Elo and Glicko-2 Rating Systems for 9-Ball Pool

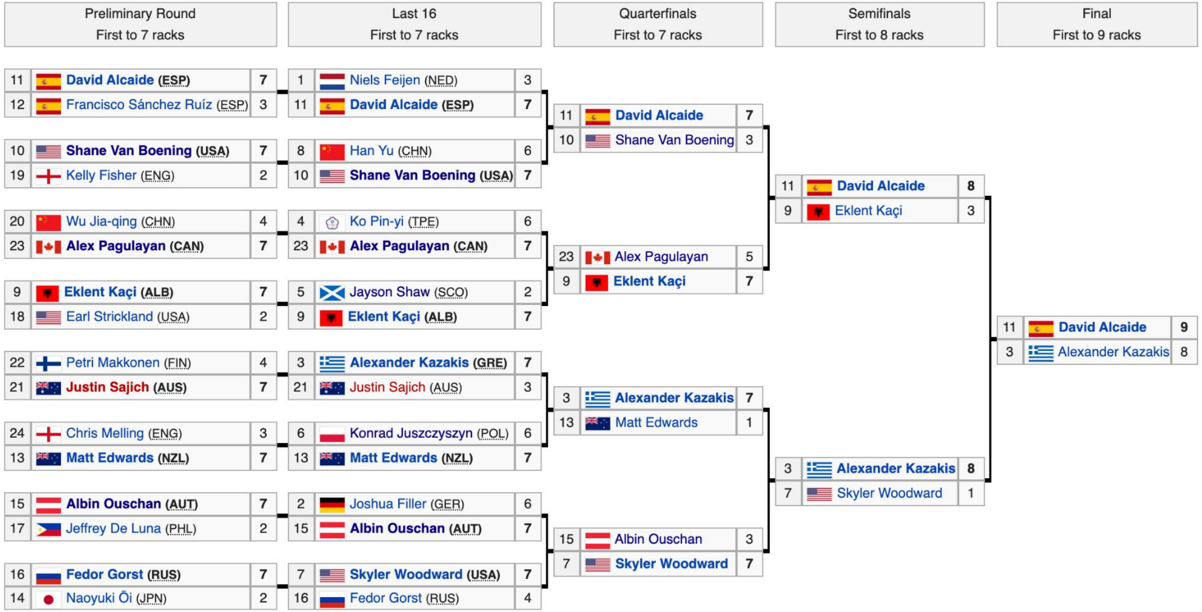
Seth Billiau, Sarah Lucioni,
Dasha Metropolitansky

Background

9-Ball is the most popular game in professional pool.

Matches consist of two players in a “race to **X**” format: the first player to win **X** games of 9-ball wins the match.

Tournaments are typically played in a bracket format with round-robin for seeding.



There are two existing rating systems:

The existing rating systems are imperfect…

Organizational Ratings	FargoRate
Organizations use their events only	Private algorithm and data
Conflicts between rating systems	Potentially too broad in scope
No probabilistic interpretation	No measures of uncertainty

We want to:

1. Create **Elo** and **Glicko-2** rating systems for 9-Ball pool
2. Compare our models with existing rating systems
3. Predict the outcome of a 2021 tournament

Data

We used 2 sets of match-level data:

Training Data	Validation Data
94 tournaments from 2007-2020	2021 Predator CLP tournament
5,085 matches (72,530 racks)	192 matches (1,418 racks)
1,248 unique players	19 unique players

Elo

Elo is a dynamically updating, relative strength rating system for players in zero-sum games.

This means that players' ratings depend on their opponents' ratings and their game results.

At a high level, here's how Elo ratings are updated:

Win	Loss	Update
Higher-rated player	Lower-rated player	A **few** points taken from the **lower-rated** player
Lower-rated player	Higher-rated player	**Many** points taken from the **higher-rated** player
Draw	Draw	A **few** points taken from the **higher-rated** player

Here's the math you need to know.

1. **Expected scores** for players A and B with ratings R_A and R_B:

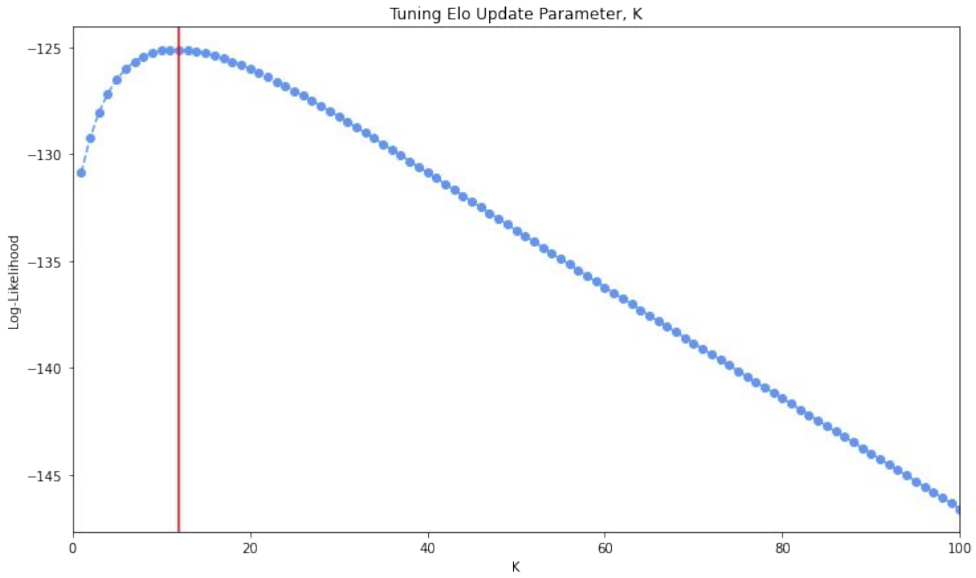
$$E_A = \frac{1}{1 + 10^{(R_B - R_A)/400}} \qquad E_B = \frac{1}{1 + 10^{(R_A - R_B)/400}}$$

2. **Update rule** if player A truly scored S_A points:

$$R'_A = R_A + K(S_A - E_A)$$

K is the maximum possible rating adjustment per game.

To tune K, we find the value of K that maximizes the log-likelihood over the validation set.



We calibrated several types of Elo systems:

Adjustment	Why
None	Baseline model to compare adjusted Elo methods
Margin of Victory	Account for score differential
Racks as Matches	Inflate the amount of data
Initial Rankings	Consider the structure of tournament data

Glicko-2

Elo: player ratings

Glicko: player ratings **+ rating deviations**

Elo: player ratings

Glicko: player ratings + rating deviations

Glicko-2: player ratings + rating deviations **+ rating volatilities**

We need to set 2 main parameters:

Change in Volatility Constraint (τ):

Ranges from 0.3-1.2, where smaller values prevent large changes in ratings based on very unlikely results.

Rating Period:

A collection of games that are treated as having occurred simultaneously. Updated ratings are computed at the end of each rating period.

Ideally, we want an average of 10-15 matches per player in a rating period, but...

Year	Avg Matches Per Player
2007	1.40
2008	1.48
2009	1.46
2010	3.74
2011	2.63
2012	2.65
2013	2.68
2014	2.56
2015	3.71
2016	4.44
2017	4.52
2018	5.01
2019	5.32
2020	4.39

Here we're treating all players with 1 match as one player, all players with 2 matches as another player, etc. up until 20 matches.

Year	Avg Matches Per Player
2007	1.56
2008	1.79
2009	1.90
2010	9.00
2011	3.77
2012	3.75
2013	3.37
2014	3.11
2015	6.44
2016	7.53
2017	12.73
2018	11.80
2019	29.28
2020	16.03

This is closer to 10-15 games per player in each rating period but only in certain years...

Year	Avg Matches Per Player
2007	1.56
2008	1.79
2009	1.90
2010	9.00
2011	3.77
2012	3.75
2013	3.37
2014	3.11
2015	6.44
2016	7.53
2017	12.73
2018	11.80
2019	29.28
2020	16.03

We find that the model with the largest log-likelihood on our validation set has the following parameters:

τ = 0.3

Rating period is each year from 2016-2020

Minimum games is 1 (i.e. no player aggregation)

Results

Method	Parameters	Log-Likelihood
Initial Rankings (Elo)	**K = 7**	**-123.4**
Margin of Victory (Elo)	K = 8	-124.7
Baseline (Elo)	K = 12	-125.1
Racks as Matches (Elo)	K = 5	-126.4
FargoRate (Elo)	N/A	-128.2
Glicko-2	$\tau = 0.3$ rating periods = 2016-20	-292.95

As a sanity check, let's look at the relative ratings of the top WPA-rated players...

Top 3

Bottom 3

Player	WPA	Fargo Rate	Baseline Elo	MOV Elo	Racks Elo	IR Elo	Glicko-2
Jayson Shaw	8	3	1	2	2	1	1
Joshua Filler	2	2	2	1	1	2	2
Shane Van Boening	3	1	4	4	7	3	3
Niels Feijen	10	8	3	3	5	4	4
Fedor Gorst	4	6	5	5	3	5	6
Alex Kazakis	9	10	6	6	4	7	8
Ko Ping-Chung	1	7	7	8	8	6	7
Chang Jung-Lin	5	4	9	9	6	9	5
Ko Pin Yi	6	5	8	7	10	8	10
Alex Pagulayan	7	9	10	10	9	10	9

Finalist
Group Winner
Top 3
Bottom 3

Player	WPA	Fargo Rate	Baseline Elo	MOV Elo	Racks Elo	IR Elo	Glicko-2
Eklent Kaci	5	1	1	1	1	1	1
Albin Ouschan	9	3	3	3	3	3	2
Niels Feijen	2	2	2	2	5	2	3
Denis Grabe	11	8	4	6	7	5	5
Ralf Souquet	3	7	5	7	11	6	4
David Alcaide	6	6	6	4	6	4	6
Mieszko Fortuński	14	9	8	8	2	8	8
Alexander Kazakis	1	5	7	5	4	7	7
Marc Bijsterbosch	16	15	10	11	14	10	9
Casper Matikainen	12	14	12	13	13	13	10
Billy Thorpe	4	13	14	15	12	14	11
Naoyuki Oi	7	4	11	10	9	11	12
Darren Appleton	15	12	9	9	10	9	13
Roberto Gomez	18	11	16	14	8	15	14
Chris Melling	10	10	13	12	15	12	15
Jasmin Ouschan	13	18	15	16	17	16	16
Kelly Fisher	8	16	18	17	18	17	17
Chris Robinson	19	17	19	18	16	18	18
Kristina Tkach	17	19	17	19	19	19	19

...and the 19 players in the Predator CLP

Drawbacks

There are several drawbacks to our approach.

Data issues	Other covariates
Survival bias	“Break” Information
Womens tournaments	All races are treated equally
Number of matches per Glicko rating period	Player-level covariates (e.g. experience, age, etc.)
Data leakage in best Elo model	Exogenous covariates (e.g. travel time, table type, etc.)

Thank You!

Possible extensions include:

Elo	Glicko-2
Change K based on round	Explore Glicko-2 at the rack-level
Set initial ratings based on player-level covariates	Fit Glicko instead of Glicko-2 for additional comparison