Elo Ratings and the Sports Model: a Neglected Topic in Applied Probability?

David Aldous

20 April 2017

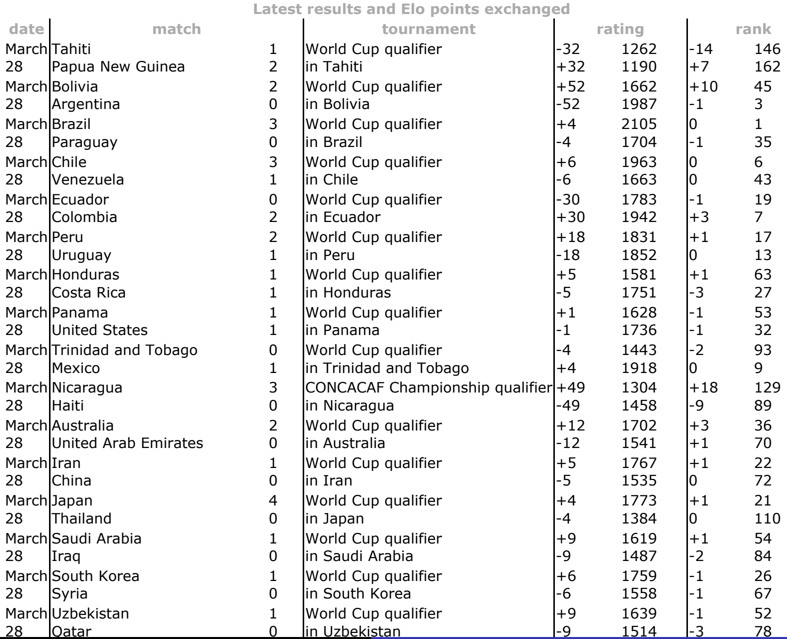
I occasionally teach a “Probability in the Real World” course, in which I give 20 lectures on maximally different topics, chosen with 3 *desiderata*.

- Not “textbook” topics taught in other courses.
- I have some“anchor data” to start the lecture.
- The topic is amenable to student projects.

Although mostly separate from my research activities, thinking about such topics sometimes uncovers research problems, and this talk is one such case.

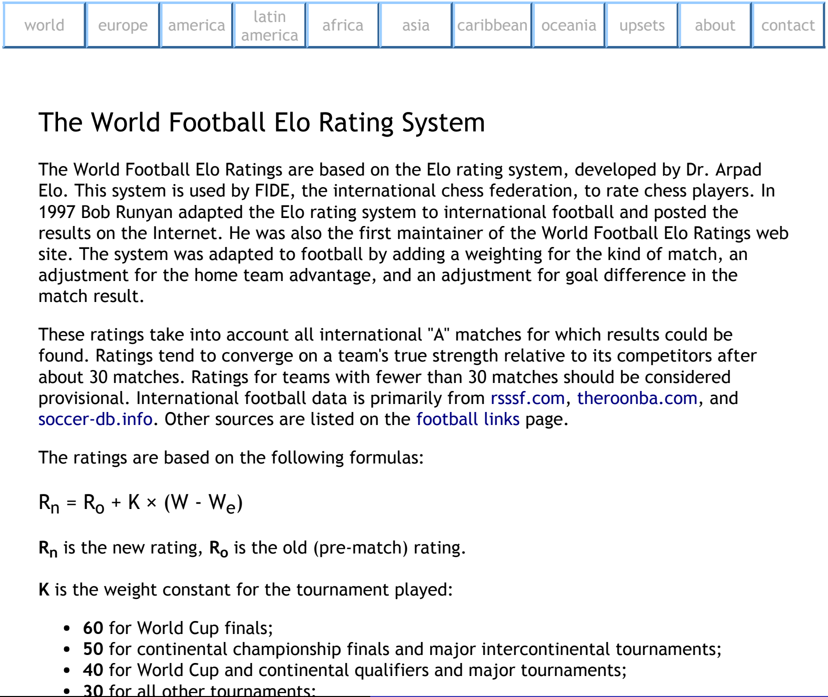
- Everyday perception of chance
- Ranking and rating
- Risk to individuals: perception and reality
- Luck
- A glimpse at probability research: spatial networks on random points
- Prediction markets, fair games and martingales
- Science fiction meets science
- Coincidences, near misses and one-in-a-million chances.
- Psychology of probability: predictable irrationality
- Mixing: physical randomness, the local uniformity principle and card shuffling
- Game theory
- The Kelly criterion for favorable games: stock market investing for individuals
- Toy models in population genetics: some mathematical aspects of evolution
- Size-biasing, regression effect and dust-to-dust phenomena
- Toy models of human interaction: use and abuse
- Short/Medium term predictions in politics and economics
- Tipping points and phase transitions
- Coding and entropy

World Football Elo Ratings



Latest results and Elo points exchanged

date	match		tournament	rating		rank	
March	Tahiti	1	World Cup qualifier	-32	1262	-14	146
28	Papua New Guinea	2	in Tahiti	+32	1190	+7	162
March	Bolivia	2	World Cup qualifier	+52	1662	+10	45
28	Argentina	0	in Bolivia	-52	1987	-1	3
March	Brazil	3	World Cup qualifier	+4	2105	0	1
28	Paraguay	0	in Brazil	-4	1704	-1	35
March	Chile	3	World Cup qualifier	+6	1963	0	6
28	Venezuela	1	in Chile	-6	1663	0	43
March	Ecuador	0	World Cup qualifier	-30	1783	-1	19
28	Colombia	2	in Ecuador	+30	1942	+3	7
March	Peru	2	World Cup qualifier	+18	1831	+1	17
28	Uruguay	1	in Peru	-18	1852	0	13
March	Honduras	1	World Cup qualifier	+5	1581	+1	63
28	Costa Rica	1	in Honduras	-5	1751	-3	27
March	Panama	1	World Cup qualifier	+1	1628	-1	53
28	United States	1	in Panama	-1	1736	-1	32
March	Trinidad and Tobago	0	World Cup qualifier	-4	1443	-2	93
28	Mexico	1	in Trinidad and Tobago	+4	1918	0	9
March	Nicaragua	3	CONCACAF Championship qualifier	+49	1304	+18	129
28	Haiti	0	in Nicaragua	-49	1458	-9	89
March	Australia	2	World Cup qualifier	+12	1702	+3	36
28	United Arab Emirates	0	in Australia	-12	1541	+1	70
March	Iran	1	World Cup qualifier	+5	1767	+1	22
28	China	0	in Iran	-5	1535	0	72
March	Japan	4	World Cup qualifier	+4	1773	+1	21
28	Thailand	0	in Japan	-4	1384	0	110
March	Saudi Arabia	1	World Cup qualifier	+9	1619	+1	54
28	Iraq	0	in Saudi Arabia	-9	1487	-2	84
March	South Korea	1	World Cup qualifier	+6	1759	-1	26
28	Syria	0	in South Korea	-6	1558	-1	67
March	Uzbekistan	1	World Cup qualifier	+9	1639	-1	52
28	Qatar	0	in Uzbekistan	-9	1514	-3	78



The World Football Elo Rating System

The World Football Elo Ratings are based on the Elo rating system, developed by Dr. Arpad Elo. This system is used by FIDE, the international chess federation, to rate chess players. In 1997 Bob Runyan adapted the Elo rating system to international football and posted the results on the Internet. He was also the first maintainer of the World Football Elo Ratings web site. The system was adapted to football by adding a weighting for the kind of match, an adjustment for the home team advantage, and an adjustment for goal difference in the match result.

These ratings take into account all international "A" matches for which results could be found. Ratings tend to converge on a team's true strength relative to its competitors after about 30 matches. Ratings for teams with fewer than 30 matches should be considered provisional. International football data is primarily from rsssf.com, theroonba.com, and soccer-db.info. Other sources are listed on the football links page.

The ratings are based on the following formulas:

$$R_n = R_o + K \times (W - W_e)$$

$\mathbf{R_n}$ is the new rating, $\mathbf{R_o}$ is the old (pre-match) rating.

K is the weight constant for the tournament played:

- **60** for World Cup finals;
- **50** for continental championship finals and major intercontinental tournaments;
- **40** for World Cup and continental qualifiers and major tournaments;
- **30** for all other tournaments;

Assertion **Ratings tend to converge on a team's true strength relative to its competitors after about 30 matches.**

By analogy a search on **seven shuffles suffice** gets you to discussions which can be tracked back to an actual theorem Bayer-Diaconis (1992).

Is there any theory or data behind this **thirty matches suffice** assertion? I haven't found any . . . . . .

History: Elo ratings originally used for chess, then for other individual games like tennis, now widely used in online games. I write "player" rather than team.

Idea 1: The basic probability model.

Each player A has some "strength" x_A, a real number. When players A and B play

$$\mathbb{P}(\text{A beats B}) = W(x_A - x_B)$$

for a specified "win probability function" W satisfying the (minimal natural?) conditions

$$\begin{gathered} W : \mathbb{R} \to (0,1) \text{ is continuous, strictly increasing} \\ W(-x) + W(x) = 1; \quad \lim_{x \to \infty} W(x) = 1. \end{gathered} \tag{1}$$

Implicit in this setup:

- each game has a definite winner (no ties);
- no home field advantage, though this is easily incorporated by making the win probability be of the form $W(x_A - x_B \pm \Delta)$;
- not considering more elaborate modeling of point difference;
- strengths do not change with time.

There is lots of discussion one could give in a long talk, but I'll stick to just three points.

- The default choice of W is $W(x) = e^x/(1+e^x)$, the logistic function. This case is equivalent to the Bradley - Terry model, widely used in statistics to give a "best fit" total ranking from inconsistent partial rankings (movie ratings etc).
- There is surprisingly little "applied probability" treatment. A 2016 Adler-Cao-Karp-Pekoz-Ross preprint *Random Knockout Tournaments* gives (for logistic and for randomly-matched tournament) upper and lower bounds for each player to win, in terms of the strengths (x_i).
- While the "unchanging strengths" assumption makes the math conceptually simple, and is appropriate for real-world tournaments, part of what makes spectator sports interesting is hope your player does better next year. The Elo scheme is implicitly designed to track changing strengths.

Easy to devise undergraduate projects based on the basic probability model. Here's an example.

What is the probability that the 2023 Australian Open Tennis Championships will be won by the second seed?

Our model says about 17%. Here we have historical data – over the last 50 years of the 4 annual Grand Slam tournaments the figure is 24% – but curious that we can get 17% from a model with no empirical data at all.

Idea 2: Elo-type rating systems

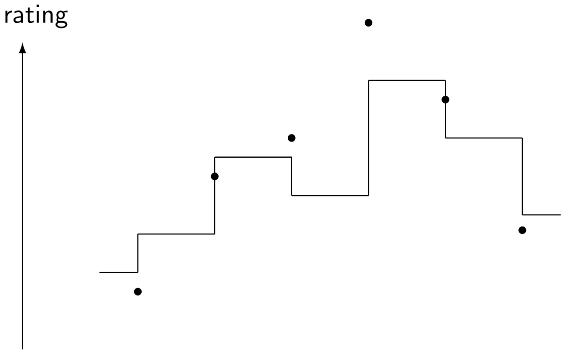
(not ELO). The particular type of rating systems we study are known loosely as Elo-type systems and were first used systematically in chess. The Wikipedia page *Elo rating system* is quite informative about the history and practical implementation. What we describe here is an abstracted "mathematically basic" form of such systems.

Each player i is given some initial rating, a real number y_i. When player i plays player j, the ratings of both players are updated using a function Υ (`Upsilon`)

$$\begin{aligned} \text{if } i \text{ beats } j \text{ then } y_i \to y_i + \Upsilon(y_i - y_j) \text{ and } y_j \to y_j - \Upsilon(y_i - y_j) \\ \text{if } i \text{ loses to } j \text{ then } y_i \to y_i - \Upsilon(y_j - y_i) \text{ and } y_j \to y_j + \Upsilon(y_j - y_i)\ . \end{aligned} \qquad (2)$$

Note that the sum of all ratings remains constant; it is mathematically natural to center so that this sum equals zero.

Schematic of one player's ratings after successive matches. The ● indicate each opponent's rating.



We require the function $\Upsilon(u), \ -\infty < u < \infty$ to satisfy the qualitative conditions

$$\Upsilon : \mathbb{R} \to (0,\infty) \text{ is continuous, strictly decreasing, and } \lim_{u\to\infty} \Upsilon(u) = 0. \tag{3}$$

We will also impose a quantitative condition

$$\kappa_\Upsilon := \sup_u |\Upsilon'(u)| < 1. \tag{4}$$

To motivate the latter condition, we want the functions

$$x \to x + \Upsilon(x-y) \text{ and } x \to x - \Upsilon(y-x)$$

the rating updates when a player with (variable) strength x plays a player of fixed strength y, to be an *increasing* function of the starting strength x.

Note that if Υ satisfies (3) then so does $c\Upsilon$ for any scaling factor $c > 0$. So given any Υ satisfying (3) with $\kappa_\Upsilon < \infty$ we can scale to make a function where (4) is satisfied.

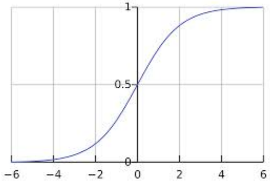
Recall that the logistic distribution function

$$L(x) := \frac{e^x}{1+e^x}, -\infty < x < \infty$$

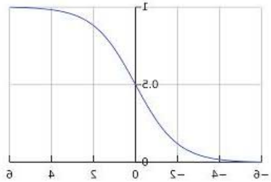
is a default choice for the "win probability" function $W(x)$ in the basic probability model; and its complement

$$1 - L(x) = L(-x) = \frac{1}{1+e^x}, -\infty < x < \infty$$

is a common choice for the "update function shape" $\Upsilon(x)$ in Elo-type rating systems. That is, one commonly uses $\Upsilon(x) = cL(-x)$.



possible $W(x)$



possible $\Upsilon(x)$

Whether this is more than a convenient choice is a central "foundational" issue in this topic.

Elo-type algorithms have nothing to do with probability, *a priori*. But there is an obvious heuristic connection between the probability model and the rating algorithm.

Consider n players with unchanging strengths $x_1, \ldots, x_n$, with match results according to the basic probability model with win probability function W, and ratings (y_i) given by the update rule with update function Υ. When player i plays player j, the expectation of the rating change for i equals

$$\Upsilon(y_i - y_j)W(x_i - x_j) - \Upsilon(y_j - y_i)W(x_j - x_i). \qquad (5)$$

So consider the case where the functions Υ and W are related by

$$\Upsilon(u)/\Upsilon(-u) = W(-u)/W(u), \quad -\infty < u < \infty.$$

In this case

> *(*) If it happens that the difference $y_i - y_j$ in ratings of two players playing a match equals the difference $x_i - x_j$ in strengths then the expectation of the change in rating difference equals zero*

whereas if unequal then (because Υ is decreasing) the expectation of $(y_i - y_j) - (x_i - x_j)$ is closer to zero after the match than before.

$$\Upsilon(u)/\Upsilon(-u) = W(-u)/W(u), \quad -\infty < u < \infty. \tag{6}$$

These observations suggest that, under relation (6), there will be a tendency for player i's rating y_i to move towards its strength x_i though there will always be random fluctuations from individual matches. So if we believe the basic probability model for some given W, then in a rating system we should use an Υ that satisfies (6).

Now "everybody (in this field) knows" this connection, but nowhere is it explained clearly and no-one seems to have thought it through (practitioners focus on fine-tuning to a particular sport). The first foundational question we might ask is

What is the solution of (6) for unknown Υ?

This can be viewed as the setup for a mathematician/physicist/statistician/data scientist joke.

Problem. For given W solve

$$\Upsilon(u)/\Upsilon(-u) = W(-u)/W(u), \quad -\infty < u < \infty.$$

Solution

- physicist (Elo): $\Upsilon(u) = cW(-u)$
- mathematician: $\Upsilon(u) = W(-u)\phi(u)$ for many symmetric $\phi(\cdot)$.
- statistician: $\Upsilon(u) = c\sqrt{W(-u)/W(u)}$ (variance-stabilizing ϕ).
- data scientist: well I have this deep learning algorithm . . . . . .

These answers are all "wrong" for different reasons. I don't have a good answer to "what Υ to use?" for given W. But the opposite question is easy: given Υ, there is a unique "implied win-probability function W" given by

$$W_\Upsilon = \frac{\Upsilon(-u)}{\Upsilon(-u) + \Upsilon(u)}$$

Conclusion: Using Elo with a particular Υ is conceptually equivalent to believing the basic probability model with

$$W_\Upsilon = \frac{\Upsilon(-u)}{\Upsilon(-u) + \Upsilon(u)}$$

Relating our math set-up to data

In published real-world data, ratings are integers, mostly in range $1000 - 2000$. Basically, 1 standard unit (for logistic) in our model corresponds to 174 rating points by convention. So the implied probabilities are of the form [football]

$$\mathbb{P}(\text{ Australia beats England }) = L((1701 - 1909)/174) = 0.23.$$

By convention a new player is given a 1500 rating. If players never departed, the average rating would stay at 1500. However, players leaving (and no re-centering) tends to make the average to drift upwards. This makes it hard to compare "expert" in different sports.

Is there any relevant non-elementary math probability?

Assume the basic probability model with non-changing strengths, and use Elo-type ratings – what happens? We need to specify how the matches are scheduled, use the mathematically simplest "random matching" scheme in which there are n players and for each match a pair of players is chosen uniformly at random. This gives a continuous-state Markov chain

$$\mathbf{Y}(t) = (Y_i(t), 1 \le i \le n), t = 0, 1, 2, \ldots$$

where $Y_i(t)$ is the rating of player i after a total of t matches have been played. We call this the *update process*. Note that this process is parametrized by the functions W and Υ, and by the vector $\mathbf{x} = (x_i, 1 \le i \le n)$ of player strengths. We center player strengths and rankings: $\sum_i x_i = 0$ and $\sum_i Y_i(0) = 0$.

The following convergence theorem is intuitively obvious; the technical point is that no further technical assumptions are needed for W, Υ.

Theorem

Under our standing assumptions (1, 3, 4) on W and Υ, for each $\mathbf{x}$ *the update process has a unique stationary distribution* $\mathbf{Y}(\infty)$*, and for any initial ratings* $\mathbf{y}(0)$ *we have* $\mathbf{Y}(t) \rightarrow_d \mathbf{Y}(\infty)$ *as* $t \rightarrow \infty$*.*

This is proved by standard methods – coupling and and Lyapounov functions. Note here we are not assuming the specific relation (6) between W and Υ. Note also that given non-random initial rankings $\mathbf{y}(0)$ the distribution of $\mathbf{Y}(t)$ has finite support for each t, so we cannot have convergence in variation distance, which is the familiar setting for Markov chains on $\mathbb{R}^n$ (Meyn -Tweedie text).

Alas these techniques do not give useful quantitative information about the stationary distribution. The theorem suggests a wide range of quantitative questions that we can't answer via theory.

To me the key question is

How well does Elo track changing strengths?

Too hard as theory – can only study via simulation. And it is not at all clear how to model changing strengths. I use several "qualitatively extreme" models.
For distribution of strengths use normal, $\sigma = 0.5$ or 1.0, which matches real data. For changes in strengths over time use

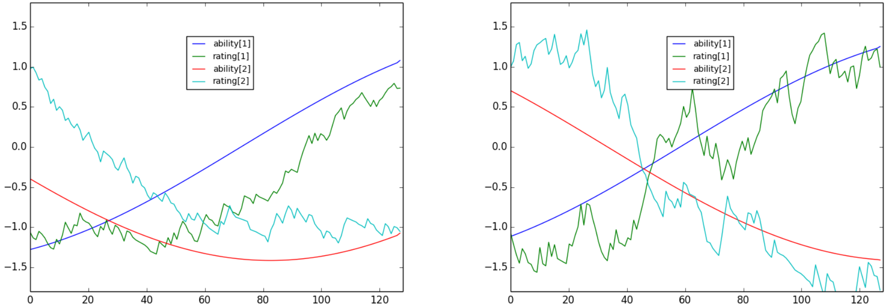
- Cyclic
- Ornstein-Uhlenbeck (ARMA)
- Hold (quite long time), jump to independent random.

each with a "relaxation time" parameter τ. For the win-probability function W and the update function Υ use

- Logistic
- Cauchy
- Linear over $[-1, 1]$

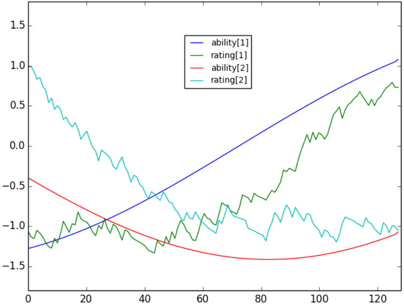
Use different scalings c for updates $c\Upsilon$.

Figure: Realizations of the cycle model: $\sigma = 1$, $\tau = 100$, logistic W and Υ.



$c = 0.17$ (left) and $c = 0.35$ (right).
This shows the intuitively obvious lag-bias versus noise effect.

In the setting above here is the optimal scaling $c = 0.26$



We need to quantify “ How well does Elo track changing strengths” in some way. Here’s my way.

Consider players A, B at a given time with actual strengths x_A, x_B and Elo ratings y_A, y_B. The actual probability (in our model) that A beats B is $W(x_A - x_B)$ whereas the probability estimated from Elo is $W(y_A - y_B)$. So we calculate

$$\text{root-mean-square of the differences } W(y_A - y_B) - W(x_A - x_B)$$

averaging over all players and all times. I call this **RMSE-p**, the “p” as a reminder we’re estimating probabilities, not strengths.

Key question: If willing to believe that (with appropriate choices) this is a reasonable model for real-world sports, what actual numerical values do we expect for RMSE-p?

Short answer; No plausible model gives RMSE-p much below 10%. This is when we are running models forever (i.e. stationary) so hard to reconcile with “30 matches suffice”.

Conceptually, there are 3 constituents of error

- **mismatch** between W and Υ.
- **lag** from changes in strengths in our past data
- **noise** from randomness of recent results.

Can do optimal trade-off between latter two by adjusting the scaling c in update $c\Upsilon$, to find the optimal c (given other parameters)

We can estimate the mismatch error from the deterministic limit in which $c \to 0$ for unchanging strengths.

Table: RMSE-p mismatch error.

W	logistic	logistic	Cauchy	Cauchy	linear	linear
Υ	linear	Cauchy	logistic	linear	logistic	Cauchy
$\sigma = 0.5$	0.9%	1.1%	1.2%	2.4%	3.2%	6.4%
$\sigma =$1.0	2.9%	2.9%	2.5%	5.4%	3.2%	6.0%

These errors are perhaps surprisingly small. Now let us take W and Υ as logistic, so no mismatch error. The next table shows the effect of changing the relaxation time τ of the strength change process.

Table: RMSE-p and (optimal c) for O-U model (top) and jump model (bottom)

	τ			
σ	50	100	200	400
0.5	12.9%	11.1%	9.5%	8.2 %
	(0.11)	(0.09)	(0.08)	(0.07)
1.0	17.0%	14.6 %	12.4%	10.4%
	(0.28)	(0.24)	(0.16)	(0.14)

	τ			
σ	50	100	200	400
0.5	12.8%	11.2%	9.8%	8.4 %
	(0.12)	(0.09)	(0.06)	(0.05)
1.0	16.9%	14.5 %	12.4%	10.5%
	0.30	0.24	0.19	0.14

This shows the intuitively obvious effect that for larger τ we can use smaller c and get better estimates. But curious that numerics in the two models are very close. One can do heuristics (and proofs if one really wanted to) for order-of-magnitude scalings as $c \downarrow 0$ but hardly relevant to real-world cases.

Bottom line from simulations: If you want RMSE-p to be noticeably less than 10% then you need to have played 400 matches and you need that strengths do not change substantially over 200 matches.

Games per year, regular season.

U.S. Football	16
Aussie Rules	22
U.K. Premier League	38
U.S. Basketball	82
U.S. Baseball	162

Slides and extended write-up on my web site.