn, R., 2020. Optimizing approximate leave-one-out cross-validation to tune hyperparameters. ArXiv, abs/2011.10218. Retrieved from http://arxiv.org/abs/2011.10218

Davidson, R.R., 1970. On extending the Bradley-Terry model to accommodate ties in paired comparison experiments, *Journal of the American Statistical Association* 65(329), 317-328. Retrieved from http://www.jstor.org/stable/2283595

Duda, R.O., Hart, P.E., & Stork, D.G., 2001. Pattern classification. John Wiley & Sons.

Egidi, L., & Torelli, N., 2021. Comparing goal-based and result-based approaches in modelling football outcomes, *Social Indicators Research* 156(2), 801-813. Retrieved from https://doi.org/10.1007/s11205-020-02293-z

eloratings.net., 2020. World football Elo ratings. Retrieved Oct. 8, 2020, from https:// www.eloratings.net/

FIFA., 2007. *FIFA/Coca-Cola women's world ranking*. Retrieved November 12, 2021, from https://digitalhub.fifa.com/m/3d9cb1decbbb2ac7/original/rxqyxdjhbs2qdtstluy6-pdf.pdf

FIFA., 2018. *Revision of the FIFA/Coca-Cola world ranking*. Retrieved February 7, 2020, from https://digitalhub.fifa.com/m/f99da4f73212220/original/edbm045h0udbwkqew35a-pdf.pdf

FIVB., 2020. *New senior world rankings*. Retrieved December 6, 2021, from https://www.fivb.com/en/volleyball/rankings

Football Rankings., 2021 Football rankings. Retrieved October 28, 2021, from http://www.football-rankings.info/

Gelman, A., Hwang, J., & Vehtari, A., 2014. Understanding predictive information criteria for Bayesian models, *Statistics and Computing* 24(6), 997-1016. Retrieved from https://doi.org/10.1007/s11222-013-9416-2

Glickman, M.E., 1999. Parameter estimation in large dynamic paired comparison experiments, *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 48(3), 377-394. Retrieved from http://dx.doi.org/10.1111/1467-9876.00159

Hastie, T., Tibshirani, R., & Friedman, J., 2009. *The elements of statistical learning*. Springer Series in Statiscs.

Herbrich, R., & Graepel, T., 2006. TrueSkill(TM): A Bayesian skill rating system (Tech. Rep.). Retrieved from https://www.microsoft.com/en-us/research/publication/trueskilltm-a-bayesian-skill-rating-system-2/

Hu, F., & Zidek, J.V., 2001. The relevance weighted likelihood with applications. In Ahmed, S.E., & Reid, N. (Eds.), Empirical Bayes and likelihood inference, New York, NY: Springer New York, pp. 211-235. Retrieved from https://doi.org/10.1007/978-1-4613-0141-713

Hvattum, L.M., & Arntzen, H., 2010. Using Elo ratings for match result prediction in association football, *International Journal of Forecasting* 26(3), 460-470. Retrieved from http://www.sciencedirect.com/science/article/pii/S0169207009001708 (Sports Forecasting)

Karlis, D., & Ntzoufras, I., 2008. Bayesian modelling of football outcomes: using the Skellam's distribution for the goal difference, *IMA Journal of Management Mathematics* 20(2), 133-145. Retrieved from https://doi.org/10.1093/imaman/dpn026

Kovalchik, S., 2020. Extension of the Elo rating system to margin of victory, *International Journal of Forecasting* 36(4), 1329-1341. Retrieved from http://www.sciencedirect.com/science/article/pii/S0169207020300157

Langville, A.N., & Meyer, C.D. (2012). Who's #1, the science of rating and ranking. Princeton University Press.

Lasek, J., & Gagolewski, M. (2020). Interpretable sports team rating models based on the gradient descent algorithm, *International Journal of Forecasting* 37(3), 1061-1071. Retrieved from http://www.sciencedirect.com/science/article/pii/S0169207020301849

[16] The True Skill algorithm uses the Rao-Kupper (Rao & Kupper, 1967) model for the draws. On the other hand, the model used in the Glicko algorithm may be recognized as the Davidson model, but development is done only for $\kappa = 1$.

Lasek, J., Szlávik, Z., & Bhulai, S., 2013. The predictive power of ranking systems in association football, *International Journal of Applied Pattern Recognition* 1(1), 27-46. Retrieved from https://www.inderscienceonline.com/doi/abs/10.1504/IJAPR.2013.052339 (PMID: 52339)

Ley, C., Van de Wiele, T., & Van Eetvelde, H. 2019. Ranking soccer teams on the basis of their current strength: A comparison of maximum likelihood approaches, *Statistical Modelling* 19(1), 55-73. Retrieved from https://doi.org/10.1177/1471082X18817650

Maher, M.J., 1982. Modelling association football scores, *Statistica Neerlandica* 36(3), 109-118. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9574.1982.tb00782.x

Myers, J.L., & Well, A.D., 2003. *Research design and statistical analysis* (2nd ed.). Mahwah, New Jersey: Laurence Erlbaum Associates.

Rad, K.R., & Maleki, A., 2020. A scalable estimate of the out-of-sample prediction error via approximate leave-one-out cross-validation, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 82(4), 965-996. Retrieved from https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12374

Rao, P.V., & Kupper, L.L., 1967. Ties in paired-comparison experiments: A generalization 25 of the Bradley-Terry model, *Journal of the American Statistical Association* 62(317), 194-204. Retrieved from https://amstat.tandfonline.com/doi/abs/10.1080/01621459.1967.10482901

The Roon Ba., 2022. Retrieved May 13, 2022, from http://www.theroonba.com/

Silver, N., 2014. Introducing NFL Elo ratings. Retrieved July 1, 2020, from https://fivethirtyeight.com/features/introducing-nfl-elo-ratings/

Soccerway., 2022. Retrieved May 13, 2022, from https://www.soccerway.com

Szczecinski, L., 2022. G-Elo: generalization of the Elo algorithm by modeling the discretized margin of victory, *Journal of Quantitative Analysis in Sports*. Retrieved from https://doi.org/10.1515/jqas-2020-0115

Szczecinski, L., & Djebbi, A., 2020. Understanding draws in Elo rating algorithm, *Journal of Quantitative Analysis in Sports* 16(3), 211-220.

Szczecinski, L., & Tihon, R., 2021. Simplified Kalman filter for online rating: one-fits-all approach. Retrieved from http://arxiv.org/abs/2104.14012

Appendix A. Derivation of Davidson algorithm

To calculate $g(z;y_t)=\frac{\mathrm{d}}{\mathrm{d}z}\ell(z;y_t)$, assume that $\eta=0$ and then (21)–(23) imply $\ell(z_t;\mathsf{A})=\ell(-z_t;\mathsf{H})$ and $\ell(z_t;\mathsf{D})=-\log\kappa+0.5\ell(z_t;\mathsf{H})+0.5\ell(z_t;\mathsf{A})$ so that

$$\ell(z_t;y_t)=\ell(z_t;\mathsf{H})\mathbb{I}[y_t=\mathsf{H}]+\ell(z_t;\mathsf{A})\mathbb{I}[y_t=\mathsf{A}]+\ell(z_t;\mathsf{D})\mathbb{I}[y_t=\mathsf{D}] \tag{A.1}$$

$$=\ell(z_t;\mathsf{H})\check{y}_t+\ell(-z_t;\mathsf{H})(1-\check{y}_t)-\mathbb{I}[y_t=\mathsf{D}]\log\kappa, \tag{A.2}$$

where we use $\check{y}_t=\mathbb{I}[y_t=\mathsf{H}]+0.5\mathbb{I}[y_t=\mathsf{D}]$ and $1-\check{y}_t=\mathbb{I}[y_t=\mathsf{A}]+0.5\mathbb{I}[y_t=\mathsf{D}]$.

It is just enough to find the derivative of $\ell(z;\mathsf{H})$

$$g(z;\mathsf{H})=-\frac{\mathrm{d}}{\mathrm{d}z}\log\frac{10^{0.5z}}{10^{0.5z}+\kappa+10^{-0.5z}}=-\ln 10\frac{0.5\kappa+10^{-0.5z}}{10^{0.5z}+\kappa+10^{-0.5z}} \tag{A.3}$$

and plug it in (A.2) to obtain

$$g(z_t;y_t)=g(z_t;\mathsf{H})\check{y}_t-g(-z_t;\mathsf{H})(1-\check{y}_t)=-\ln 10(\check{y}_t-F_\kappa(z_t)), \tag{A.4}$$

where $F_\kappa(z_t)$ is the same as (26) (if $\eta=0$).

Now, we can go back to an arbitrary η which requires replacing z_t with $z_t+b_t\eta$ and then (A.4) is the same as (25).

Appendix B. Approximate leave-one-out cross-validation

Our goal is to calculate in a simple manner the terms $\boldsymbol{x}_t^\top\hat{\boldsymbol{\theta}}_{\setminus t}$ that appear in the scoring functions in (32) and in (33).

We start by approximating the objective function (35) using the Taylor series

$$J_{\setminus t}(\boldsymbol{\theta})=J(\boldsymbol{\theta})+\xi_{c_t}\log L(\boldsymbol{x}_t^\top\boldsymbol{\theta}/s;y_t) \tag{B.1}$$

$$\approx J(\hat{\boldsymbol{\theta}})+\xi_{c_t}\log L(\boldsymbol{x}_t^\top\hat{\boldsymbol{\theta}}/s;y_t)$$
$$-\frac{\xi_{c_t}}{s}g_t\boldsymbol{x}_t^\top(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})+\frac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^\top\left[\hat{\mathbf{H}}-\frac{\xi_{c_t}}{s^2}h_t\boldsymbol{x}_t\boldsymbol{x}_t^\top\right](\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}), \tag{B.2}$$

where $g_t\equiv g(\boldsymbol{x}_t^\top\hat{\boldsymbol{\theta}}/s;y_t)$ is defined in (25),

$$\hat{\boldsymbol{\theta}}=\arg\min_{\boldsymbol{\theta}}J(\boldsymbol{\theta}) \tag{B.3}$$

is the optimal solution for all data, the Hessian at the optimum is given by

$$\hat{\mathbf{H}}=\nabla^2_{\boldsymbol{\theta}}J(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}=\sum_{t\in\mathcal{T}}\frac{\xi_{c_t}}{s^2}h_t\boldsymbol{x}_t\boldsymbol{x}_t^\top+\frac{\alpha}{s^2}\boldsymbol{I}, \tag{B.4}$$

and we use the second derivative $h_t\equiv h(\boldsymbol{x}_t^\top\hat{\boldsymbol{\theta}}/s)$ where (Szczecinski & Tihon, 2021, [Sec. IV])

$$h(z)=\frac{\mathrm{d}}{\mathrm{d}z}g(z;y)=\frac{(\ln 10)^2}{4}\frac{\kappa 10^{0.5(z+\eta b)}+4+\kappa 10^{-0.5(z+\eta b)}}{(10^{0.5(z+\eta b)}+\kappa+10^{-0.5(z+\eta b)})^2}. \tag{B.5}$$

By equating the gradient of (B.2) to zero, we find the approximate solution to the optimization problem

$$\hat{\boldsymbol{\theta}}_{\setminus t}\approx\arg\min_{\boldsymbol{\theta}}J_{\setminus t}(\boldsymbol{\theta})$$
$$=\hat{\boldsymbol{\theta}}+\frac{\xi_{c_t}g_t}{s}\left[\hat{\mathbf{H}}-\frac{\xi_{c_t}}{s^2}h_t\boldsymbol{x}_t\boldsymbol{x}_t^\top\right]^{-1}\boldsymbol{x}_t \tag{B.6}$$

and the terms $\boldsymbol{x}_t^\top \hat{\boldsymbol{\theta}}_{\setminus t}$, $t \in \mathcal{T}$ which appear as arguments of the metrics (32) and (33) can now be efficiently calculated for all $t \in \mathcal{T}$ once $\hat{\boldsymbol{\theta}}$ is known (Rad & Maleki, 2020; Burn, 2020)

$$\boldsymbol{x}_t^\top \hat{\boldsymbol{\theta}}_{\setminus t} \approx \boldsymbol{x}_t^\top \hat{\boldsymbol{\theta}} + \frac{\xi_{c_t} g_t}{s} \boldsymbol{x}_t^\top \left[\hat{\mathbf{H}} - \frac{\xi_{c_t}}{s^2} h_t \boldsymbol{x}_t \boldsymbol{x}_t^\top \right]^{-1} \boldsymbol{x}_t \tag{B.7}$$

$$= \boldsymbol{x}_t^\top \hat{\boldsymbol{\theta}} + \frac{\xi_{c_t} g_t}{s} \boldsymbol{x}_t^\top \left[\hat{\mathbf{H}}^{-1} + \frac{\xi_{c_t} h_t}{s^2 - \xi_{c_t} h_t \boldsymbol{x}_t^\top \hat{\mathbf{H}}^{-1} \boldsymbol{x}_t} \hat{\mathbf{H}}^{-1} \boldsymbol{x}_t \boldsymbol{x}_t^\top \hat{\mathbf{H}}^{-1} \right] \boldsymbol{x}_t \tag{B.8}$$

$$= \boldsymbol{x}_t^\top \hat{\boldsymbol{\theta}} + \frac{\xi_{c_t} g_t a_t s}{s^2 - \xi_{c_t} h_t a_t}, \tag{B.9}$$

where $a_t = \boldsymbol{x}_t^\top \hat{\mathbf{H}}^{-1} \boldsymbol{x}_t$ and to pass from (B.7) to (B.8) we used the matrix inversion lemma (Barber, 2012, [Ch. A.1.8]).

The advantage of this formulation is clear: instead of solving T times the optimization problem (34), we only need to solve once the optimization defined in (B.3). Compared to the latter, the remaining operations of the inversion of the matrix $\hat{\mathbf{H}}$ and the multiplication required to calculate a_t, $t \in \mathcal{T}$, have a very small complexity.

The identical approach may be used to apply the ALO to the problem (42) but, we have to replace ξ_{c_t} in (B.9) with $\xi_{c_t} \zeta_{v_t}$.

In order to apply the ALO to the problem (50), we need a second derivative of (48) which is given by

$$h(z) = \frac{\mathrm{d}}{\mathrm{d}z} g(z; d) = \mathrm{e}^c (\mathrm{e}^{z+b\eta} + \mathrm{e}^{-z-b\eta}). \tag{B.10}$$