Article

Stochastic Extensions of the Elo Rating System

Gonzalo Gómez-Abejón * and J.Tinguaro Rodríguez

Posted Date: 25 July 2024

doi: 10.20944/preprints202407.1974.v1

Keywords: Rating Systems; Elo; Stochastic Processes; Sports Forecasting

 doi:10.20944/preprints202407.1974.v1

Article

Stochastic Extensions of the Elo Rating System

Gonzalo Gómez-Abejón [1,*] and J. Tinguaro Rodríguez [2]

[1] Department of Computational and Applied Mathematics and Operations Research (CMOR), Rice University, Houston, Texas

[2] Department of Statistics and Operations Research, Complutense University of Madrid, Madrid, Spain; jtrodrig@mat.ucm.es

* Correspondence: gg51@rice.edu

Abstract: This work studies how the Elo rating system can be applied to score-based sports, where it's gaining popularity, and in particular for predicting the result at any point of a game, extending its statistical basis to stochastic processes. We derive some new theoretical results for this model and use them to implement Elo ratings for basketball and soccer leagues, where the assumptions of our model are tested and found to be mostly accurate. We showcase several metrics for comparing the performance of different ratings systems and determine whether adding a feature has a statistically significant impact. Finally, we propose an Elo model based in a discrete process for the score that allows us to obtain draw probabilities for soccer matches, and has a performance competitive with alternatives like SPI ratings.

Keywords: rating systems; Elo; stochastic processes; sports forecasting

1. Introduction

Rating systems track and predict the performance of competitors in pairwise zero-sum games. They were initially developed to objectively measure the strength of chess players, with the first successful system proposed by Arpad Elo in 1960, and adopted by the United States Chess Federation (USCF) adopted it to replace the more problematic Harkness rating system. The International Chess Federation (FIDE) also started publishing ratings of its players in 1970, and continues to publish them today. The Elo rating system, or variants of it, are also used by most chess websites, where users are automatically matched against other players of similar strength, as well as by federations of go and scrabble players, and several videogames.

The motivation, mathematical basis and problems of the elo rating system are explained in detail in Elo's book [2], and the academic literature is summarized in [1]. The system gives each player a real parameter called their rating, and defines an algorithm to update it with the result of each new game. Winning a game increases the player's rating, and losing decreases it, but the increase depends on the rating of the rival. This allows for the ratings to change as the strength of the players also does. More sophisticated systems like *glicko*, proposed by Mark Glickman [3], include two parameters for each player, representing the strength and its variability, to allow for players gaining strength more quickly than others.

In his thesis, Glickman also extends Elo ratings to sports like American Football to make predictions about the score (not just the result), and more recently, the idea of applying elo ratings to team sports has gained popularity. For instance, fivethirtyeight.com has been keeping computing the ratings of NBA teams since 2014, and other websites like eloratings.net do the same for national soccer teams.

Our contributions are a more formal study of the model underlying the Elo system, with a few results regarding unbiased estimators for the ratings that are used in our computational analysis of league games. We derive how the natural extension of the Elo model for games involving a scoreboard can be used to predict the result during the game, and develop a systematic system for evaluating different systems.

Finally, we show how a discrete stochastic process can be used to model the score, and integrated into a novel type of Elo rating system. For comparison, we implement the rating system behind Nate Silver's soccer-specific SPI ratings, and show that our system has similar predicting power for the result of the games, despite only tracking one parameter for each team.

 doi:10.20944/preprints202407.1974.v1

The remainder of this paper is organized as follows: Section 2 recalls the main aspects of the (static) Elo rating system. The proposed stochastic extensions of the Elo system are then introduced in Section 3. Their performance at predicting the results of team sports, namely basketball and soccer, is analyzed through the computational study presented in Section 4. Finally, Section 5 is devoted to shed conclusions and expose possible lines of future work.

2. Preliminaries

Although the Elo system has been extensively studied, it's worth going through its mathematical and statistical basis in order to see how its assumptions can be extended to stochastic processes later. We also go through some basic accuracy metrics that can be used to asses its performance, and a statistical estimator to obtain ratings for players without a preexisting rating.

2.1. The Elo Rating System

Elo's rating system assumes that in a game between two players A and B, the result can be expressed giving points $p_A, p_B \in [0,1]$ to A and B respectively so that $p_B + p_A = 0$. The most simple case is a game with two results, in which p_A is 1 if A wins and 0 if B wins, but in chess, a draw is represented by $p_A = p_B = \frac{1}{2}$, since, in traditional chess tournaments, the player with the higher number of wins plus half the number of draws wins. Then, if A and B respectively have Elo ratings r_A and r_B, the updated ratings after the game are given by

$$r'_A := r_A + K(p_A - \mathbb{E}[p_A|r_A, r_B]) \qquad \text{and} \qquad r'_B := r_B + K(p_B - \mathbb{E}[p_B|r_B, r_A]) \tag{1}$$

where $\mathbb{E}[p_A|r_A, r_B]$ is the expected score of player A in a game against B, given by

$$E[p_A|r_A, r_B] := \mathbf{F}_X(r_A - r_B) \tag{2}$$

Here, K is a positive constant, and $\mathbf{F}_X$ is the distribution function of some random variable X with mean zero. The distribution function of X must also be symmetrical around zero so that $\mathbb{E}[p_A|r_A, r_B] + \mathbb{E}[p_B|r_A, r_B] = 1$. X was originally a normal variable, but it was later changed to follow a logistic distribution in the FIDE's ratings. This has been argued to produce better predictions [2, ch. 8.41], and it also leads to an explicit and more meaningful expression for $\mathbb{E}[p_A]$. In fact, sometimes the update rule is given simply as

$$r'_A := r_A + K\left(p_A - \frac{1}{1 + c^{(r_A - r_B)}}\right) = r_A + K\left(p_A - \frac{a^{r_B}}{a^{r_B} + a^{r_A}}\right) \tag{3}$$

for some $c = \log(a) \in \mathbb{R}^+$.

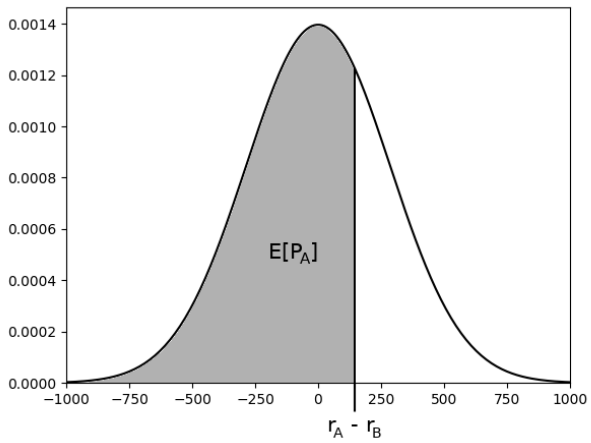


Figure 1. Expected score versus rating difference.

 doi:10.20944/preprints202407.1974.v1

It's also possible to update the ratings with the results of a set of matches (for instance the results of a tournament). Given a sample $M = \{m_k = (a_k, b_k, p_k^a, p_k^b)\}$ of matches with n intervening players $i = 1, \ldots, n$, where the k-th match is between players a_k and b_k and ends with scores p_k^a for a_k and $p_k^b = 1 - p_k^a$ for b_k, and given initial ratings $R := (r_1, r_2, \ldots, r_n)$, we similarly compute:

$$r_i' := r_i + K \cdot \sum_{k|a_k=i} (p_k^a - \mathbb{E}[p_k^a|R]) + K \cdot \sum_{k|b_k=i} \left(p_k^b - \mathbb{E}[p_k^b|R]\right) \tag{4}$$

Obviously, if a player outperforms their expected score, his rating increases, and otherwise it decreases, but the total sum of ratings doesn't change, as we will see later. However, before discussing the update rule, we need to review the model where the expected score originally comes from, which we will refer to as the static Elo model.

2.2. Basis of the Static Elo Model

The difficulty in rating chess players versus, for instance, olympic runners, is that in the latter sport there is a magnitude, the finish time, which doesn't depend on the rivals. For instance, taking a weighted mean of the latest times of each athlete would be enough to compare them, and objectively establish which are better and by how much.

However, Elo considers [2](ch. 8.23) what happens if we don't know the finish times, and can only compare athletes by looking at who finishes first in head to head races. If the finish times of runners A and B follow distributions $T_A \sim N(\mu_A, \sigma^2)$ and $T_B \sim N(\mu_B, \sigma^2)$, and are independent, then:

$$\mathbb{P}[A \text{ wins}] = \mathbb{P}[T_A < T_B] = \mathbb{P}[T_A - T_B < 0] = \Phi\left(\frac{\mu_B - \mu_A}{\sqrt{2\sigma^2}}\right) \tag{5}$$

In general, if the times follow the same continuous distribution Y with different means, i.e. $T_A \sim Y + \mu_A$ and $T_B \sim Y + \mu_B$, we can define $X = T_A - T_B + \mu_B - \mu_A$ and

$$\mathbb{P}[A \text{ wins}] = \mathbb{P}[T_A - T_B < 0] = \mathbf{F}_{T_A - T_B}(0) = \mathbf{F}_X(\mu_B - \mu_A) \tag{6}$$

which is exactly Eq. (2) for $r_i = -\mu_i$ (note that the lower the time, the better the athlete, and the higher the rating should be). Also, by definition X has mean zero and follows a symmetrical distribution. Therefore, Elo's system simply tries to estimate those underlying means μ_A, μ_B from several match results and a given $\mathbf{F}_X$.

It also follows that $\mathbf{F}_X(\mu_A - \mu_B) = \mathbf{F}_X((\mu_A + h) - (\mu_B + h))$, so the μ_i can only be determined up to addition of a constant (we can only estimate their pairwise differences). Finally, $\mathbf{F}_{\lambda X}(\lambda\mu_A - \lambda\mu_B) = \mathbf{F}_X(\mu_A - \mu_B)$, so scaling X doesn't affect the model either (only scales the ratings). This means that both the average of the ratings and the variance of X are an arbitrary choice, and only the shape of $\mathbf{F}_X$ affects the behavior of the model.

Traditionally, 1500 is chosen as the average rating [1](p. 621) although Elo's initial choice was 2000 as the midpoint and $\sigma_X = 200\sqrt{2} \approx 282.84$ [2](ch. 8.21). This was replaced by the logistic distribution $\mathbf{F}_X(x) = \left(1 + 10^{x/400}\right)^{-1}$, which has a similar standard deviation of $\frac{400\pi}{log(10)\sqrt{3}} \approx 315.09$.

2.3. Statistical Inference in the Static Model

For the model presented above, the next logical step is to estimate the ratings of the players given a sample of matches (for instance a tournament). Again, we denote the sample by $M = \{m_1, \ldots, m_{|M|}\}$ with $m_k = (a_k, b_k, p_k^a, p_k^b)$, and if player i played match k, we will use as shorthand $p_{i,k} = p_i^a$ if $i = a_k$ and $p_{i,k} = p_i^b$ if $i = b_k$. We also denote the indices of games played by j with $M^j := \{1 \leq k \leq |M| : j =$

 doi:10.20944/preprints202407.1974.v1

a_k or $j = b_k\}$. Then, an interesting estimator $\hat{R} = (\hat{r}_1, \dots \hat{r}_n)$ is the one that reduces to zero the errors ε_j defined by

$$\varepsilon_j(M, \hat{R}) := \sum_{k|a_k=j} (p_k^a - \mathbb{E}[p_k^a|\hat{R}]) + \sum_{k|b_k=j} (p_k^b - \mathbb{E}[p_k^b|\hat{R}]) =: \sum_{k \in M^j} (p_{j,k} - \mathbb{E}[p_{j,k}|\hat{R}])$$

As we will see later, this estimator is related with the adjustment formula in Eq.(4), and in particular, if it exists, it's a fixed point when adjusting with that same sample M:

$$\hat{r}_j = \hat{r}'_j = \hat{r}_j + K \sum_{k \in M^j} (p_{j,k} - E[p_{j,k}|\hat{R}]) \iff 0 = \sum_{k \in M^j} (p_{j,k} - E[p_{j,k}|\hat{R}]) = \varepsilon_j$$

In fact, given a sample with n players, there are $n-1$ errors to minimize (since they add up to zero) and $n-1$ ratings to adjust (since the sum doesn't matter), so we should expect the estimator with zero error to be unique, and thus the only fixed point of the adjustment formula. In Appendix A, we characterize it's existence and prove uniqueness in the case that it exists. Convergence is easy to see for the case of two players i and j, because if the results of a game follow Elo's model, i.e. Eq.2, and $\hat{r}_j^m, \hat{r}_i^m$ are the estimators produced by the sample of the first m games, then by the law of large numbers

$$\varepsilon_j = 0 \Rightarrow \mathbf{F}_X(\hat{r}_j^m - \hat{r}_i^m) = \frac{1}{m}\sum_{k=1}^{m} \mathbf{F}_X(\hat{r}_j^m - \hat{r}_i^m) = \frac{1}{m}\sum_{k=1}^{m} p_{j,k} \xrightarrow[m\to\infty]{a.s.} E[p_{j,k}] = \mathbf{F}_X(r_j - r_i)$$

and since $\mathbf{F}_X$ is continuous and has an inverse, $\hat{r}_j^m - \hat{r}_i^m$ converges to $r_j - r_i$. Note that if the average estimated rating is the same at every step m, then $\hat{r}_j^m \xrightarrow[m\to\infty]{a.s.} r_j + C$ for every j. Therefore, when we look for numerical methods to compute it, it makes sense to keep the average $\frac{1}{m}\sum_{j=1}^{n} \hat{r}_j$ constant, and only worry about convergence of the pairwise differences.

Finally, there's another natural way to estimate the ratings for a sample, if X is a logistic variable and there are only two results in the game ($p_A \in \{0, 1\}$). In that case, the sample can be encoded as a set of independent variables $X^1, ..., X^n$ where $X_k^j = 1$ if $j = a_k$, $X_k^j = -1$ if $j = b_k$, and otherwise (if $k \notin M^j$) $X_k^j = 0$. The static model then gives

$$\mathbb{P}(p_k^a = 1) = \mathbb{E}[p_k^a] = \mathbf{F}_X(r_{a_k} - r_{b_k}) = \mathbf{F}_X\left(\sum_{j=1}^{n} X_k^j r_j\right) = \frac{1}{1 + e^{\lambda \sum_{j=1}^{n} r_j X_k^j}} \tag{7}$$

which is precisely the formula of the logistic regression of $Y := (p_k^a : 1 \le k \le m)$ against X. Since λ only depends on the variance of the distribution, and the choice of variance doesn't affect the model, we can just estimate $r_1, \dots r_n$ by doing a logistic regression in the X^j.

2.4. Basis of the Elo Adjustment Formula

So far we have worked assuming Eq.(2) holds and each player has a constant "true" rating r_j. However, in practice the strength of chess players or soccer teams changes over time, and hence the need for a system that updates the ratings with each new game and not with the whole history.

This is the reasoning behind the update formula in Eq.(1): if we consider only one match between A and B, with estimated ratings $\hat{r}_A, \hat{r}_B$ and real ratings r_A, r_B (and for simplicity we pick $\hat{r}_A + \hat{r}_B = r_A + r_B = 0$), then using Taylor's expansion, $\exists \xi \in \mathbb{R}$ such that

$$\mathbb{E}[\hat{r}'_A - \hat{r}_A] = K \cdot \mathbb{E}[p_A - \mathbf{F}_X(\hat{r}_A - \hat{r}_B)] = K \cdot \mathbf{F}_X(2r_A) - K \cdot \mathbf{F}_X(2\hat{r}_A) = 2K(r_A - \hat{r}_A) f_X(\xi)$$

Therefore, for small enough K, in particular for $K < (2\max_{x\in\mathbb{R}} f_X(x))^{-1}$, this gives

$$|\mathbb{E}[\hat{r}'_A - r_A]| = |\mathbb{E}[\hat{r}'_A - \hat{r}_A] + \hat{r}_A - r_A| = |1 - 2Kf(\xi)||\hat{r}_A - r_A| < |\hat{r}_A - r_A|$$

 doi:10.20944/preprints202407.1974.v1

and thus if r_A is fixed, $E[\hat{r}_A]$ converges to r_A as $\hat{r}_A$ is updated over an increasing number of games. In practice, K is chosen much smaller than $\frac{1}{2\max f(\mathbb{R})}$ - for instance, FIDE uses $K = 10$ for players with ratings above 2400, and in this case $\mathbf{F}_X(x) = \frac{1}{1+10^{-x/400}}$ so a better K for accelerating the convergence of the expectancy would be $\frac{1}{2f_X(0)} = \frac{4 \cdot 400}{2\log(10)} \approx 347$.

The reason for the difference is that faster convergence is at the expense of a much higher variance of $\hat{r}'_A$ (the variance is proportional to K^2). If we apply the update formula over an infinite number of games, even if the true ratings of the model stay constant, the computed ratings don't converge, and they approach a limiting distribution [5], the variance of which we would like to minimize.

Therefore, the choice of K depends on the dynamic properties of the strength of the players. If we expect this strength to change significantly from one game to the next, K should be bigger, and if we expect the skill of the players to be relatively stable (for instance, if many games are played in a small lapse of time), we choose a lower K. The FIDE uses $K = 40$ for younger players, who may improve quickly, and $K = 10$ as said above for players with rating above 2400, which are usually grandmasters and don't improve their play that fast.

2.5. Asymmetric Games

Until now we have assumed that for two players of equal strength each has expected score $1/2$, but in practice this is not true: in chess, for instance, the player with white pieces makes the first move, and has a slight advantage. In the team sports we will consider later, it is well known that if a team plays in the home field, it also has an advantage over its rival. To incorporate this in the Elo rating system, when A has a systemic advantage against B, we can use the following expectancy formula [3][p. 36]:

$$\mathbb{E}[p_A] = \mathbf{F}_X(r_A - r_B + L) = 1 - \mathbf{F}_X(r_B - r_A - L) = 1 - \mathbb{E}[p_B] \tag{8}$$

Here, L is a positive constant, which should obviously be bigger for a bigger advantage of A, since $\mathbb{E}[p_A]$ is increasing in L, and $L = 0$ reduces to our symmetric model. We can estimate it given a sample of matches, provided of course that we know which player has the systematic advantage in each match. From now on, we suppose that in the k-th match $m_k = (a_k, b_k, p_k^a, p_k^b)$ of a sample M, a_k has the advantage. In particular, if X is logistic, L can be estimated as the independent term in the regression formula in Eq.(7).

2.6. Accuracy Metrics for Elo Rating Systems

In order to determine the predictive accuracy of a rating system, we will use several tools. The first one, provided by Arpad Elo [2, ch. 2.6], is a statistical test of normality for a sample in which the players have preexisting ratings R and each player plays m games. Originally the sample was a chess tournament, or a set of tournaments with the same m [2, ch. 2.7].

Elo proposes to represent the values of the residues $\varepsilon_j(R)$ in a histogram and to compare them with the frequencies of a normal variable with mean 0 and the sample variance. Since $\varepsilon_j(R) = \sum_{k \in M^j} \mathbb{E}[p_{j,k}|R]$ is the sum of m variables, if m is high enough and the rating differences aren't too high, by the CLT, each ε_j should approximately follow a normal distribution, with variance $\sum_{k \in M^j}(p_{j,k}|R) \leq 1/2\sqrt{m}$. More sophisticated normality tests, like the normal probability plot, can be carried out for the same variables.

In order to compare the accuracy of different systems in the same sample, we propose looking at the average of the squared residues across all matches, i.e. the mean squared error of p:

$$MSE := \frac{1}{|M|} \sum_{k=1}^{|M|} \left(p_{a_k,k} - E[p_{a_k,k}|R]\right)^2 \tag{9}$$

 doi:10.20944/preprints202407.1974.v1

We can decompose this sum using the known expression of the mean square error, which is the square of the bias plus the mean variance:

$$E[(X-a)^2] = (E[X]-a)^2 + Var(X) \tag{10}$$

In particular, when applied to the results of the games, we obtain:

$$MSE = \frac{1}{|M|}\sum_{k=1}^{|M|}\left(E[p_{a_k,k}] - E[p_{a_k,k}|R]\right)^2 + \frac{1}{|M|}\sum_{k=1}^{|M|} Var(p_{a_k,k}) \tag{11}$$

The advantage of this measure is that we don't require any of the assumptions of a normality test. The sample can be extended in time (we can update the ratings between games using Eq.(1)) and we don't need every player to play the same number of games. Also, if we apply two different rating systems to the same number of games, we can look at the variance reduction and extrapolate p-values to determine if one system is better than other.

However, unlike in a linear regression model, the results p_k^a don't have a constant variance, so the expression $MRV := \frac{1}{|M|}\sum_{k=1}^{|M|} Var(p_{a_k,k})$ (mean result variance) can only be understood as the expected variance of a game, i.e. $\mathbb{E}[Var(p_k^a|r_{a_k} - r_{b_k})]$, dependent on some distribution of the rating differences $r_{a_k} - r_{b_k}$. We can only compare results from different samples if we assume this average variance to be similar, but this isn't the case in general (for instance, MRV decreases as the dispersion of the ratings increases).

If we assume the results are binary ($p_k^a \in \{0,1\}$) we can use other approaches. For instance, if $\mathbf{F}_X$ is logistic, as we saw in Eq.(7), the model is equivalent to that of a logistic regression, so we can derive the statistics of this type of regression. For instance, the deviance D, which is distributed as a chi-squared variable,

$$D = -2 \cdot \ln\left(\frac{\text{likelyhood}_{Elo}}{\text{likelyhood}_{saturated}}\right)$$

Finally, the effectiveness of the model can be gauged more visually by plotting the receiver operating characteristic (ROC) curve. Again, this doesn't work with other possible results besides win or loss, such as a draw ($p_k^a = \frac{1}{2}$), unless the non-decisive results are removed or imputed (for instance, to 0 and 1 with probabilities $1 - p_k^a$ and p_k^a).

3. Stochastic Elo Models

In many competitive team sports, we don't just observe the result of the game, but also a stochastic process S_t, which determines the result at time T (we will assume T is positive constant, although it could be any stopping time of S_t). In practice, S_t will represent the score difference at time t, positive when the home team is winning, so the home team wins when $S_T > 0$, while the visiting team wins when $S_T < 0$.

Our goal is to extend the Elo model to incorporate S_t, so that the expected result of the game at time $t = 0$ matches the static model. In other words, we want to obtain an expression for $\mathbb{E}[p_A|r_A, r_B, S_t] = g(r_A, r_B, S_t, t)$ such that $g(r_A, r_B, 0, 0) = F_X(r_A - r_B + L) = \mathbb{E}[p_A|r_A, r_B]$, i.e. our initial guess matches that of the original Elo model.

We will study several possible models, but there are some properties they should all verify. First, since (in the sports we will see) the past states of the scoreboard aren't relevant to the game, and only the final score determines the result, S_t should have the Markov property. As a consequence, both $X_t := \mathbb{E}[p_A|S_t] = g(r_A, r_B, S_t, t)$ and $Y_t := \mathbb{E}[S_T|S_t]$ should be martingales, by the Tower Property of Conditional Expectation [4](p. 46).

 doi:10.20944/preprints202407.1974.v1

Furthermore, for $s > t$, since S_s gives us as much information as S_s and S_t together, our estimation at time s should be better than the one at time t, and therefore

$$\mathbb{E}[Var(p_A|S_s)] < \mathbb{E}[Var(p_A|S_t)] \qquad \text{and} \qquad \mathbb{E}[Var(S_T|S_s)] < \mathbb{E}[Var(S_T|S_t)]$$

3.1. Continuous Models

Recall that in the original Elo model, team A wins if $X + r_A - r_B + L > 0$, and that should match the odds of S_T being positive, so we can assume S_T is $X + r_A - r_B + L$ times some constant. Since the incentives of the teams are the same at every point of the game (score the most points and have the opponent score the least), we can also assume the increments $S_{t+h} - S_t$ are independent (for disjoint intervals) and identically distributed.

For this to make sense, $S_T = S_{T/n} + (S_{2T/n} - S_{T/n}) + \cdots + (S_T - S_{T-T/n})$ must be infinitely divisible. Although the logistic distribution is infinitely divisible [6], the increments aren't logistic, so we will suppose that X and S_T are normal, and in that case the increments must also follow the normal distribution by Cramér's decomposition theorem [7], so S_t has i.i.d. Gaussian increments and it's a Brownian motion with drift:

$$S_t = \mu t + \sigma B_t \quad \Rightarrow \quad \mathbb{E}[p_A] = \mathbb{P}[N(\mu T, \sigma^2 T) > 0] = \Phi\left(\frac{\mu T}{\sigma\sqrt{T}}\right) = \Phi\left(\frac{\mu}{\sigma}\sqrt{T}\right)$$

where B_t is the standard Brownian motion, i.e. $B_0 = 0$, and B has independent and normal increments $B_t - B_s \sim N(0, |t-s|)$. If we want this expectation to match the one given by the static model, we must have

$$\Phi\left(\frac{\mu}{\sigma}\sqrt{T}\right) = \mathbf{F}_X(r_A - r_B + L) = \Phi\left(\frac{r_A - r_B + L}{T}\right) \quad \Rightarrow \quad \frac{\mu}{\sigma}\sqrt{T} = \frac{r_A - r_B + L}{\sigma_X} \tag{12}$$

And from this we immediately obtain the expected score conditioned on $S_t = 0$:

$$\mathbb{E}[p_A|S_t = 0] = \mathbb{P}\Big[N(\mu(T-t), \sigma^2(T-t)) > 0\Big] = \Phi\left(\frac{\mu}{\sigma}\sqrt{T-t}\right) =$$

$$= \Phi\left(\frac{r_A - r_B + L}{\sigma_X}\frac{\sqrt{T-t}}{\sqrt{T}}\right) = \mathbf{F}_X\left((r_A - r_B + L)\frac{\sqrt{T-t}}{\sqrt{T}}\right) \tag{13}$$

Similarly, if we condition on an arbitrary score difference at time $t < T$, we obtain

$$\mathbb{E}[p_A|S_t = S] = \mathbb{P}\Big[S + N(\mu(T-t), \sigma^2(T-t)) > 0\Big] = \mathbb{P}\Big[N(\mu(T-t) + S, \sigma^2(T-t)) > 0\Big]$$

$$= \Phi\left(\frac{\mu}{\sigma}\sqrt{T-t} + \frac{S}{\sigma}\sqrt{T-t}^{-1}\right) = \mathbf{F}_X\left((r_A - r_B + L)\frac{\sqrt{T-t}}{\sqrt{T}} + \frac{\sigma_X}{\sigma}\frac{S}{\sqrt{T-t}}\right)$$

Denoting by $u_t := (T-t)/T$ the fraction of the time that remains, and by C the constant coefficient $\frac{\sigma_X}{\sigma\sqrt{T}}$ we have:

$$\mathbb{E}[p_A|S_t = S] = \mathbf{F}_X\left((r_A - r_B + L)\sqrt{u_t} + C\frac{S}{\sqrt{u_t}}\right) \tag{14}$$

In the end, we obtain an expression where the term $(r_A - r_B + L)\sqrt{u_t}$ depends on the ratings, and decreases to zero as the time t approaches T, while the term $C\sqrt{u_t}^{-1}S$ depends on the score and becomes larger as the game nears its end, as we would want (unless $S = 0$). In other words, the behavior or $\mathbb{E}[p_A|S_t]$ as $t \to T$ is exactly what we would expect.

Note that setting $t = 0 \Rightarrow \sqrt{u_t} = 1$, and assuming a game starts with a score difference of S points, the expected score is $\mathbf{F}_X(r_A - r_B + L + CS)$. This allows us to interpret the constant C as the handicap (measured in rating points) that each goal or point entails at the start of the game. We could

 doi:10.20944/preprints202407.1974.v1

in fact use this to estimate C from game data, although we don't need to pick $t \approx 0$. For any $t \in (0,t)$, C should minimize the mean square error

$$MSE(t) := \frac{\sum_{k\in M}(p_{a_k,k} - \mathbb{E}[p_{a_k,k}|S_{t,k}])^2}{|M|} = \frac{1}{|M|}\sum_{k\in M}\left(p_{a_k,k} - \mathbf{F}_X\left(\sqrt{u_t}(r_{a_k} - r_{b_k} + L) + C\frac{S_{t,k}}{\sqrt{u_t}}\right)\right)^2$$

as long as the model is correct, and the minimum should be the same for every t. Conversely, if the function above is minimized for $C = C^*$ for every t, our model is optimal among a certain class, as the following result showcases.

Lemma 1. *If there are functions $f : \mathbb{R} \to \mathbb{R}\setminus\{0\}$ and $V : \mathbb{R}^2 \to \mathbb{R}$ and there exists a $C^* \in \mathbb{R}$ such that $V(x, C^*f(x)) \leq V(x, Cf(x))$ for every $x, C \in \mathbb{R}$, then every $g : \mathbb{R} \to \mathbb{R}$ verifies*

$$V(x, C^*f(x)) \leq V(x, g(x)) \qquad \forall x \in \mathbb{R}$$

Proof. $V(x,g(x)) = V\left(x, \frac{g(x)}{f(x)}f(x)\right) \geq V(x, C^*f(x))$ □

In particular, for $V(t,g) = \frac{1}{|M|}\sum_{k\in M}(p_{a_k,k} - \mathbf{F}_X(\sqrt{u_t}(r_A - r_B + L) + gS_{t,k})^2$ and the function $f : t \mapsto 1/\sqrt{u_t}$, this means that our model is optimal among the ones with a prediction of the form $\mathbb{E}[p_A] = \mathbf{F}_X(\sqrt{u_t}(r_A - r_B + L) + Sg(t))$. If we further suppose that the true expectancy has the form $\mathbf{F}_X(f(t)(r_A - r_B + L) + Sg(t))$, we can estimate f by looking at the times in the sample where $S_{t,k} = 0$, and if $f : t \mapsto \sqrt{u_t}$ minimizes the mean square error over that restricted sample, our model is optimal among this wider class.

Finally, assuming we already know or have estimated $C = \frac{\sigma_X}{\sigma\sqrt{T}} \Rightarrow \sigma = \frac{\sigma_X}{C\sqrt{T}}$, we can obtain the drift $\mu = \frac{\sigma}{\sqrt{T}}\frac{r_A - r_B + L}{\sigma_X}$ and express S_t in terms of known quantities as

$$S_t = \mu t + \sigma B_t = \frac{r_A - r_B + L}{C\cdot T}t + \frac{\sigma_X}{C\sqrt{T}}B_t \tag{15}$$

Finally, notice also that $\mathbb{E}[S_T] = \frac{r_A - r_B + L}{C}$, which gives another way to estimate C.

3.2. Accuracy Metrics for the Stochastic Model

If we take $t = T$ in Eq. (15) we obtain

$$C\cdot S_T \sim r_A - r_B + L + \frac{\sigma_X}{\sqrt{T}}N(0,T) \Leftrightarrow C\cdot S_T - (r_A - r_B + L) \sim N(0,\sigma_X^2) \sim X \tag{16}$$

Therefore, the residuals $C\cdot S_T - (r_A - r_B + L)$ follow the same law as X (the variable used for the static model), and we can do a normality test on them, for instance through a QQ plot or comparing a histogram with the predicted frequencies for X. We could in principle use Eq.(14) with any distribution of X, and in that case $C\cdot S_T$ minus the rating difference follows the law of X and the same test works.

From Eq.(15) we can also reconstruct the standard Brownian motion S_t in the model:

$$\frac{1}{\sqrt{T}}B_t = \frac{1}{\sigma_X}\left(C\cdot S_t - \frac{t}{T}\cdot(r_A - r_B + L)\right) \tag{17}$$

and since $\frac{1}{\sqrt{a}}B_{at}$ is a standard Brownian motion if and only if B_t is, taking $s_t = 1 - u_t = t/T$ we get that $\frac{1}{\sqrt{R}}B_t = \frac{1}{\sqrt{T}}B_{Ts_t} = C\cdot S_{Ts_t} - s_t(r_A - r_B + L)$ is a standard Brownian motion in $[0,1]$ as a function of s_t, which is just the fraction of the game time elapsed at t.

From Eq.(15) we can also deduce

$$Var(S_T|S_t) = Var(S_T - S_t) = Var(\mu(T-t) + \sigma B_T - B_t) = \sigma^2(T-t) = C^{-2}\sigma_X^2\cdot u_t \tag{18}$$

 doi:10.20944/preprints202407.1974.v1

That is, if we compute the mean squared error of S_T instead of p_A at each time t, we should obtain a linear function in t, and we can test this hypothesis via linear regression.

Recall that we are assuming the increments of S_t are independent and identically distributed, and in particular the variance of $S_{t+h} - S_t$ only depends on t. We can test this directly by looking at the points scored by either team at each interval between 0 and T, or, if S_t can increase or decrease by amounts other than one, we can see if the sample variance $\frac{1}{|M|-1}\sum_{k\in M}(S_{t+1,k} - S_{t,k})^2 \xrightarrow[|M|\to\infty]{a.s.}$ $\mathbb{E}[(S_{t+1} - S_t)^2]$ depends on t.

Finally, this continuous model implies expressions for $\mathbb{E}[p_A|S_t]$ and $\mathbb{E}[S_T|S_t]$ for each t, and we can check if they behave like martingales in a sample of games.

3.3. Non-Homogeneous Process

In practice , S_t doesn't always behave like an homogeneous process in time, that is, more points may be scored in some sub-intervals of $[0, T]$ than others. In that case, if $Var(S_t) = m(t)$ for some increasing function m, we can model the process as

$$S_t = \int_0^t \mu dm(t) + \int_0^T m'(s)dB_s = \mu \cdot m(t) + \sigma B_{m(t)} = S'_{m(t)}$$

where S' behaves like the score in the first model. Since this is a map of the process that we were using before, we can use the homogeneous model to compute

$$\mathbb{E}[p_A|S_t = 0] = \mathbb{E}[p_A|S'_{m(t)} = S] = \mathbb{P}[S'_{m(T)} > 0|S'_{m(t)} = S] =$$

$$= \mathbf{F}_X\left(\sqrt{\frac{m(T) - m(t)}{m(T)}}(r_A - r_B + L) + C\sqrt{\frac{m(T)}{m(T) - m(t)}}S\right)$$

which is exactly the same as before, for $u_t = \frac{m(T)-m(t)}{m(T)}$.

3.4. A Discrete Model

The main difference between this model and the actual score difference is that the former is a continuous process, but the latter is discrete in value ($S_t \in \mathbb{Z}$) and continuous in time. The process that fulfills this conditions best is a Skellam process, as described in [11], i.e. the difference of two Poisson processes with different rates, which model the score of each team or player during the game.

If P_1 and P_2 are two independent Poisson distributions with mean μ_1 and μ_2, then their difference is said to follow a Skellam distribution, $P_1 - P_2 \sim Sk(\mu_1 - \mu_2)$. If $N_1(t)$ and $N_2(t)$ are independent Poisson processes with rates λ_1 and λ_2, the process

$$Z_t := N_1(t) - N_2(t) \sim P(t\lambda_1) - P(t\lambda_2) = Sk(t\lambda_1, t\lambda_2)$$

has i.d.d. increments as well.

However, in this case, modeling the score with Z_t is not so straightforward. For starters, if we multiply Z_t by a constant, we don't get another Skellam process. Adding a real-valued function of t also changes the domain of the process from $\mathbb{Z}$ to $\mathbb{R}$. Therefore, the only option is to assume that

$$S_t \sim Z_t(\mu_1(r_A, r_B), \mu_1(r_A, r_B)) \Rightarrow S_T \sim Sk(T\mu_1(r_A, r_B), T\mu_2(r_A, r_B))$$

This in turn means that we can't assume homocedasticity of the process, because a Skellam distribution with parameters μ_1 and μ_2 has mean $\mu_1 - \mu_2$ and variance $\mu_1 + \mu_2$. Since μ_1 and μ_2 are non-negative, the mean is less or equal to the variance, and if we fix $\mu_1 + \mu_2 = \sigma$, we would put a bound on the expected final score $\mathbb{E}[S_T] = \mu_1 - \mu_2 \leq \sigma$, which is not realistic.

 doi:10.20944/preprints202407.1974.v1

However, the distribution of S_T, which should have the same shape as X, now depends on two parameters, and not just on the rating difference $r_A - r_B$. We can remove this degree of freedom fixing a relation between μ_1 and μ_2, and for simplicity's sake, since the probability that $Sk(\mu_1, \mu_2)$ takes the value k is

$$\mathbb{P}[Sk(\mu_1, \mu_2) = k] = e^{-(\mu_1+\mu_2)} \left(\frac{\mu_1}{\mu_2}\right)^{k/2} I_k(2\sqrt{\mu_1\mu_2})$$

we can fix $2\sqrt{\mu_1\mu_2} = H$ to compute I_k only once for each $k \in \mathbb{Z}$. Here $I_k(\cdot)$ is the modified Bessel function of the first kind [9], which doesn't have a closed form expression.

To determine how μ_1 and μ_2 depend on the ratings, we can take $\mathbb{E}[S_T] = \mu_1 - \mu_2$ to be linear in the rating gap $\Delta_r = r_A - r_B + L$ as in Eq.(15), and then if we pick σ_X so that $C = 1$, from the product and difference of μ_1 and μ_2 we obtain a second degree equation

$$(x + \mu_1)(x - \mu_2) = x^2 + \Delta_r x - \frac{H^2}{4}$$

$$\Rightarrow \quad \mu_1 = \frac{1}{2}\left(\Delta_r + \sqrt{\Delta_r^2 + H^2}\right) \qquad \text{and} \qquad \mu_2 = \frac{1}{2}\left(-\Delta_r + \sqrt{\Delta_r^2 + H^2}\right)$$

from which we finally obtain not just a prediction for p_A, but also a probability for each possible final score, and in this case the probability that $S_T = 0$ is positive, so we can include draws in our model. In particular, if we assign a result of $p_A = p_B = 1/2$ for draws,

$$E[p_A] = \mathbb{P}[S_T < 0] + \frac{1}{2}\mathbb{P}[S_T = 0] = \mathbb{P}[Sk(\mu_1, \mu_2) < 0] + \frac{1}{2}\mathbb{P}[Sk(\mu_1, \mu_2) = 0] \tag{19}$$

Similarly, if we consider that expression as a function of Δ_r, it can be shown that it's increasing (see Appendix B), and with limits 0 and 1 at $-\infty$ and ∞. Therefore, $\mathbb{E}[p_A] = \mathbf{F}_X(r_A - r_B + L)$ for some random variable X, and then this model is also an extension of a static Elo model.

4. Computational Study

Next, we evaluate the performance of the proposed stochastic extensions of the Elo system in a computational study implemented in Python. All the results discussed can be replicated with the code available at github using real data from reference databases for each sport, which are also included for download through the previous link. In the case of soccer, a comparison with the more complex methodology of the Soccer Power Index (SPI) developed by Nate Silver [13] is also carried out.

4.1. Experimental Setup

In order to evaluate a rating model with any of the accuracy metrics exposed in Section 2.6, a sample of matches M for which both players have a rating is needed. Furthermore, the update formula in Eq. (1) needs preexisting ratings, so we will have to use some of the games in our dataset to estimate the starting ratings of each team, using the estimator $\hat{R}$ defined in Section 2.3.

On the other hand, the proposed stochastic models are designed for sports or games where the result depends on a numerical score (S) and the match has a fixed duration (T), such as soccer, basketball or ice hockey, although we will omit the latter in the experimental results for brevity. Moreover, we will focus on league competitions, where each team plays more games and we can use a bigger sample to obtain $\hat{R}$. For instance, in the Premier League, 20 teams play 380 matches, while in the World Cup, 32 teams play 64 matches.

Most national leagues have a similar format, consisting in yearly seasons in which most of the participating teams remain the same from one year to the next, and all teams play the same number of games each week. We can design an algorithm to extract rated games from any of these sports as follows:

- The expected result is determined by Eq. (8) with constant L.

 doi:10.20944/preprints202407.1974.v1

- During each season, we denote by "entering teams" to the teams that didn't play the previous season (during the first season, they are all entering teams).
- We divide each season in two parts (I and II), the former comprised of the games that start before every entering team has played at least m games.
- During part I, in each match between two non-entering teams, we update their ratings (from the previous season) according to Eq. (1) using a fixed factor K.
- When part I ends, we compute the rating estimator $\hat{R}$ for each entering team using the games in part I and the current ratings of non-entering teams.
- During part II, since we have ratings for all teams, we can just update them every match using Eq. (1) with the same K-factor.

In this way, we have a rating before the start of the game for the matches between non-entering teams in part I of each season, and all matches in part II, and we can use them to evaluate the rating system.

This algorithm has (hyper-) parameters m, K and L, that is, respectively the length of part I of each season, the sensitivity of the Elo system to new results, and the home field advantage. We will fix m and estimate K and L by minimizing the mean square error defined in Eq. (9).

4.2. Basketball Results

Basketball has several advantages regarding the Elo system. First, it only admits two results (win or loss), so the result follows a Bernoulli distribution and the logistical model in Eq. (7) can be used. In addition, the score variable S_t has a relatively large range (each team scores around 100 points) and changes quickly (by 1, 2 or 3 points each time), so we don't lose too much by approximating it by a continuous variable.

On the other hand, an NBA basketball match lasts 48 minutes divided in 4 12-minute quarters, but if the scoreboard is even at the end of that time, additional 5-minute quarters are played until one team is ahead. We will ignore these extra quarters for the sake of simplicity, since we assume T is fixed, but our prediction at the end of the game will be slightly wrong when $S_t = 0$.

Our dataset for basketball will consist of the NBA league games between the seasons 2000-01 and 2023-24, obtained from www.basketball-reference.com, which registers every change in the score and its time (in seconds).

4.2.1. Static Elo

We implement the algorithm described above for $m = 20$ and F_X corresponding to a normal variable $X \sim N(0, 200)$, in a training set of seasons 2000-01 to 2009-10, and in four different ways to test the effectiveness of the Elo model and the significance of its parameters:

- Without Elo, simply assuming $E[p_{k,a_k}] = \overline{p_{home}} := |M|^{-1} \sum_{k=1}^{|M|} p_{k,a_k} =: F_X(L)$.
- Fixing $K = 0$ and minimizing MSE as a function of L (no change in strength).
- Fixing $L = 0$ and minimizing MSE as a function of K (no home advantage).
- Minimizing MSE as a function of K and L (standard Elo system).

To avoid overfitting in K and L, we also implement the algorithm for the optimal values K^* and L^* in a validation set consisting of seasons 2010-11 to 2023-24. The mean squared errors in the training and testing sets are summarized in Table 1:

Table 1. Mean squared error of different Elo models vs. sample variance.

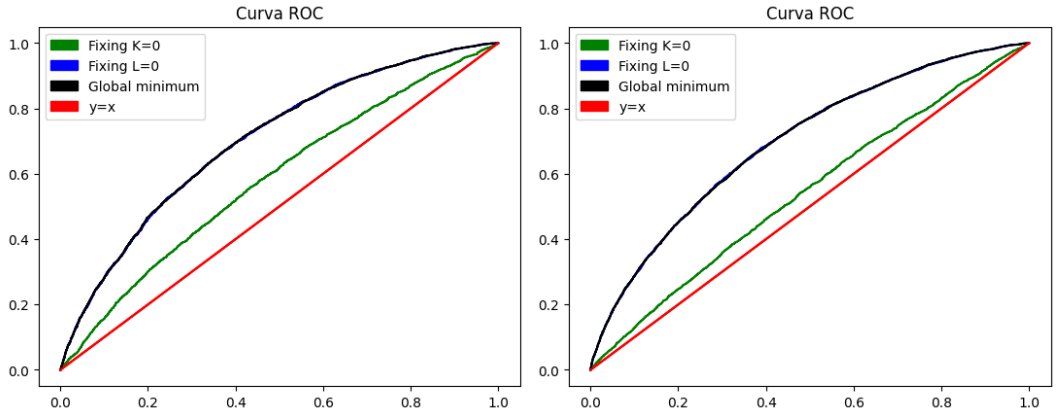
K^*	L^*	MSE	MSE_{test}
No Elo	-53.78	0.23876	0.24394
0	-60.89	0.26498	0.29590
14.71	0	0.22333	0.22308
16.19	-60.74	0.21179	0.21744

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 July 2024 doi:10.20944/preprints202407.1974.v1

We can see straight away that the Elo explains a significant part of the variance (around 11%), and both parameters are useful, since setting them to zero increases the mean squared error. The reduction is similar in the training and testing sets, so cross-validation shows the robustness of the algorithm in Section 4.1.

Notice also that the K-factor and home field advantage are similar to those of chess, where $|L^*| \approx 50$ [2](ch. 8.93). The sign of L^* only reflects that in basketball (and American sports in general) the visiting team is listed first.

We note that setting $K = 0$ (which amounts to leaving the rating of a team unchanged until it's relegated) increases the variance, but still has predictive power, as evidenced by the ROC curves shown in Figure 2.



(a) Training set. **(b)** Test set.

Figure 2. ROC curve for each of the Elo system implementations.

And finally, setting $L = 0$ barely changes the curve, because the map from $F_X(z)$ to $F_X(z + L)$ is a monotone increasing bijection, and the order of the expected result under the two models are the same for equal K^* (and indeed the K-factors are very similar).

After fitting our model and obtaining the sample, we could expect that for a subsample of matches with rating difference $r_A - r_B \approx r$, the average result would be close to $F_X(r + L)$, but this isn't the case as evidenced by Figure 3.

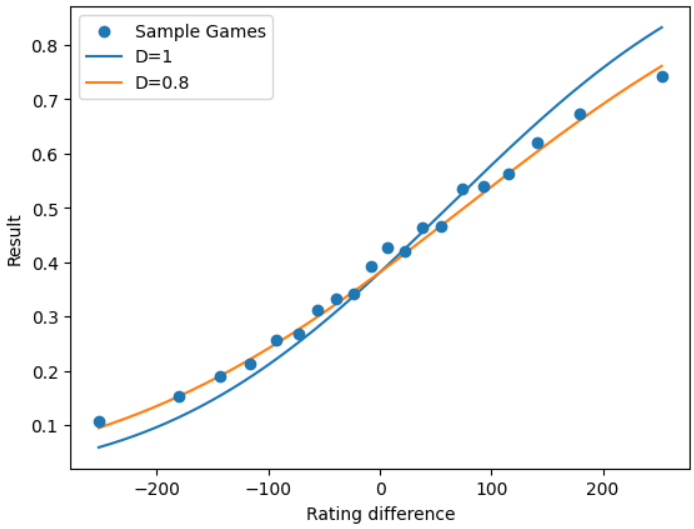


Figure 3. Observed expectation of p_A vs. $r_A - r_B$ versus predicted value $F_X(D(r_A - r_B) + L)$.

 doi:10.20944/preprints202407.1974.v1

We observe that the normal distribution function F_X overestimates the effect of the rating difference in the expected result of a match. We can obtain a better predictor if we multiply $r_A - r_B$ by a "dampening" constant $D \leq 1$, just like in Fig. 3 of [8].

If we minimize the MSE on K, L and D (only using D for the calculation of the mean squared error, since otherwise it's equivalent to scaling X), we obtain optimal parameters $K^* \approx 19.61$, $L^* \approx -58.9$ and $D^* \approx 0.786$, reducing the MSE to 0.21064 in the test set.

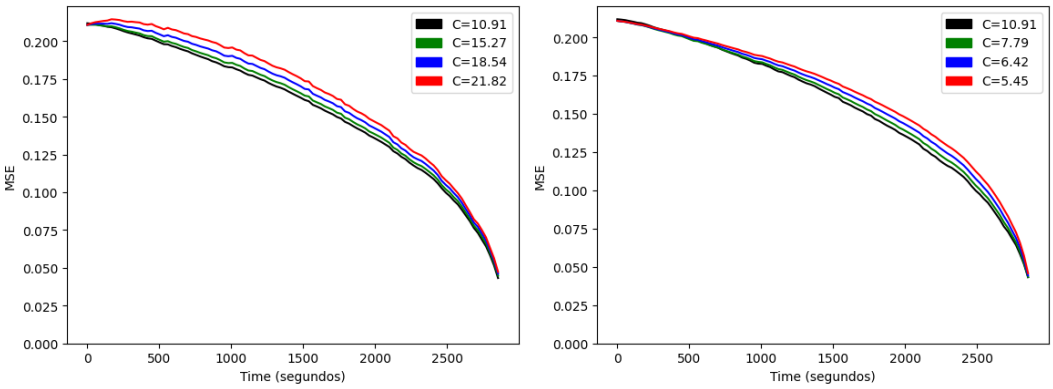
This seems a small improvement in MSE, but it's straightforward to test the significance of the reduction in variance by looking at the variables $X := (p_A - \mathbf{F}_X(L + r_A - r_B))^2$ and $Y := (p_A - \mathbf{F}_X(L + Dr_A - Dr_B))^2$ for a random game.

In order to check if $Z := X - Y$ has a positive mean from our sample of games, we sample the difference $z_k = (p_k^a - \mathbf{F}_X(L + r_{a_k} - r_{b_k}))^2 - (p_k^a - \mathbf{F}_X(L + Dr_{a_k} - Dr_{b_k}))^2$ and argue that by the central limit theorem, the sample mean of Z approximately follows a normal distribution with n times the sample variance of Z. In our testing data, Z has negative mean with $p \approx 1.2 \cdot 10^{-6}$, so D significantly improves our predictions.

4.2.2. Stochastic Elo

Our algorithm leaves us with 12500 rated games out of the 12933 games in the training set, and 17437 out of the 17820 originally in the testing set. These 17437 matches are the ones used for the analysis of our stochastic model.

We start by estimating the only parameter of the stochastic model, C. As we showed in Lemma 1, if the model is correct, the optimal value C^* should minimize the error $MSE(t)$ at every time t, so we minimize the average of that function at different times for a more robust estimation of C^*. In the training set, we obtain $C^* \approx 10.91$, and we can compare the function $MSE(t)$ for higher and lower values, as shown in Figure 4. Since every curve lies above the curve for $C = 10.91$, Lemma 1 suggests that our formula for the expected result is optimal among the family $E[p_A] = \mathbf{F}_X(\Delta r\sqrt{u_t} + g(t)S_t)$. We also note that the error doesn't reach zero at the final time $t = T$ of 2880 seconds, since a tied score at that time results in an extra time being played. Our model assumes a fixed duration, so in order to predict the result beyond minute 48 we would need to model another (shorter) match.



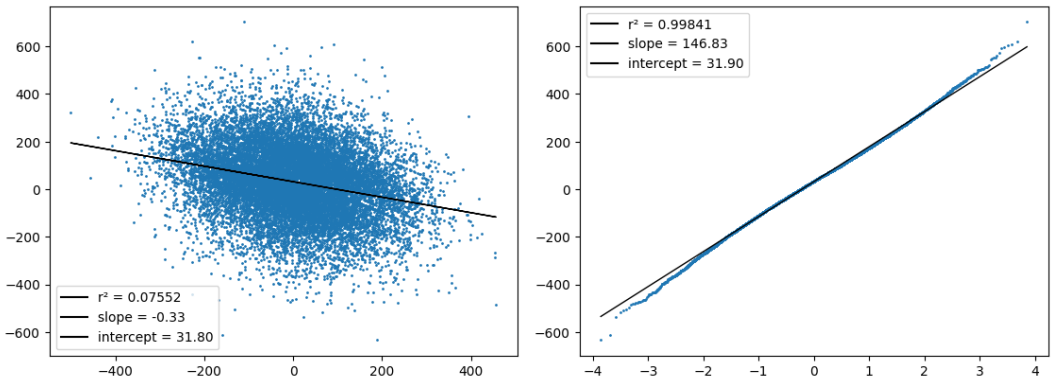
(a) $C > C^*$. **(b)** $C < C^*$.

Figure 4. Mean squared error as a function of time.

Next, Figure 5 shows the result of checking the normality of the score process by computing the normal residues $C \cdot S_T - L - D(r_A - r_B)$ and comparing them with the quantiles of a normal distribution. Here, again, S_T is the score after the first 48 minutes, not the final score of the game. As expected from Eq. (16), the residuals more or less follow a normal distribution, but the mean is not zero, meaning the home team scores better (in terms of S_T) than the Elo model predicts. The standard deviation is also smaller than the expected value of the model, namely $\sigma_X = 200$. From the first plot

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 July 2024 doi:10.20944/preprints202407.1974.v1

we can also tell that the residuals are somewhat correlated to the Elo difference, but not to an extent that would invalidate the model.



(a) Residues vs. Elo difference.

(b) Residues vs. standard normal quantiles.

Figure 5. Tests for the model residues $C \cdot S_T - L - D(r_A - r_B)$.

Finally, Figure 6 checks the lineal relationship between the score variance and time, as described in 18, computing the quantities

$$MSE_S(t) := \frac{1}{|M|}\sum_{i=1}^{|M|}(S_T - \mathbb{E}[S_T|S_t])^2 \xrightarrow[|M|\to\infty]{a.s.} \mathbb{E}[(S_T - \mathbb{E}[S_T|S_t])^2] = Var_t(S_T) = \frac{\sigma_X^2}{C^2}u_t$$

The resulting function is close to a line, but the variance still drops more quickly at the end of the match than at the beginning, suggesting S_T is not completely homogeneous in time. Figure 7 plots the average number of points scored at each 10-second interval of the game by either team, in order to approximate the variance of that interval (they are in fact equal if we assume the score is the difference of two independent Poisson processes as in the discrete model). However, Figure 7 shows that the process is very homogeneous in time, except for brief periods before the end of each quarter.

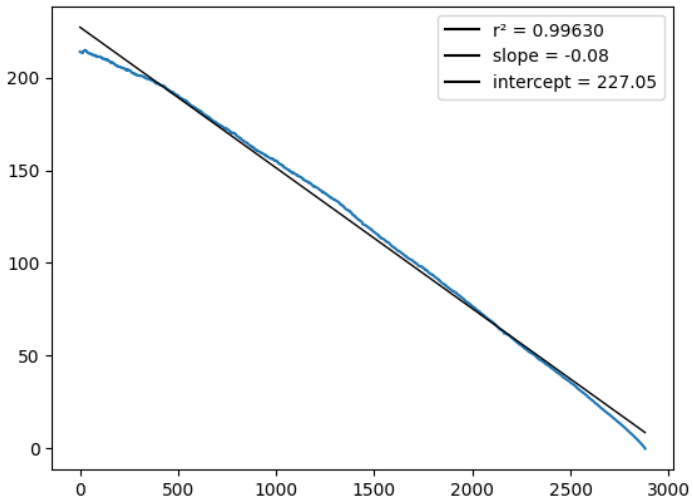


Figure 6. Score prediction variance $MSE_S(t)$ vs. match time.

 doi:10.20944/preprints202407.1974.v1

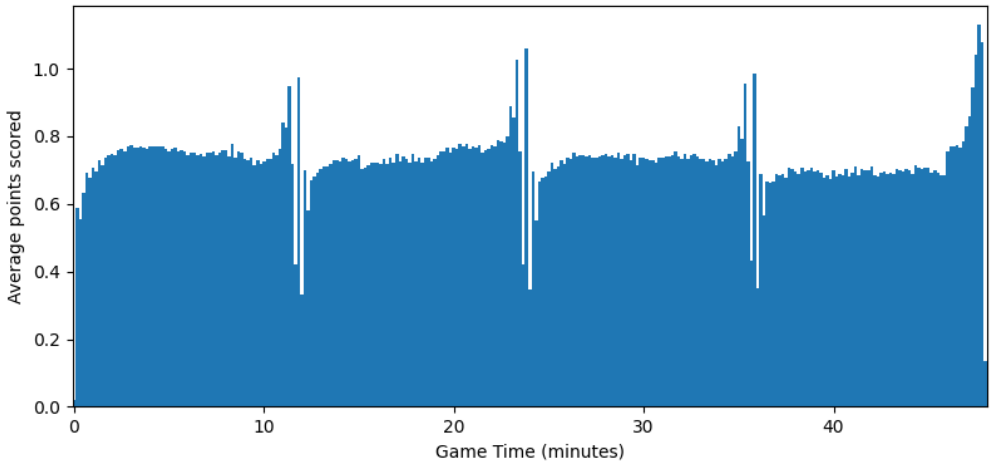


Figure 7. Average increase in combined score vs. match time.

4.3. Soccer Results

Soccer is somewhat more challenging than basketball when it comes to applying Elo models. First, since our database consists of league games, we don't have extra times, at the expense of allowing for draws ($p_A = p_B = \frac{1}{2}$) if the score is tied at the end of the match. We should note that in both the Premier League and the Spanish First Division, teams are awarded 3 points for a victory, 1 for a tie and 0 for each loss, so this isn't entirely a zero-sum game, whereas in the Elo model $p_A + p_B = 1$.

Although there are no extra times, there is time added at the end of each half-time, but our database only keeps track of the minute in which each goal is scored, and every goal scored in the added time will appear in the minute 90 or 45. Finally, S_t is relatively small (usually $S_T < 5$), so approximating it by a continuous variable is problematic.

Our dataset for soccer consists of the seasons 2003-04 to 2023-24 of the Spanish and English first division leagues, both counting 5320 games. We use the former for training and the latter for testing. The data was obtained from fbref.com.

4.3.1. Static Elo

Since the league structure is similar to the NBA, we will use the same algorithm in Section 4.1 to obtain rated games, this time with a sample of $m = 12$ for part I and the same $X \sim N(0, 200)$, now adding also the dampened model to the table along with the optimal D. The corresponding results are presented in Table 2.

Table 2. Mean squared error of different Elo models vs. sample variance.

K^*	L^*	D^*	MSE	MSE_{test}
No Elo	48.11	-	0.17881	0.18188
0	52.12	1	0.17248	0.19252
9.90	0	1	0.16346	0.16136
10.80	52.68	1	0.15420	0.15396
11.82	52.50	0.874	0.15387	0.15341

Just like we saw for basketball, both K and L are clearly significant, but the reduction in variance for adding D is much smaller. The statistical test described for basketball this time gives us that D is significant with $p \approx 0.0013$, and the optimum D^* is closer to 1.

4.3.2. Stochastic Elo

After implementing the static Elo system, we are left with 7176 matches out of the 7980 in our testing set, and we will use these for the analysis of the stochastic model.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 July 2024 doi:10.20944/preprints202407.1974.v1

However, in this case the irregularity of the process versus the recorded time is stronger. As shown in Figure 8, the number of goals scored by every team each minute increases as the game nears its end, and has two spikes in the added time of each half. This makes the simple model we used for basketball less effective. If we estimate C by minimizing the average of $MSE_S(t)$ at evenly spaced points t, with linear $u_t = \frac{T-t}{T}$, the resulting functions $MSE(t)$ and $MSE_S(t) \approx Var(S_T|S_t)$ are depicted in Figures 9 and 10, respectively.

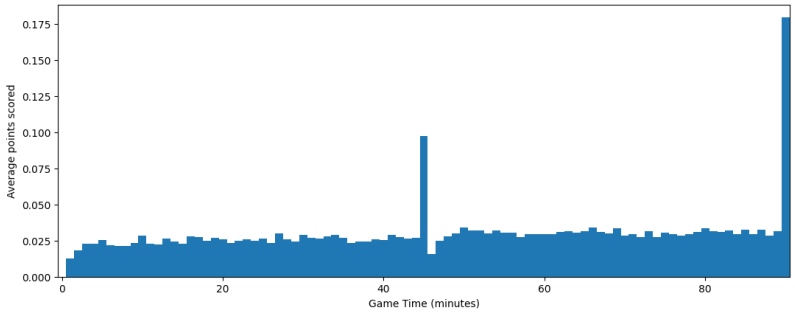
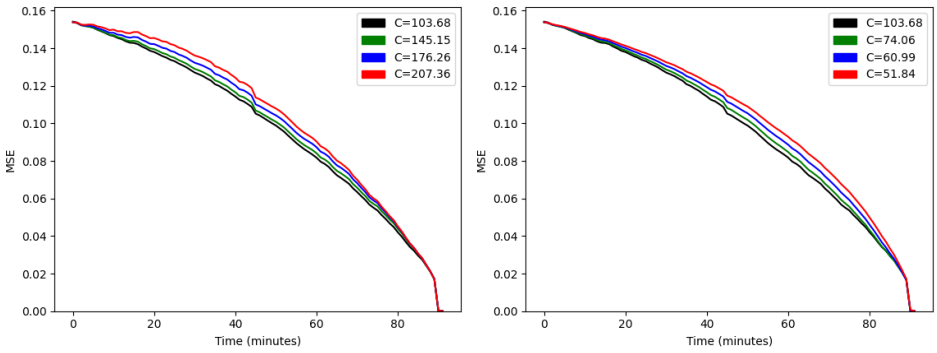


Figure 8. Average number of goals each minute.



(a) $C > C^*$. **(b)** $C < C^*$.

Figure 9. Mean squared error as a function of time.

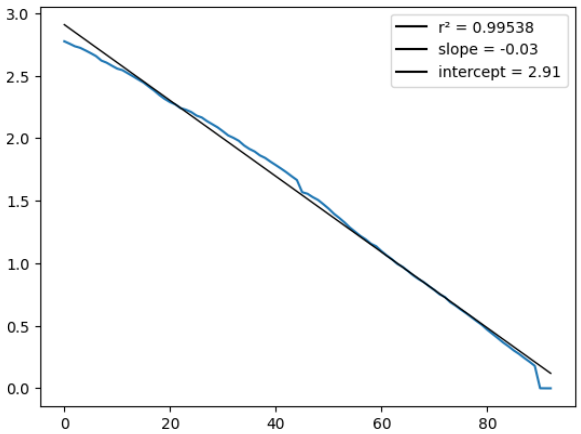
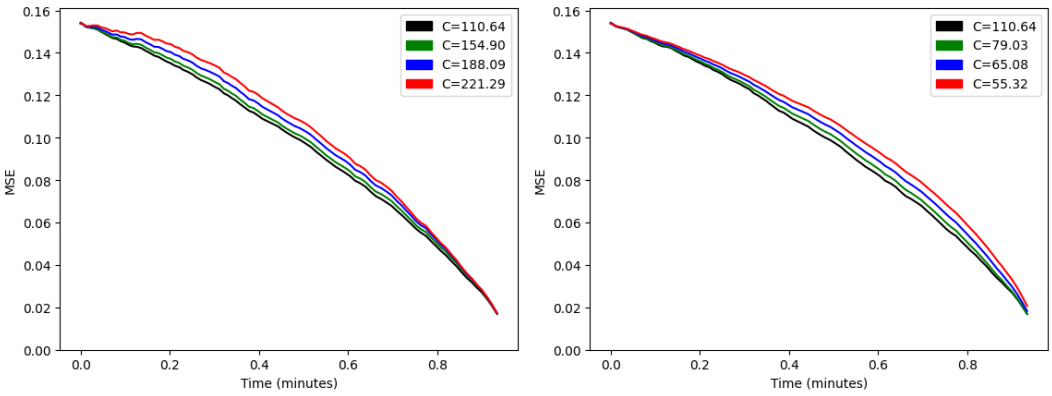


Figure 10. Score prediction variance $MSE_S(t)$ vs. match time.

To tackle this problem, we model the process as a non-homogeneous process, with u_t equal to the fraction of goals in our sample scored after time t. To show the results with this modification,

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 July 2024 doi:10.20944/preprints202407.1974.v1

Figures 11 and 12 replaces the match time in the x axis with the modified time $1 - u_t$. The last plot showcases that the score variance (conditioned on S_t) is almost exactly proportional to u_t, and that our assumptions on the process S_t are reasonable.



(a) $C > C^*$. **(b)** $C < C^*$.

Figure 11. Mean squared error as a function of $1 - u_t$.

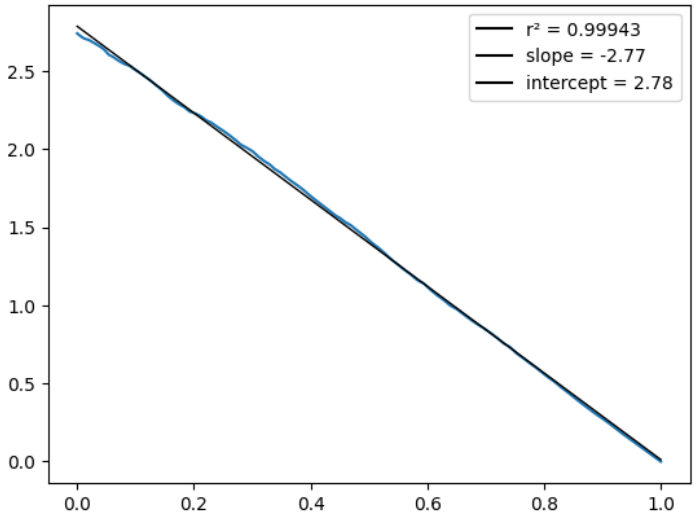
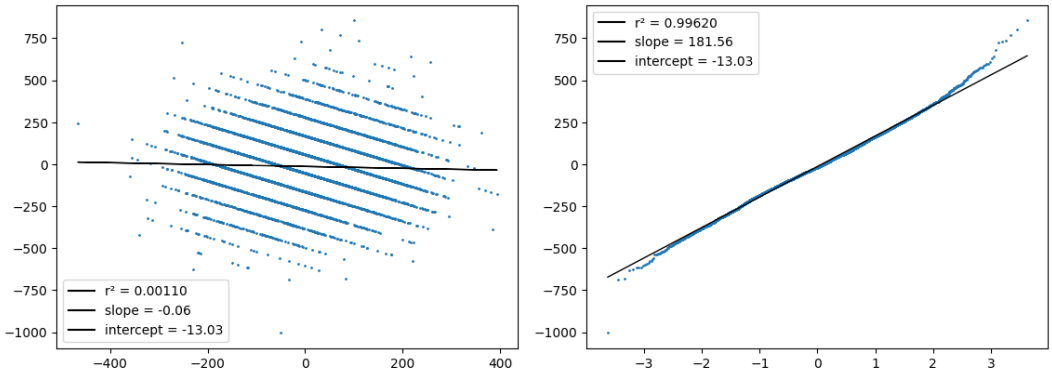


Figure 12. Score prediction variance $MSE_S(t)$ vs. modified time $1 - u_t$.

Finally, Figure 13 plots the residues $C \cdot S_T - L - D(r_A - r_B)$ and compare them to the standard normal quantiles and their respective rating differences $L + D(r_A - r_B)$. The distribution of the residues is close to normal, except for slightly thicker tails. The correlation between the residues and the rating differences is virtually zero compared with the one we saw for basketball.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 July 2024 doi:10.20944/preprints202407.1974.v1



(a) Residues vs. Elo difference.

(b) Residues vs. standard normal quantiles.

Figure 13. Tests for the model residues $C \cdot S_T - L - D(r_A - r_B)$.

4.3.3. Discrete Elo model

Lastly, we implement the discrete score model based on the Skellam process. As we showed, this is a particular case of an Elo model, and thus the same algorithm can be used to estimate the parameters K, L and D. In this case we use $m = 12$ and the function F_X defined in Eq.(19). To estimate the only non-Elo parameter of the model,

$$H = 2\sqrt{\mu_1\mu_2} = 2\sqrt{\mathbb{E}[goals_{home}] + \mathbb{E}[goals_{away}]} = 2\sqrt{\mathbb{E}[goals_{home} \cdot goals_{away}]}$$

we use the sample mean of the product of the goals in our training sample, obtaining $H \approx 2.578$. The corresponding results are presented in Table 3.

Table 3. Mean squared error of restricted Elo models vs. sample variance.

K^*	L^*	D^*	MSE	MSE_{test}
No Elo	0.56115	-	0.17881	0.18188
0	0.61525	1	0.17256	0.19258
0.11702	0	1	0.16348	0.16141
0.12888	0.61560	1	0.15424	0.15397
0.14781	0.62004	0.86536	0.15389	0.15334

Despite the distribution function being different, the mean squared error is very close to that of the standard Elo model with normal F_X. However, this model also predicts probabilities for the three possible results, and thus it allows computing metrics like log-loss [10].

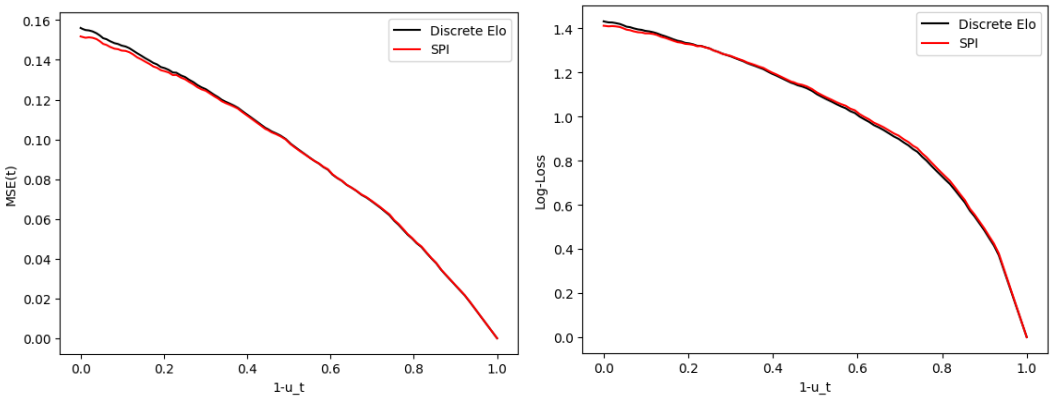
For comparison, we implement the rating system used by the Soccer Power Index (SPI) developed by Nate Silver [13], which uses 2 rating parameters for each team: one for their offense (their capacity to score goals) and another for their defense. A concise but complete explanation of the SPI system is given in Appendix C.

The SPI system achieves a MSE of 0.1518 versus the 0.1533 of the discrete Elo system, and the squared errors are significantly lower, with a p-value of $5 \cdot 10^{-5}$. The average log-loss at time zero is also better, with 1.411 beating the 1.432 for the discrete Elo and a p-value of $8 \cdot 10^{-6}$. However, both systems also give probabilities for the result being a victory, loss or draw at any time t during the game, namely

$$\mathbb{P}[p_A = 1] = \mathbb{P}[Sk(\mu_1 \cdot u_t, \mu_2 \cdot u_t) + S_t > 0]$$
$$\mathbb{P}[p_A = 1/2] = \mathbb{P}[Sk(\mu_1 \cdot u_t, \mu_2 \cdot u_t) + S_t = 0]$$
$$\mathbb{P}[p_A = 0] = \mathbb{P}[Sk(\mu_1 \cdot u_t, \mu_2 \cdot u_t) + S_t < 0]$$

 doi:10.20944/preprints202407.1974.v1

where μ_1 and μ_2 are the expected goals by the home team (A) and the visiting team, respectively. Computing these for every time t, we obtain the results shown in Figure 14. Note that in the log-loss curve, the Elo system beats SPI in the second half - for instance, at $t = 72$ minutes, the log-losses are 0.857 vs. 0.841 with a p-value of $2 \cdot 10^{-7}$ showing that the Elo loss achieves a lower loss. This suggests that despite being a 1-parameter model and its relative simplicity, the discrete Elo system is on par with SPI in terms of predictive power when it comes to mid-game predictions (conditioned on S_t).



(a) Mean Squared Error. **(b)** Log-Loss.

Figure 14. Evolution of model error metrics vs. modified time $1 - u_t$.

5. Conclusions

In this work, it has been shown how the model underlying the Elo system has a natural extension for fixed-duration sports with which it possible to model the score of each team. The proposed stochastic Elo system can be easily implemented for league competitions using the algorithm in Section 4.1, which in turn uses the estimator defined in Section 2.3, and supported by the results proved in Appendix A.

A discrete model for soccer has been also proposed, which is also an Elo system at time $t = 0$, but models the goal difference as a discrete process, thus providing probabilities for each possible outcome of a match. Moreover, we saw that this system has accuracy comparable with that of a biparametric rating system like SPI across the duration of a soccer game. This suggests that the simplicity of the Elo system may favor it as an alternative to sport-specific models, and its use will continue to increase.

5.1. Future Work

These results certainly apply to games or sports with a score board and fixed duration in time, but they could be replicated for other sports with different scoring systems, like table tennis, where the game ends when the first player gets a score of 21, regardless of how long it takes. It's possible that an Elo system can be derived from these type of "race" processes as we did for our Skellam process.

The concept of predicting the result of a match mid-game could also be applied to chess, where there is no objective score S_t but it's common to obtain an "evaluation" from the state of the board, using chess engines. Some chess websites store extremely large databases of computer evaluated games in which a study of this kind could be performed.

Another question we didn't consider when optimizing the parameters of the Elo system, such as K or L, is whether the distribution function F_X can also be inferred from a large enough sample of games. Obviously, the space of distribution functions isn't finite-dimensional, but a method for choosing F_X hasn't been proposed, even among a finite-dimensional family of distributions.

Finally, there are games (such as checkers) for which there is a concept of perfect play, i.e. a strategy that guarantees a victory or a draw. It's easy to see that an expected result above $1/2$ implies an upper bound in ratings from the original model, but the Elo system doesn't have an upper bound

 doi:10.20944/preprints202407.1974.v1

built in. However, some models derived from race processes not only produce odds for a win, draw or loss, but also have a maximum or minimum rating associated with perfect play.

For instance, given ratings $r_1, r_2 \in \mathbb{R}_{\geq 0}$, we consider Poisson processes $P_1(t)$ and $P_2(t)$ with rates r_1 and r_2, and arrival times $T_1(n)$ and $T_2(n)$. Suppose player 1 wins when $T > T_2(k) < T_1(k)$, player 2 wins when $T > T_1(k) < T_2(k)$ and they tie when $T_1(k) > T < T_2(k)$, for fixed $k \in \mathbb{Z}$, $T \in \mathbb{R}$. Then the strength of a player decreases with r, but if $r_1 = 0$ player 1 will never lose, and 0 is a lower bound for ratings. To the best of our knowledge, this type of systems haven't been studied computationally at all.

Supplementary Materials: The computer code for our computational study, as well as the basketball and soccer databases, can be downloaded at https://github.com/gonzalogomezabejon/StochasticElo.

Author Contributions: Conceptualization, G.G.; methodology, G.G. and T.R.; software, G.G; validation, G.G.; formal analysis, G.G; investigation, G.G; resources, G.G; data curation, G.G.; writing—original draft preparation, G.G; writing—review and editing, T.R.; visualization, G.G.; supervision, T.R.; project administration, T.R.; funding acquisition, T.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Government of Spain grant number PID2021-122905NB-C21.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Proofs Regarding the Static Rating Estimators

In this appendix we characterize the existence of the rating estimators $\hat{R}$ that give zero errors $\varepsilon_i(\hat{R}, M)$ for a sample of matches M (assuming F_X is continuous), and prove uniqueness of $\hat{R}$ (up to adding a constant). We also prove that the vector $\varepsilon(R, M)$ is a descent direction for the problem $\min\{\|\varepsilon(R, M)\|_1 : R \in \mathbb{R}^m\}$.

Recall that our sample is given by $M = \{(a_k, b_k, p_k^a, p_k^b) : k = 1, \dots, |M|\}$, where $a_k \in \{1, \dots, m\}$ is the "home" player or team, $b_k \in \{1, \dots, m\}$, is the "visiting" player, and $p_k^a = p_{a_k,k}$ is the score of player a_k, that is, 1 if a_k wins and 0 if a_k loses, or an intermediate value like $\frac{1}{2}$ if there is a draw. Conversely, the score of b_k is $p_k^b = 1 - p_k^a$.

We want to characterize the existence of the static estimator, that is, the set of ratings $\hat{R} = (\hat{r}_1, \hat{r}_2, \dots, \hat{r}_m)$ that verifies, for every player j,

$$0 = \varepsilon_j(\hat{R}, M) := \sum_{k|j=a_k} (p_k^a - \mathbb{E}[p_k^a|\hat{R}]) + \sum_{k|j=b_k} (p_k^b - \mathbb{E}[p_k^b|\hat{R}]) = \sum_{k \in M^j} (p_{j,k} - \mathbb{E}[p_{j,k}|\hat{R}])$$

where $M^j := \{1 \leq k \leq |M| : j = a_k \text{ or } j = b_k\}$ and $p_{a_k,k} = p_k^a$, $p_{b_k,k} = p_k^b$. For the expectancy we use $\mathbb{E}[p_k^a|R] = \mathbf{F}_X(r_a - r_b + L)$, where L is a non-negative constant ($L = 0$ is the standard symmetric case). Finally, we define a digraph $G_M = (\{1, \dots n\}, A_M)$ where $ij \in A_M$ if and only if for some k, $\{i, j\} = \{a_k, b_k\}$ and $p_{i,k} > 0$. We want to show:

Theorem A1 (Existence). *For any sample M with G_M weakly connected (connected as an undirected graph), there exists an estimator $\hat{R}$ with $\varepsilon(\hat{R}, M) = 0$ if and only if G_M is strongly connected (that is, for every pair of players i, j there is a directed (i, j)-path).*

Theorem A2 (Uniqueness). *If there are two estimators $R = (r_1 \dots r_n)$ and $R' = (r'_1 \dots r'_n)$ such that $\varepsilon_j(M, R) = \varepsilon_j(M, R') = 0 \; \forall j = 1 \dots n$, then $r_1 - r'_1 = r_2 - r'_2 = \cdots = r_n - r'_n$*

Lemma A1.

$$\sum_{j=1}^{n} \varepsilon_j(R, M) = 0$$

Proof.

$$\sum_{j=1}^{n} \varepsilon_j(R, M) = \sum_{j=1}^{n} \sum_{k \in M^j} (p_{j,k} - \mathbb{E}[p_{j,k}|R]) =$$

 doi:10.20944/preprints202407.1974.v1

$$= \sum_{k=1}^{|M|} \left(p_{a_k,k} + p_{b_k,k} - \mathbb{E}[p_{a_k,k}|R] - \mathbb{E}[p_{b_k,k}|R]\right) = \sum_{i=1}^{|M|}(1-1) = 0$$

□

Lemma A2. *If the players are partitioned in two sets* $\{1,\dots n\} = I \cup J$ *with* $I \cap J = \emptyset$*, then*

$$\sum_{i\in I}\varepsilon_i(R,M) - \sum_{j\in J}\varepsilon_j(R,M) = 2\sum_{a_k\in I}^{b_k\in J}\left(p_{a_k,k} - \mathbb{E}[p_{a_k,k}|R]\right) - 2\sum_{b_k\in I}^{a_k\in J}\left(p_{a_k,k} - \mathbb{E}[p_{a_k,k}|R]\right)$$

Proof. We sort the games of M in the sets $M = M^{II} \cup M^{IJ} \cup M^{JI} \cup M^{JJ}$ where the set $M^{PQ} := \{1 \le k \le |M| : a_k \in P, b_k \in Q\}$ contains the indices of the matches where the first player is in the set P and the second in Q. Then, by Lemma A1,

$$\sum_{i\in I}\varepsilon_i(R,M) - \sum_{j\in J}\varepsilon_j(R,M) = 2\sum_{i\in I}\varepsilon_i(R,M) = 2\sum_{i\in I}\sum_{k\in M^i}(p_{i,k} - \mathbb{E}[p_{i,k}|R]) =$$

$$= 2\sum_{i\in I}\sum_{k\in M^i\cap M^{II}}(p_{i,k} - \mathbb{E}[p_{i,k}|R]) + 2\sum_{i\in I}\sum_{k\in M^i\cap M^{IJ}}(p_{i,k} - \mathbb{E}[p_{i,k}|R]) + 2\sum_{i\in I}\sum_{k\in M^i\cap M^{JI}}(p_{i,k} - \mathbb{E}[p_{i,k}|R])$$

but since M^{II} only contains games between players in I, again by Lemma A1,

$$\sum_{i\in I}\sum_{k\in M^i\cap M^{II}}(p_{i,k} - \mathbb{E}[p_{i,k}|R]) = \sum_{i\in I}\varepsilon_i(R,M^{II}) = 0$$

and finally, since every game in M^{IJ} has exactly one player in I,

$$\sum_{i\in I}\sum_{k\in M^i\cap M^{IJ}}(p_{i,k} - \mathbb{E}[p_{i,k}|R]) = \sum_{k\in M^{IJ}}(p_{a_k,k} - \mathbb{E}[p_{a_k,k}|R]) = \sum_{a_k\in I}^{b_k\in J}\left(p_{a_k,k} - \mathbb{E}[p_{a_k,k}|R]\right)$$

and we use the same argument in M^{JI} to obtain

$$\sum_{i\in I}\sum_{k\in M^i\cap M^{JI}}(p_{i,k} - \mathbb{E}[p_{i,k}|R]) = \sum_{k\in M^{JI}}(p_{b_k,k} - \mathbb{E}[p_{b_k,k}|R]) = -\sum_{b_k\in I}^{a_k\in J}\left(p_{a_k,k} - \mathbb{E}[p_{a_k,k}|R]\right)$$

as we wanted (since $p_{a_k,k} + p_{b_k,k} = 1 = \mathbb{E}[p_{a_k,k}|R] + \mathbb{E}[p_{b_k,k}|R]$). □

Proof of theorem A1. ($\Rightarrow$) If the static estimator $\hat{R}$ exists but G_M isn't strongly connected, then there are players i and j such that there are no (i,j)-paths in G_M. In that case, if we define $I := \{1 \le p \le n : \exists P\ (i,p)\text{-path in } G_M\}$ as the set of players reachable from i in G_M, and $J := \{1\dots n\} \setminus I$ to be the set of players not reachable from I. By construction, in a game $a_k \in I$ and $b_k \in J$, if $p_{a_k,k} > 0$ we would have an edge between $a_k \in I$ and $b_k \in J$, which contradicts the definition of I and J, so $p_{a_k,k} = 0$.

Note that $i \in I$ and $j \in J$ so the sets aren't empty, and since G_M is weakly connected, there must be a game between a player in J to another in I, let's say w.l.o.g. that $(a_l, b_l) \in I \times J$, and in that case, by lemma A2,

$$0 = \sum_{i\in I}\varepsilon_i(\hat{R},M) - \sum_{j\in J}\varepsilon_j(\hat{R},M) = 2\sum_{a_k\in I}^{b_k\in J}\left(p_{a_k,k} - \mathbb{E}[p_{a_k,k}|\hat{R}]\right) - 2\sum_{b_k\in I}^{a_k\in J}\left(p_{a_k,k} - \mathbb{E}[p_{a_k,k}|\hat{R}]\right) =$$

$$= 2\sum_{a_k\in I}^{b_k\in J}\left(0 - \mathbf{F}_X(\hat{r}_{a_k} - \hat{r}_{b_k} + K)\right) - 2\sum_{b_k\in I}^{a_k\in J}\left(1 - \mathbf{F}_X(\hat{r}_{a_k} - \hat{r}_{b_k} + K)\right) \le 2(0 - \mathbf{F}_X(\hat{r}_{a_l} - \hat{r}_{b_l} + K)) < 0$$

 doi:10.20944/preprints202407.1974.v1

since $F_X(x) \in (0,1)$ for any $x \in \mathbb{R}$, and that's not possible.

($\Leftarrow$) On the other hand, suppose G_M is strongly connected. Then we show that $\hat{r}$ exists by proving a stronger proposition: if we fix $r_1, r_2, \ldots, r_{n-m}$, there are $\hat{r}_{n-m+1}(a), \ldots, \hat{r}_n \in \mathbb{R}$ such that for every $j \in \{n-m+1, \ldots, n\}$, we have $\varepsilon_j((r_1, \ldots, \hat{r}_n), M) = 0$. To do this, we show by induction in m that there exist continuous functions $\phi_m : \mathbb{R}^{n-m} \to \mathbb{R}^m$ for $1 \leq m \leq n-1$ such that $\varepsilon_h((\vec{r}, \phi_m(\vec{r})), M) = 0$ for any $\vec{r} \in \mathbb{R}^{n-m}$ and $h \geq n-m+1$, and also, that every component of $\phi_m()$ is non-decreasing in each of its arguments (that is, $\vec{r}' \geq \vec{r} \Rightarrow \phi_m(\vec{r}') \geq \phi_m(\vec{r})$). For brevity, we write $\varepsilon_i(\vec{r})$ instead of $\varepsilon_i(\vec{r}, M)$.

First, we note that since G_M is weakly connected, $\varepsilon_j((r_1, r_2, \ldots, r_n))$ is strictly decreasing in r_j, but non-decreasing in any other rating r_i for $i \neq j$:

$$r'_j > r_j \Rightarrow \varepsilon_j((r_1 \ldots r'_j \ldots r_n)) = \sum_{a_k=j} (p^a_k - \mathbf{F}_X(r'_j - r_{b_k} + L)) + \sum_{b_k=j} (p^b_k - \mathbf{F}_X(r'_j - r_{a_k} - L)) <$$

$$< \sum_{a_k=j} (p^a_k - \mathbf{F}_X(r_j - r_{b_k} + L)) + \sum_{b_k=j} (p^b_k - \mathbf{F}_X(r_j - r_{a_k} - L)) = \varepsilon_j((r_1 \ldots r_j \ldots r_n))$$

since one of the sums is non-empty and F_X is strictly increasing. From the same expression it's easy to see that $r'_i \geq r_i \Rightarrow \varepsilon_j((r_1 \ldots r'_i \ldots r_n)) \geq \varepsilon_j((r_1 \ldots r_i \ldots r_n))$

For the base case of induction, $m = 1$, it's easy to show that the increasing function $f : x \mapsto \varepsilon_n((r_1, \ldots, r_{n-1}, x))$ has a unique zero, because

$$f(x) = \sum_{k|a_k=n} (p^a_k - \mathbf{F}_X(x - r_{b_k} + L)) + \sum_{k|b_k=n} (p^b_k - \mathbf{F}_X(x - r_{a_k} - L)) \xrightarrow[x\to\infty]{} \sum_{k \in M^n} (p_{n,k} - 1) < 0$$

(since G_M is strongly connected, player n scores less than 1 in some match k) and similarly

$$f(x) = \sum_{k|a_k=n} (p^a_k - \mathbf{F}_X(x - r_{b_k} + L)) + \sum_{k|b_k=n} (p^b_k - \mathbf{F}_X(x - r_{a_k} - L)) \xrightarrow[x\to-\infty]{} \sum_{k \in M^n} p_{n,k} < 0$$

and f is obviously continuous (F_X is continuous), so just applying the intermediate value theorem allows us to define $\phi_1((r_1, \ldots, r_{n-1}))$ as the only zero of f. To see that ϕ_1 is non-decreasing in every coordinate of its argument, we check that

$$\vec{r}' \geq \vec{r} = (r_1 \ldots r_{n-1}) \Rightarrow \varepsilon_j(\vec{r}', \phi_1(\vec{r}')) = 0 = \varepsilon_j(\vec{r}, \phi_1(\vec{r})) \leq \varepsilon_j(\vec{r}', \phi_1(\vec{r})) \Rightarrow \phi_1(\vec{r}) \leq \phi_1(\vec{r}')$$

and the last implication is a consequence of ε_n being strictly decreasing in r_n.

To see that ϕ_1 is continuous, we note that for any convergent sequence $\{\vec{r}_k\}_{k\in\mathbb{N}}$ with $\vec{r}^* = \lim_k \vec{r}_k$, we have

$$\lim_k \varepsilon_n(\vec{r}^*, \phi_1(\vec{r}_k)) = \lim_k \varepsilon_n(\vec{r}_k, \phi_1(\vec{r}_k)) = \lim_k 0 = 0 = \varepsilon_n(\vec{r}^*, \phi_1(\vec{r}^*))$$

and since $\varepsilon_n((\vec{r}^*, \cdot)) = f(\cdot)$ is strictly monotone and continuous, and therefore bijective, $\phi_1(\vec{r}^*) = \lim_k \phi_1(\vec{r}_k)$.

Now for the induction step: suppose ϕ_{m-1} exists and is non-decreasing in and with respect to every coordinate, and $m < n$. For fixed $\vec{r} = (r_1, r_2, \cdots r_{n-m})$, we will show that the continuous function $f : x \mapsto \varepsilon_{n-m+1}((\vec{r}, x, \phi_{m-1}((\vec{r}, x))))$, which inherits continuity from $\mathbf{F}_X$ and ϕ_{m-1}, is strictly decreasing.

Suppose $x' > x$. Then, by our induction hypothesis $\phi_{m-1}((\vec{r}, x')) \geq \phi_{m-1}((\vec{r}, x'))$. Let's define $I := \{i | (\vec{r}, x', \phi_{m-1}(\vec{r}, x'))_i > (\vec{r}, x, \phi_{m-1}(\vec{r}, x))_i\}$ and $J := \{1 \ldots n\} \setminus I = \{j | (\vec{r}, x', \phi_{m-1}(\vec{r}, x'))_j = (\vec{r}, x, \phi_{m-1}(\vec{r}, x))_j\}$, and note that $1 \in J$ since $\dim(\vec{r}) = n - m \geq 1$, and $n - m + 1 \in I$. Note that $i \in I \Rightarrow i = n - m + 1$ or $i > n - m + 1$, and in the latter case $\varepsilon_i(\vec{r}, x, \phi_{m-1}(\vec{r}, x)) = 0$ by definition of ϕ_{m-1}. Using that fact, and then lemmas A1 and A2,

 doi:10.20944/preprints202407.1974.v1

$$f(x) = \varepsilon_{n-m+1}(\vec{r}, x, \phi_{m-1}(\vec{r}, x)) = \sum_{i\in I} \varepsilon_i(\vec{r}, x, \phi_{m-1}(\vec{r}, x)) =$$

$$\frac{1}{2}\left(\sum_{i\in I}\varepsilon_i(R) - \sum_{j\in J}\varepsilon_j(R)\right) = \sum_{a_k\in I}^{b_k\in J}(p_{a_k,k} - \mathbb{E}[p_{a_k,k}|R]) - \sum_{a_k\in J}^{b_k\in I}(p_{a_k,k} - \mathbb{E}[p_{a_k,k}|R])$$

where $R = (\vec{r}, x, \phi_{m-1}(\vec{r}, x))$. Therefore,

$$f(x') - f(x) = \sum_{a_k\in I}^{b_k\in J}(\mathbb{E}[p_{a_k,k}|R] - \mathbb{E}[p_{a_k,k}|R']) - \sum_{a_k\in J}^{b_k\in I}(\mathbb{E}[p_{a_k,k}|R] - \mathbb{E}[p_{a_k,k}|R']) < 0$$

since at least one of the sums is non-empty (G_M is connected) and $\mathbf{F}_X(R_i - R_j \pm L) - \mathbf{F}_X(R'_i - R'_j \pm L) = \mathbf{F}_X(R_i - R_j \pm L) - \mathbf{F}_X(R'_i - R_j \pm L) < 0$ for any game between $j \in J$ and $i \in I$.

We also check that $f(x)$ approaches a positive number as x goes to infinity. Since $\phi_{m-1}(\vec{r}, x)$ is non-decreasing in x, its coordinates either have a limit as x goes to infinity, or they also go to infinity (by the monotone convergence theorem), and we can define

$$I := \{i|\lim_{x\to\infty}(\vec{r}, x', \phi_{m-1}(\vec{r}, x'))_i = \infty\} \qquad \text{and} \qquad J := \{j|\lim_{x\to\infty}(\vec{r}, x', \phi_{m-1}(\vec{r}, x'))_j \in \mathbb{R}\}$$

for which again, $n - m + 1 \in I$ and $1 \in J$, and as before,

$$f(x) = \sum_{a_k\in I}^{b_k\in J}(p_{a_k,k} - \mathbb{E}[p_{a_k,k}|R]) - \sum_{a_k\in J}^{b_k\in I}(p_{a_k,k} - \mathbb{E}[p_{a_k,k}|R]) \xrightarrow[x\to\infty]{} \sum_{a_k\in I}^{b_k\in J}(p_{a_k,k} - 1) - \sum_{a_k\in J}^{b_k\in I}(p_{a_k,k} - 0) < 0$$

since G_M is connected and therefore one of the sums is non-empty. An analogous argument shows that as $\lim_{x\to-\infty} f(x) > 0$, and thus, for any given $\vec{r}$, $f(x)$ has a unique zero, which we denote by $g(\vec{r})$, and with that we can define $\phi_m := (g(\vec{r}), \phi_{m-1}((\vec{r}, g(\vec{r}))))$.

Now, we can show that $g(\cdot)$ is non-decreasing, because if $\vec{r}' \geq \vec{r}$, then by induction hypothesis, $\phi_{m-1}(\vec{r}', g(\vec{r})) \geq \phi_{m-1}(\vec{r}, g(\vec{r}))$, and hence

$$\varepsilon_{n-m+1}(\vec{r}, g(\vec{r}), \phi_{m-1}(\vec{r}, g(\vec{r}))) = 0 = \varepsilon_{n-m+1}(\vec{r}, g(\vec{r}), \phi_{m-1}(\vec{r}, g(\vec{r}))) \geq \varepsilon_{n-m+1}(\vec{r}', g(\vec{r}), \phi_{m-1}(\vec{r}', g(\vec{r})))$$

which implies (by the fact that $x \mapsto \varepsilon_{n-m+1}((\vec{r}, x, \phi_{m-1}((\vec{r}, x))))$ is strictly decreasing and continuous, i.e. bijective) that $g(\vec{r}') \geq g(\vec{r})$.

This in turn implies that ϕ_m is non-decreasing by composition of g and ϕ_{m-1}, and to conclude the induction, we only need to check that g is continuous. This is similar to the base case, using that for any convergent sequence $\{\vec{r}_k\}_{k\in\mathbb{N}}$ with $\vec{r}^* = \lim_k \vec{r}_k$, we have

$$\lim_k \varepsilon_{n-m+1}(\vec{r}^*, g(\vec{r}_k), \phi_{m-1}(\vec{r}^*, g(\vec{r}_k))) = \lim_k \varepsilon_{n-m+1}(\vec{r}_k, g(\vec{r}_k), \phi_{m-1}(\vec{r}_k, g(\vec{r}_k))) = 0 = \varepsilon_{n-m+1}(\vec{r}^*, g(\vec{r}^*), \phi_{m-1}(\vec{r}^*, g(\vec{r^*})))$$

and invoking that $f^* : x \mapsto \varepsilon_{n-m+1}((\vec{r}, x, \phi_{m-1}((\vec{r}, x))))$ is bijective again.

Our induction works for $m = 1 \ldots n-1$, and to complete the proof (for $m = n$) we consider $\hat{R} = (0, \phi_{n-1}(0))$. By definition of ϕ_{n-1}, we know that $\varepsilon_2(\hat{R}) = \varepsilon_3(\hat{R}) = \cdots = \varepsilon_n(\hat{R}) = 0$, and by lemma A1, $\varepsilon_1(\hat{R}) = \sum_{i=2}^n \varepsilon_i(\hat{R}) = 0$, as we wanted. □

Proof of theorem A2 (uniqueness). Let's suppose there are two rating vectors $R = (r_1, \ldots, r_n)$ and $Q = (q_1, \ldots, q_n)$ such that $\varepsilon(R, M) = \varepsilon(Q, M) = 0$. The residues ε are the same for the normalized ratings with mean zero $R' = R - \frac{1}{n}\sum i = 1^n r_i$ and $Q\sum_{i=1}^n r_i = \sum_{i=1}^n q_i = 0$. Therefore, if $R - Q$ isn't

 doi:10.20944/preprints202407.1974.v1

constant, $R' - Q' \neq 0$, and there are $i, j \in \{1, \dots, n\}$ such that $r'_i < q'_i$ and $r'_j > q'_j$. In that case, let $I := \{1 \leq i \leq n : r'_i < q'_i\}$ and $J := \{1 \leq j \leq n : r'_j \geq q'_j\}$.

$$0 = \sum_{i \in I} \varepsilon(R', M) - \sum_{j \in J} \varepsilon(R', M) - \left(\sum_{i \in I} \varepsilon(Q', M) - \sum_{j \in J} \varepsilon(Q', M) \right) =$$

$$= 2 \sum_{k \in M^{IJ}} p_{a_k,k} - \mathbb{E}[p_{a_k,k}|R'] - (p_{a_k,k} - \mathbb{E}[p_{a_k,k}|Q']) - 2 \sum_{k \in M^{JI}} p_{a_k,k} - \mathbb{E}[p_{a_k,k}|R'] - (p_{a_k,k} - \mathbb{E}[p_{a_k,k}|Q']) =$$

$$= 2 \sum_{k \in M^{IJ}} \Big(F_X(q'_{a_k} - q'_{b_k} + L) - F_X(r'_{a_k} - r'_{b_k} + L) \Big) + 2 \sum_{k \in M^{JI}} \Big(F_X(r'_{a_k} - r'_{b_k} + L) - F_X(q'_{a_k} - q'_{b_k} + L) \Big)$$

but every term in the sums is positive, because for any players $(i, j) \in I \times J$, we have that $r'_i - q'_i < 0 \leq r'_j - q'_j \Rightarrow r'_i - r'_j < q'_i - q'_j \Rightarrow F_X(r'_i - r'_j + L) < F_X(q'_i - q'_j + L)$, and similarly $F_X(r'_j - r'_i + L) > F_X(q'_j - q'_i + L)$. Since one of the sums is non-empty (otherwise I and J would be disconnected in the undirected G_M), the one of the two sums is positive and the other is non-negative, which is a contradiction. Therefore, $\varepsilon(R, M) = 0 = \varepsilon(Q, M) \Rightarrow R' = Q'$ and therefore R must be equal to Q plus a constant. □

Theorem A3. *$\varepsilon(R, M)$ is a descent direction for the problem* $\min_{R \in \mathbb{R}^n} \|\varepsilon(R, M)\|_2^2$

Proof. We want to show that $(\nabla \|\varepsilon(R, M)\|_2^2)^T \varepsilon(R, M) < 0$ for $\varepsilon \neq 0$. First, the partial derivatives of ε_i are, for $i \neq j$,

$$\frac{d\varepsilon_i(R, M)}{dr_j} = \frac{d}{dr_j} \sum_{k}^{a_k = i} p_{i,k} - F_X(r_i - r_{b_k} + L) + \frac{d}{dr_j} \sum_{k}^{b_k = i} p_{i,k} - F_X(r_i - r_{a_k} - L) =$$

$$= \sum_{a_k = i}^{b_k = j} f_X(r_{a_k} - r_{b_k} + L) + \sum_{a_k = j}^{b_k = i} f_X(r_{b_k} - r_{a_k} - L) = \sum_{k \in M^i \cap M^j} f_X(r_{a_k} - r_{b_k} + L)$$

since f_X is symmetric around zero, and for $i = j$,

$$\frac{d\varepsilon_j(R, M)}{dr_j} = \frac{d}{dr_j} \sum_{k}^{a_k = j} -f_X(r_j - r_{b_k} + L) + \frac{d}{dr_j} \sum_{k}^{b_k = j} -f_X(r_j - r_{a_k} - L) = - \sum_{k \in M^j} f_X(r_{a_k} - r_{b_k} + L)$$

so the scalar product we want is

$$(\nabla \|\varepsilon\|_2^2)^T \varepsilon = \sum_{i=1}^{n} \varepsilon_i(R, M) \cdot \frac{d}{dr_i} \|\varepsilon(R, M)\|_2^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} 2\varepsilon_i(R, M) \varepsilon_j(R, M) \frac{d\varepsilon_j(R, M)}{dr_i} =$$

$$= -2 \sum_{i=1}^{n} \varepsilon_i(R, M)^2 \sum_{k \in M^i} f_X(r_{a_k} - r_{b_k} + L) + 4 \sum_{i \neq j} \varepsilon_i(R, M) \varepsilon_j(R, M) \sum_{k \in M^i \cap M^j} f_X(r_{a_k} - r_{b_k} + L) =$$

$$= 2 \sum_{k=1}^{|M|} \Big(2\varepsilon_{a_k}(R, M) \varepsilon_{b_k}(R, M) - \varepsilon_{a_k}(R, M)^2 - \varepsilon_{b_k}(R, M)^2 \Big) f_X(r_{a_k} - r_{b_k} + L) =$$

$$= -2 \sum_{k=1}^{|M|} \big(\varepsilon_{a_k}(R, M) - \varepsilon_{b_k}(R, M) \big)^2 f_X(r_{a_k} - r_{b_k} + L) \leq 0$$

And if $\varepsilon(R, M) \neq 0$, by lemma **A1** there must be strictly negative and non-negative components of $\varepsilon(R, M)$, and a game between one of each (by weak connectedness of G_M), so at least one pairing (a_k, b_k) of M has $\varepsilon_{a_k} \neq \varepsilon_{b_k}$, and the associated term of the sum above is strictly positive, giving $(\nabla \|\varepsilon\|_2^2)^T \varepsilon < 0$ □

 doi:10.20944/preprints202407.1974.v1

Appendix B. Proof That the Discrete Model Is an Elo Model

Here we prove that the discrete model proposed in section 3.4 is a particular type of Elo model, or in other words, that if $\mathbb{E}[p_A] = \mathbb{P}[Sk(\mu_1, \mu_2) > 0] + \frac{1}{2}\mathbb{P}[Sk(\mu_1, \mu_2) = 0]$ for

$$\mu_1 = \frac{1}{2}\left(\Delta_r + \sqrt{\Delta_r^2 + H^2}\right) \qquad \text{and} \qquad \mu_2 = \frac{1}{2}\left(-\Delta_r + \sqrt{\Delta_r^2 + H^2}\right)$$

there exists a continuous distribution X with domain $\mathbb{R}$ for which $\mathbb{E}[p_A] = F_X(\Delta r)$

Proof. Let's denote

$$F(x) := \mathbb{P}[Sk(\mu_1(x), \mu_2(x)) > 0] + \frac{1}{2}t\mathbb{P}[Sk(\mu_1(x), \mu_2(x)) = 0] =$$

$$= \mathbb{P}\left[Sk\left(\frac{\sqrt{x+H^2}-x}{2}, \frac{\sqrt{x+H^2}-x}{2}\right) > 0\right] + \frac{1}{2}\mathbb{P}\left[Sk\left(\frac{\sqrt{x+H^2}-x}{2}, \frac{\sqrt{x+H^2}-x}{2}\right) = 0\right]$$

Here we are assuming $H^2 > 0$, since $\frac{1}{4}H^2$ is the expectation of a non-negative quantity (the product of the goals scored by each team). This means that μ_1 is strictly increasing in Δr, and $\mu_2 = \frac{H^2}{4\mu_1}$ is strictly decreasing in Δr.

On the other hand, $g_k(a, b) := \mathbb{P}[Sk(a, b) \geq k]$ is increasing in a and decreasing in b, because for any $\varepsilon > 0$, we have

$$g_k(a+\varepsilon, b) = \mathbb{P}[Sk(a+\varepsilon, b) \geq k] = \mathbb{P}[Sk(a, b) + P(\varepsilon) \geq k] > \mathbb{P}[Sk(a, b) \geq k] = g_k(a, b)$$

$$g_k(a, b) = \mathbb{P}[Sk(a, b) \geq k] > \mathbb{P}[Sk(a, b) - P(\varepsilon) \geq k] = g_k(a, b+\varepsilon)$$

where $P(\varepsilon)$ is a Poisson variable with mean ε.

This two propositions imply that $F(\Delta r) = \frac{1}{2}\mathbb{P}[Sk(\mu_1, \mu_2) \geq 1] + \frac{1}{2}\mathbb{P}[Sk(\mu_1, \mu_2) \geq 0] = \frac{g_1(\mu_1,\mu_2)+g_0(\mu_1,\mu_2)}{2}$ is strictly increasing in μ_1 and decreasing in μ_2, and therefore increasing as a function of Δr. To finish, we only need to show that $\mathbb{E}[p_A]$ goes to 1 as Δr approaches $+\infty$, and 0 when Δr goes to $-\infty$, since any monotone function with that asymptotic behavior is a distribution function.

As Δr goes to infinity, μ_1 goes to infinity and μ_2 goes to zero, so

$$F(\Delta r) = \mathbb{P}[P(\mu_1) > P(\mu_2)] + \frac{1}{2}\mathbb{P}[P(\mu_1) = P(\mu_2)] \geq \mathbb{P}[P(\mu_1) > 0 = P(\mu_2)] =$$

$$= (1 - e^{-\mu_1})e^{-\mu_2} \xrightarrow[\Delta r \to \infty]{} (1 - e^{-\infty})e^0 = 1$$

And of course $F(\Delta r) = \frac{1}{2}(\mathbb{P}[Sk(\mu_1, \mu_2) \geq 1] + \mathbb{P}[Sk(\mu_1, \mu_2) \geq 0]) \geq 1$, so by the sandwich rule,

$$F(\Delta r) \xrightarrow[\Delta r \to \infty]{} 1$$

A similar argument shows that when Δr goes to $-\infty$,

$$F(\Delta r) \leq \mathbb{P}[P(\mu_1) > 0] + \frac{1}{2}\mathbb{P}[P(\mu_2) = 0] + \frac{1}{2}\mathbb{P}[P(\mu_1) > 0] \leq$$

$$\leq \frac{3}{2}(1 - e^{-\mu_1}) + \frac{1}{2}e^{-\mu_2} \xrightarrow[\Delta r \to \infty]{} \frac{3}{2}(1 - e^0) + \frac{1}{2}e^{-\infty} = 0$$

and F is bounded below by zero, so both limits hold and F is the distribution function of some variable X. □

 doi:10.20944/preprints202407.1974.v1

Appendix C. Implementation of the Rating System from SPI

The SPI rating system that we used as a benchmark to compare our discrete model is described in [12] and [13]. The system uses two rating parameters for each team instead of only one, namely OFF_A, which is higher the more goals team A is likely to score, and DEF_A, which is higher the more goals A will allow its rivals to score.

In a match between A and B, in which A scores G_A goals and B scores G_B, we define the Adjusted Goals Scored of A as

$$AGS_A := (G_A - DEF_B)\frac{0.424 \cdot AVG_{BASE} + 0.548}{\max(0.25, 0.424 \cdot DEF_B + 0.548)} + AVG_{BASE}$$

where AVG_{BASE} is the average number of goals scored by each team in a game. Similarly, the Average Goals Allowed of A is defined as

$$AGA_A := (G_B - OFF_B)\frac{0.424 \cdot AVG_{BASE} + 0.548}{\max(0.25, 0.424 \cdot OFF_B + 0.548)} + AVG_{BASE}$$

and both quantities are similarly defined for B. The model then states that $\mathbb{E}[AGS_A] = OFF_A$ and $\mathbb{E}[AGA_A] = DEF_A$. From this, given some starting values for both parameters for both teams, we can infer AGS_A, AGA_A, AGS_B and AGA_B for each game, and correct the parameters by

$$\begin{aligned} OFF'_A &= \lambda AGS_A + (1-\lambda)OFF_A & DEF'_A &= \lambda AGA_A + (1-\lambda)DEF_A \\ OFF'_B &= \lambda AGS_B + (1-\lambda)OFF_B & DEF'_B &= \lambda AGA_B + (1-\lambda)DEF_B \end{aligned} \tag{A1}$$

Since AGA_A and AGS_A are linear in G_B and G_A respectively, we obtain the expectation of G_B and G_A from them as follows:

$$\mathbb{E}[G_A] = f_{A,B} := (OFF_A - AVG_{BASE})\frac{\max(0.25, 0.424 \cdot DEF_B + 0.548)}{0.424 \cdot AVG_{BASE} + 0.548} + DEF_B$$

$$\mathbb{E}[G_B] = g_{B,A} := (DEF_A - AVG_{BASE})\frac{\max(0.25, 0.424 \cdot OFF_B + 0.548)}{0.424 \cdot AVG_{BASE} + 0.548} + OFF_B$$

There are two issues with this: first, $g_{A,B}$ and $f_{B,A}$ should also match $\mathbb{E}[G_A]$ and $\mathbb{E}[G_B]$ respectively, but they don't necessarily match: in practice, we use $\mathbb{E}[G_A] = \frac{f_{A,B}+g_{A,B}}{2}$ for estimating the win and draw odds in our code. The second issue is that the resulting value of $\mathbb{E}[G_A]$ or $\mathbb{E}[G_B]$ could be negative, and in that case, we impute 10^{-6} as the expected goals of that team.

Finally, in order to account for the home field advantage, supposing A and B play on A's field, we modify the formulas above as follows:

$$AGS_A := (G_A - DEF_B)\frac{0.424 \cdot AVG_{BASE} + 0.548}{\max(0.25, 0.424 \cdot DEF_B + 0.548)} + AVG_{AWAY}$$

$$AGA_A := (G_B - OFF_B)\frac{0.424 \cdot AVG_{BASE} + 0.548}{\max(0.25, 0.424 \cdot OFF_B + 0.548)} + AVG_{HOME}$$

where AVG_{HOME} is the average number of goals scored by home teams in our dataset, and AVG_{AWAY} the average goals scored by the visiting player. The formulas for $f_{A,B} = (OFF_A - AVG_{AWAY})\frac{\max(0.25, 0.424 \cdot DEF_B + 0.548)}{0.424 \cdot AVG_{BASE} + 0.548} + DEF_B$, $f_{B,A}, g_{A,B}$ and $g_{B,A}$ are modified accordingly.

Finally, in order to use this model in the same dataset as the Elo system, we use the same algorithm as in 4.1 replacing the unbiased estimator for ratings by 50 iterations of the update formulas A1 from the arbitrary starting point $OFF_i = DEF_i = 1$, which quickly converges to a fixed point.

 doi:10.20944/preprints202407.1974.v1

The only parameter of the model is the update sensibility λ in A1, which plays a similar role as K does for the Elo system, and we also introduce a dampening parameter D so that the result predictions are done with $D \cdot OFF_A, D \cdot DEF_A, D \cdot OFF_B$ and $D \cdot DEF_B$ instead of the original parameters. In our training dataset for soccer, we find the minimum MSE at the point $\lambda^* \approx 0.02$ and $D \approx 0.9$.

References

1. Aldous D., Elo Ratings and the Sports Model: A Neglected Topic in Applied Probability?. *Statistical Science* **2017**, *Vol. 32, No. 4*, 616—629.
2. Elo, A.E. *The Rating of Chessplayers, Past and Present*, 2nd ed.; Arco Publishing Inc.: New York, USA, 1986;
3. Glickman, M.E. Paired Comparison Models with Time-Varying Parameters. PhD, Harvard University, Cambridge, Massachussets, May 1993.
4. Steele, J.M. *Stochastic Calculus and Financial Applications*, 1st ed.; Springer-Verlag: New York, USA, 2010;
5. Jabin, P.E.; Junca, S. A Continuous Model For Ratings. *SIAM Journal on Applied Mathematics*, **2015**, *75 (2)*, 420–442.
6. Bondesson L. Generalized gamma convolutions and related classes of distributions and densities, Lecture Notes in Stat. 76, Springer, New York, 1992.
7. Cramér, H. Über eine Eigenschaft der normalen Verteilungsfunktion. *Math Z*, **1936**, *Vol 41* 405—414
8. Glickman, M.E.; Jones, A.C. Rating the Chess Rating System. *Chance* , **1999**, *12 (2)0* 21–28
9. Skellam, J. G. The frequency distribution of the difference between two Poisson variates belonging to different populations. *Journal of the Royal Statistical Society*, **1946**, *Vol. 109 Issue 3* 296
10. Ferri, C.; Hernandez-Orallo, J.; Modroiu, R. An Experimental Comparison of Performance Measures for Classification *Pattern Recognition Letters*, **2009**, *30* 27–38
11. Karlis. D.; Ntzoufras, I. Analysis of sports data by using bivariate Poisson models *The Statistician*, **2003**, *52 Part 3* 381-–393
12. How Our Club Soccer Predictions Work. Available online: https://fivethirtyeight.com/methodology/how-our-club-soccer-predictions-work/ (accessed on 14 July 2024).
13. A guide to ESPN's SPI ratings. Available online: https://www.espn.com/world-cup/story/_/id/4447078/ce/us/guide-espn-spi-ratings (accessed on 14 July 2024).