University College Dublin

School of Mathematics and Statistics

STAT40710: Dissertation

Student: Gavin Tynan

Supervisor: Gabrielle Kelly

Date: 14/8/2020

Table of Contents

Title

Rethinking the ranking of women's tennis.

Abstract

Throughout this dissertation we explore problems with the Women’s Tennis Association ranking model, explain its importance within the game, the advantages generated by being highly ranked and propose an improved and fairer ELO model. Starting with the current best tennis prediction model publicly available we explore extensions such as including matches from lower tiered events, considering a player’s performance in a game not just the outcome, including a players past pedigree to prevent over fitting to form and proactively predicting upsets when a player is outperforming their current predicted ability. The final model contains 10 parameters which are shared by players and use to generate their abilities over time. This model drastically outperforms the current ranking model in predictive performance and is competitive with bookmaker’s models.

1 - Introduction

Tournament seeding is common practice across most professional sports. In essence, its main purpose is to increase the probability that later rounds are more competitive than earlier rounds and that, by the time a tournament reaches its finale, spectators will be treated to the most competitive and thrilling matchup. Simply put, it segregates the best players from each other early in a tournament in order to increase the chances of them meeting in later rounds.

In America's top 3 sports leagues - the NBA, NFL, and MLB - each season is broken up into two parts; the regular season and the postseason. The regular season's sole purpose is to generate seedings for the postseason. This is done by allowing the regular season to function as a league whereby all teams play the same number of games against similar opposition. The best performers are then entered into the tournament-based postseason and the draw is seeded to reflect each team's regular-season performance.

Tennis is unique in that its entire season is tournament based and thus at no point is there a level playing field whereby seeds are earned. Instead, seeding is determined based on ranking points, and ranking points are earned based on performances at seeded tournaments. This format creates a recursive systematic bias towards seeded players whereby being seeded increases your chances of performing well in a tournament, which in turn increases your ranking points which increases your chances of being seeded for the next tournament. If tennis seedings aren't earned on a level playing field like in other sports it is imperative they are as fair as possible such that the advantage generated by being seeded is warranted.

The main aim of this dissertation is to establish a fairer basis for women's tennis rankings than the current system. Women's tennis has been chosen specifically as most of the literature review with respect to tennis focused on the Men's game. Women's tennis is also more interesting in that the top players are not quite as static as the men's, where Roger Federer, Andy Murray, Rafael Nadal and Novak Djokovic have dominated for so many years. While Serena Williams has been a constant, her supporting cast has been anything but. With Williams at the age of 38 and with the emergence of young players like Ashleigh Barty, Naomi Osaka, and Bianca Andreescu, it would appear that a changing of the guard is taking place.

The main challenge in constructing a fair ranking, is that the final model must be fully explainable such that a player knows how many ranking points are on the line in a given match and also can understand how their current ranking was generated. This explainability constraint leads to the challenge of not being able to include information that is outside of a player's control in the model such as betting odds.

1.1 - Women's Tennis Association ranking

Before considering improvements to the tennis ranking system, one must first understand how the system works. The Women's Tennis Association (WTA) ranking points are calculated on a rolling 52-week cumulative basis (Rankings, 2020). Ranking points are awarded at each round of each major tournament.

A player's total ranking points at any given time are based on their best performances at 16 tournaments within the previous 52-week window. Within the 16 tournaments the following tournaments must be included:

- 4 Grand Slams

- 4 Premier Mandatory
- Best 2 performances at Premier 5 events

This leaves 6 tournaments up to a player's discretion. To appear on the WTA rankings, a player must earn a minimum 10 points at a single tournament or participate in at least 3 tournaments within the rolling window.

The WTA rankings are then used to determine the player seedings for all tournaments. This is simply done by assigning a seed to the top X players at the tournament. Each tournament seeds a different number of players and thus X varies depending on the tournament. Grand Slams reserve the right to alter seedings slightly should they so wish with Wimbledon often adjusting theirs taking into account a player's past performances on grass.

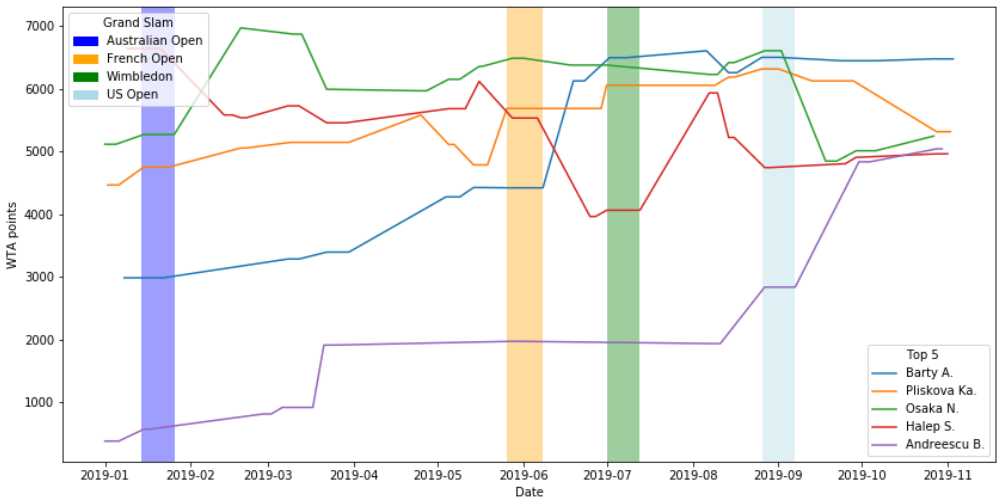

Figure 1: WTA year-end Top 5 Players - ranking points over the 2019 season

Figure 1 illustrates the main issue with ranking points which is that they don't capture form/current ability. This can clearly be seen following Simona Halep's trend (red). The stark decrease in her trend following the French Open is due to her exiting in the quarter-final; to most this would be a respectable finish but, because she won the Open the previous year, her win must be removed from her ranking points before her new points are added. While penalising her for underperformance relative to her past performances makes some sense, it should not be taken as indicative of current ability relative to other players on tour; this is demonstrated the following month when she went on to win the most prestigious Grand Slam, Wimbledon.

In the current system, ranking points are a reflection of cumulative ability and thus it makes no sense to use these points to seed tournaments when the aim of seeding is to increase the chances of the current best players facing off in later rounds. If ranking points do not reflect current ability, then they should not be used for tournament seeding. It is thus necessary to examine what other models could be used to indicate the best current players.

2 - Background models

2.1 - Bradley Terry

The Bradley-Terry model (Bradley and Terry, 1952) was designed for the analysis of experiments involving paired comparisons. The fundamental concept proposed was that when two items are being compared in a ranking experiment a test of no difference can be performed based on the binomial distribution. Thus, a probability over which item will be preferred can be estimated and these probabilities can be used to form an overall ranking to compare multiple items.

The model takes the form of a generalized linear regression with a logit link function and no intercept:

Let β_i = player i, π_{ij} = the probability player i is preferred to player j

$$\log(\frac{\pi_{ij}}{\pi_{ji}}) = \beta_i - \beta_j \qquad (1)$$

The model's input data is organized such that each row represents a paired comparison with player i denoted with the value 1, player j denoted with -1 and the outcome is a binary representation with respect to player i:

Let X = paired comparisons, y = the outcome of the comparisons with respect to player i.

$$X = \begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & -1 & 1 & 0 \end{bmatrix} \qquad y = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

Above shows that in the first paired comparison (row 1) the player in column 1 (player i) was preferred to player j (column 4) and in the second paired comparison (row 2) player j (column 2) was preferred to player i (column 3).

The model is fitted using a maximum likelihood estimate and the number of parameters to be estimated is equal to the number of items to be ranked in the experiment (the number of columns in the X matrix). Thus the coefficient fitted to each item (β_i) can be seen as a relative strength measure in comparison to the other items, meaning the item with the largest fitted coefficient is seen as the top-ranked item and these coefficients can be used to generate probabilities of preference when comparing two items:

$$\pi_{ij} = \frac{\exp(\beta_i)}{\exp(\beta_i) + \exp(\beta_j)} \qquad (2)$$

One of the issues with the Bradley-Terry model in the context of tennis rankings is that the parameter estimates are static. This means that once the model is fit to the data, the strength estimates of each player would not change and could not be updated without refitting the entire model to the updated data. Another problem when considering long term ranking is that because parameter estimates are based on paired comparisons, a parameter must be estimated for each player involved in any paired comparison being considered. This means that within the period being ranked, any player that played a match must have a parameter estimated, leading to a scenario whereby the number of parameters to be fitted is constantly growing with many non-significant estimates due to many players having appeared only a few times.

2.2 - Dynamic Bradley Terry

The Dynamic Bradley-Terry model (Cattelan, Varin, and Firth, 2012) attempted to solve the Bradley-Terry Model's static and constantly growing parameter problem. They hypothesized that in a sporting matchup, a team's ability at time t was based on their previous performances up to time t and thus is assumed to evolve over time as their form (past performances) varies.

They also hypothesized that parameters used to estimate form were not unique to each team and that instead common parameters could be estimated and used to generate each team's ability at any given time. They decided to treat form as an exponentially weighted moving average process (EWMA). This creates a scenario whereby only two parameters need to be estimated irrespective of the number of participants: a smoothing parameter for the EWMA and a coefficient to represent max ability. The smoothing parameter determines the importance of newer results; a high smoothing parameter has a recency bias and a low parameter is slower to adapt to new data. The max ability coefficient then interacts with the form at time t to give each team an ability, meaning depending on their form each team is performing at some percentage of the theoretical max.

While this paper offers some interesting proposals, it also doesn't quite work in the context of tennis ranking. Both of the examples used within the paper for soccer and basketball were used in the context of a league: in a league, every team plays the same number of games and, over the course of a season, plays against similar opposition. This is not true in tennis where each player's schedule is completely different, and the quality of opposing players is also highly variable. Thus, when comparing performance in tennis, the quality of opposition must be considered otherwise it will be easier for players to inflate their rankings by choosing tournaments with weaker opposition.

2.3 - High dimensional dynamic model

The High Dimensional Dynamic Model (Gorgi, Koopman and Lit, 2019) continued the exploration of a dynamic model based on the Bradley Terry Model but specifically with regards to tennis. Each player's ability now varies depending on time and over time a player's ability is treated as a random walk.

Before a game is played each player is assigned a starting ability:

Let $\beta_{i,0}$ = player i's ability at time 0, α = the weight assigned to the log of their WTA points

$$\beta_{i,0} = \log(\text{WTA points}) \times \alpha \qquad (3)$$

Unlike previously in the Dynamic Bradley Terry Model, this random walk considers not just the outcome of the match but also the prior likelihood associated with the match. This means before a match, the difference between the two players' abilities are used to create a prior probability for player i beating player j (equation 2).

Following the match, this probability is used to determine the extent to which both players' abilities changed depending on the outcome. The difference between the prior probability and the outcome is known as the score ($s_{i,t}$):

Let $y_{ij,t}$ = the binary match outcome at time t with respect to player i against player j [0, 1].

$$s_{i,t} = y_{ij,t}(1 - \pi_{ij,t}) - (1 - y_{ij,t})\pi_{ij,t}, \qquad s_{j,t} = -s_{i,t} \qquad (4)$$

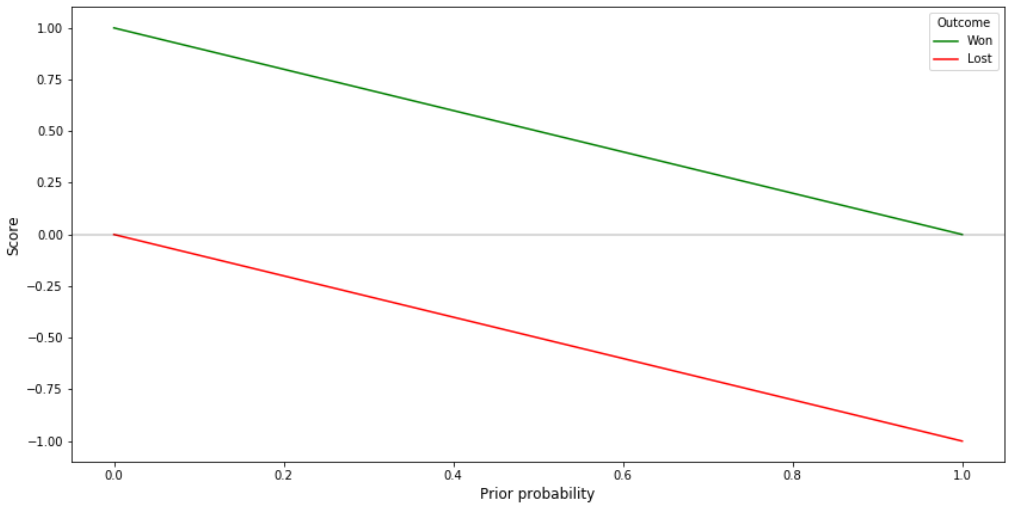

Figure 2: Score function (equation 4).

The essence of the scoring function is that a player is given the probability of the outcome that did not occur. The winning player is given a positive score equal to the probability of them losing and the losing player is given a negative score equal to the probability of them winning. This rewards players less for winning games in which they were expected to win but rather penalises them more should they underachieve and lose, which is imperative to consider in a system like a tennis where schedules are not equal.

The score is then multiplied by a constant τ and added to a player's old ability to form their new ability:

$$\beta_{i,t+1} = \beta_{i,t} + \tau s_{i,t} \qquad (5)$$

The paper also shows extensions to this model considering other factors such as surface type, age and home-field advantages which are all also interesting in the context of tennis.

3 - Data

The dataset was obtained from https://github.com/JeffSackmann/tennis_wta. The years 2010 → 2017 were used to fit the proposed models with 2018 → 2019 used for model evaluation. Following the data cleaning steps the training dataset had 22,274 (161,427 including ITF) matches and the test dataset had 5,958 (48,998 including ITF).

3.1 - Cleaning

The dataset provided did not have individual dates for each match but instead an overall tournament start date, meaning all matches within a specific tournament shared the same date. As the models proposed treat player strength as time-varying, individual dates for each match must be created.

To overcome this, tournament rounds were treated as ordered, categorical variables and each match date was increased relative to their round i.e. matches played in round 1 maintained the tournament start date with those in round 2 getting a date = tournament start date + 1. This workaround unfortunately meant round-robins all received the same date, meaning information could not be passed forward in round-robins thus players' ability only changed after round-robins, not during.

Match score was given as a string with tiebreak points bracketed i.e. '6-0 7-5(6)'. If a player retired, it would give the match score up to that point followed by 'RET' and a walkover was denoted as 'W/O'. It was decided not to include walkovers or retirements in the dataset as they were not reflective of performance.

3.2 - Feature engineering

3.2.1 - Score

Score's string data was transformed into set and game info. This meant extracting the number of games (games are played within a set i.e. 6-0 means 6 games to the winner 0 to the loser) won and lost by the winning player as well as the number of sets.

3.2.2 - Bookmakers winning probability

Like most sports, tennis has a betting market whereby a variety of bookmakers offer odds on matches allowing people to bet. Bookmakers' odds are seen somewhat as the holy grail of predictions due to it being extremely difficult for people to outperform them. The combination of state-of-the-art models and adjusting odds to the wisdom of the crowd as people place more bets makes their odds generally the best in class.

A second dataset to get bookmaker odds was used from http://www.tennis-data.co.uk/alldata.php. In total, this dataset had 4,941 WTA games between 2018 and 2019, containing the bookmaker Bet365's odds for each match in decimal form i.e. 1.91.

Decimal odds were converted to percentages by dividing 1 by the odds i.e. 1 / 1.91 = 0.524. The two percentages (the winning and losing percentages) were then summed together to get the "book percentage" which is always greater than 100% to include the bookmakers' margin. The winning percentages were then divided by the book percentage to get the normalised winning probability which removed the bookmakers' margin.

Of the 4,941 games in the new dataset only 4,337 could be used with our test dataset due to a variety of formatting issues.

4 - Theory and methods

4.1 - ELO - FiveThirtyEight

FiveThirtyEight (538) is an American blog that focuses on various statistical analysis across politics, economics and sports. To date their ELO implementation (Morris and Bialik, 2015) is the best performing tennis prediction model publicly available (Kovalchik, 2016). Further background research yielded no improvements on this model to date, with the exception of 538's own newer model to include playing surfaces.

4.1.1 - Basic model

Each player is given a starting ability at time 0 where i refers to player i and there are N players in total:

$$\beta_{i,0} = 1500 \qquad i = 1,2,..,N \qquad (6)$$

1,500 is an arbitrary number, the key being that before a match is observed all players start at the same ability. Similar to the High Dimensional Model, before each match a prior probability reflecting the expected outcome is generated:

Let $\pi_{ij,t}$ = the probability of player i beating player j at time t.

$$\pi_{ij,t} = 1 - \frac{1}{1 + 10^{\frac{\beta_{i,t} - \beta_{j,t}}{400}}} \qquad (7)$$

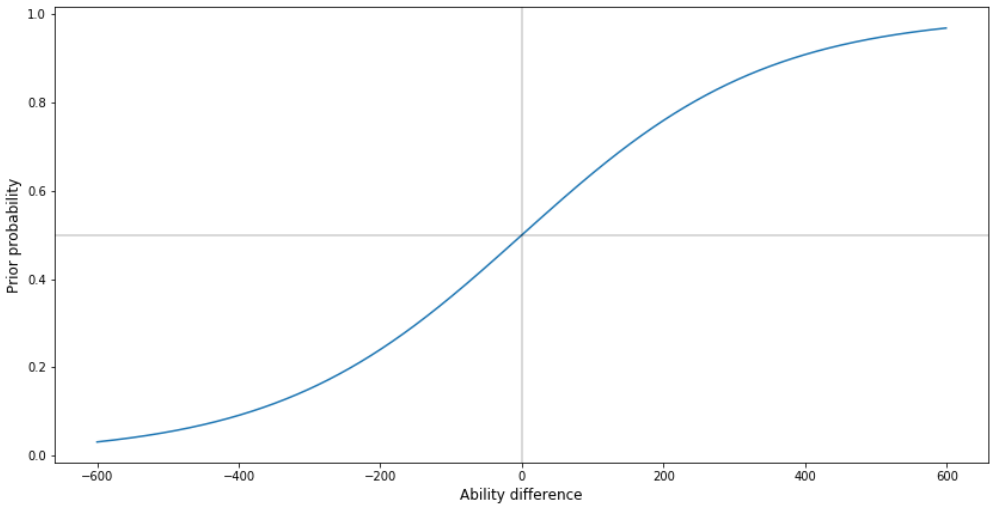

Figure 3: Prior probability sigmoid curve (equation 7).

After the match has concluded, the prior probability is used to create a score (equation 4) for both players to reflect their performance.

Like the Giorgi et al model, the magnitude the scoring function has on a player's new ability is determined by the τ function (refereed to as the K factor in ELO literature). Unlike previously, 538 treats τ as dynamic, varying depending on how many games a player has

played previously. The idea is that when a player is new their τ is high in an effort to get them to their true ability as quickly as possible and it decreases as we become more certain of their ability.

$$\tau = \frac{K}{(no.\ matches + offset)^{shape}} \qquad (8)$$

The offset parameter is designed to ensure new players don't get too high an initial τ with the shape ensuring τ doesn't become too small as players play more matches. 538 propose the following parameters K = 250, offset = 5, shape = 0.4 (Figure 4).

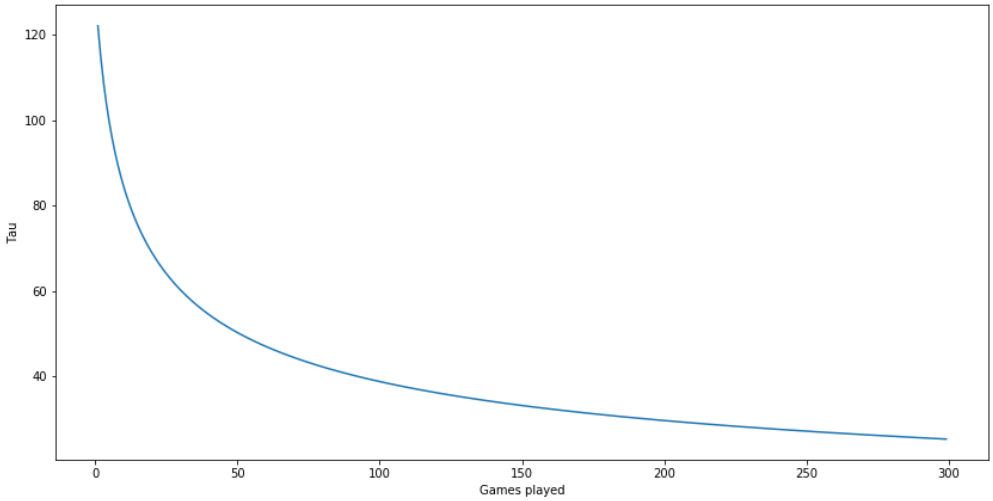

Figure 4: Tau function (equation 8) fitted to 538's proposed parameters.

Each player's ability is then updated for the next time period using equation 5.

4.1.2 - Surface model

Following on from the success of their initial ELO Model, 538 released an updated model (Morris, Bialik and Boice, 2016) to include surface effects. The hypothesis was that a certain player may be stronger or weaker on one surface than another (think Rafael Nadal's 12 Grand Slams on clay compared to 2 on grass).

Previously each player had one numeric value reflecting their ability which was used irrespective of surface. This new model proposes maintaining 4 numeric values for each player, one for each surface (hard, clay, grass) and an overall value:

Let $\beta_{i,t}^{o}$ denote overall ability and $\beta_{i,t}^{s}$ denote surface specific ability

$$\beta_{i,0}^{o} = 1500$$

$$\beta_{i,0}^{s} = [1500,1500,1500] \qquad (9)$$

Prior to each match on a particular surface a player's current ability is a weighted (η) version of their overall and surface-specific ability.

An aspect regarding the implementation of this new model, which was unclear from Morris, Bialik and Boices's write-up is the surface weight. It is mentioned $\eta = 0.29$ when referring to hard surfaces but the authors fail to clarify if η is surface-agnostic and thus constant, or if η is a vector of 3 values, a separate weight for each surface.

Let η = the surface weight, I^s = surface indicator (1 for the playing surface else 0 i.e. [1 (hard), 0 (clay), 0 (grass)] for hard)

If η is a constant:

$$\beta_{i,t} = (1-\eta)\beta^o_{i,t} + \eta \sum_{s\in[h,c,g]} I^s_{i,t}\,\beta^s_{i,t} \qquad (10)$$

If η is a vector:

$$\beta_{i,t} = \sum_{s\in[h,c,g]} (1-\eta_s) I^s_{i,t}\beta^o_{i,t} + \eta_s I^s_{i,t}\beta^s_{i,t} \qquad (11)$$

The weighted ability $\beta_{i,t}$ is used as before using equation 7 to generate a prior probability, equation 4 to calculate a score to reflect performance and equation 8 to establish a τ value. Equation 5 however is now performed twice:

$$\beta^o_{i,t+1} = \beta^o_{i,t} + \tau s_{i,t}$$

$$\beta^s_{i,t+1} = \beta^s_{i,t} + \tau s_{i,t} \qquad (12)$$

4.2 - Improvements

4.2.1 - Model fit

For the purpose of this research all models discussed were refit to the specific train data using the following maximum likelihood equation to minimise prediction error:

$$L(v) = \sum_t \sum_{i,j} \log(y_{ij,t}(\pi_{ij,t}) + (1-y_{ij,t})(1-\pi_{ij,t})) \qquad (13)$$

With v being a vector containing the parameters to be estimated, in the case of the basic model for example (section 4.1.1) these are K, offset and shape.

All models were fit using the minimise function in Python's scipy package (we therefore took the negative of equation 13). This function did not return stable standard error estimates and thus models were compared to one another using a log likelihood ratio test:

Let r = the number of additional params, L_0 = the simpler model log likelihood and L_1 = the more complex log likelihood

$$-2(L_0 - L_1) > \chi^2_{0.05}(r) \qquad (14)$$

4.2.2 - International Tennis Federation

When comparing both the basic and surface models' predictions to the bookies' odds it became apparent our model may be missing data the bookies had. When both players had

played numerous previous games our predictions generally resembled the bookies' odds; however, when a player was new our model struggled while the bookies continued to excel.

The International Tennis Federation (ITF) is the circuit below the WTA. Top players generally stick exclusively to the WTA due to the increased competition, prize money and exposure however lower ranked players will often float between both, playing WTA tournaments when they qualify and ITF events in the meantime. By not observing this data, we are missing crucial information about players' performances while they were not on the main tour.

Let I^c= circuit indicator (1 when ITF else 0), ρ = ITF weight deduction for τ (0,1)

$$\tau = \frac{K}{(no.\ matches + offset)^{shape}}(1 - I^c_{i,t}\rho) \qquad (15)$$

Equation 15 gives ITF games a lower τ than WTA to ensure the top ITF players are not overestimated compared to WTA players.

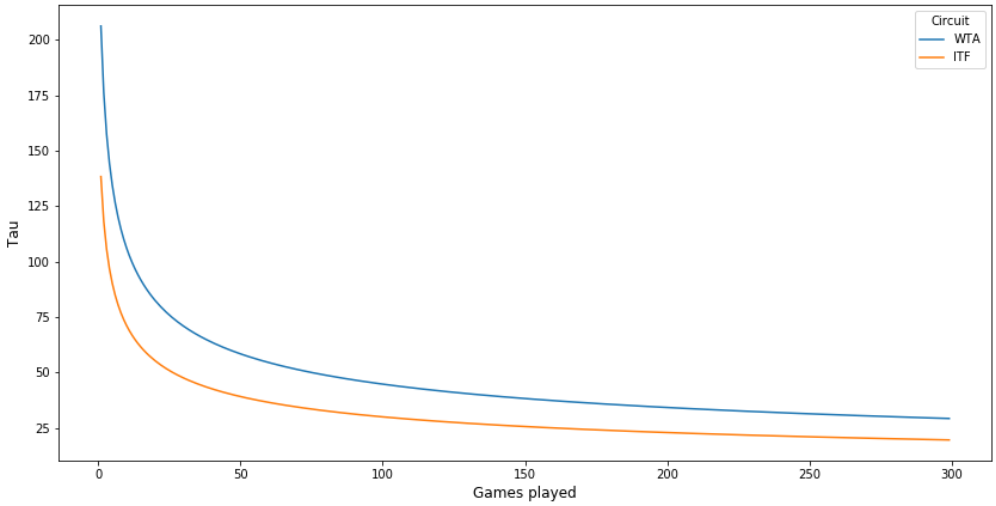

Figure 5: Tau function (equation 15) fitted to model 4's MLE.

The purpose of this dissertation remains focused on WTA matches, thus ITF games were not considered in the likelihood calculation ensuring maximum likelihood estimations remained comparable across models.

4.2.3 - Performance

When one thinks of great players, they rarely focus solely on the outcomes but instead on specific performances. Rafael Nadal fans don't just cite his 2008 French open win as proof of his all-time great status, but instead focus on the manner in which he won the final: 6–1, 6–3, 6–0 against the then world number 1 and top seed, Roger Federer. Similarly, for Simona Halep in the 2019 Wimbledon final; facing Serena Williams who was on course to break the women's record for total number of Grand Slams won, she defeated her 6-2, 6-2 in under 1 hour. It feels wrong that under the current implementation, no matter the manner in which Nadal or Halep won their respective finals, their ELO change would be the same.

When describing the implementation for their surface-specific ELO, 538 make reference to their reasoning for not including score/performance to date: "But the question when switching to a more granular metric than wins and losses is whether you gain more in detail than you lose in accuracy. So far, in this case, switching to sets or games has tested out to be unhelpful predictively" also mentioning, "We could change this if we find a more accurate method". To date they have not released a revised version of their ELO to incorporate score.

One of the problems with incorporating match score in an ELO model is how close the score in tennis tends to be. Jeff Sackman writes about this in more detail (Sackman, 2019) where he explains how winning 55% of the points is in-fact a lot less close than it may sound. He also proposes instead to think about match scores as though from a binomial distribution where our assumption is that two players are equal, and we then use then cumulative binomial distribution to calculate the probability of winning the same amount or less points than the winner did. This is then thought of as the probability the winning player was superior on the day.

Let Bin = cumulative binomial distribution, x = the number of games won by the winner, n = the total number of games played, p = the probability the winner wins a game, I^{ss} = straight sets indicator, ss = straight sets coefficient (acts as a performance boost)

$$wP = Bin(x; n, p) + I^{ss} ss$$

$$lP = 1 - wP \qquad (16)$$

We propose using this probability (equation 16) as a dynamic ceiling, this being instead of rewarding a winning player with the probability of them losing we reward them with their performance minus the probability of them winning. The same is also true for the losing player - their score equaling their performance minus the probability of them losing. This change rewards players for outperforming their prior probability irrespective of whether they win or lose. To the best of our knowledge this approach has not been documented elsewhere.

$$s_{i,t} = y_{ij,t}(wP - \pi_{ij,t}) + (1 - y_{ij,t})(lP - (1 - \pi_{ij,t})) \qquad (17)$$

4.2.4 - Class

There's a saying in sport, "form is temporary, class is permanent" - the idea being sometimes great players don't play so great. 2016 was Angelique Kerber's year. An Australian Open, a US Open, Wimbledon and Olympic silver medals as well as a year-end WTA number 1 ranking ensured Kerber was the woman in tennis entering the 2017 season.

2017 however failed to pan out, with Kerber only making it to the 1st round of the French and US Opens and 4th round of Wimbledon and the Australian Open. These performances left Kerber outside the Top 20 come the WTA year-end rankings.

Our initial predictions for the 2018 season demonstrated ELO's lost faith in Kerber as we consistently underestimated her, predicting her to lose 6 of her 1st 10 games (she won 10/10). Interestingly, the bookies maintained their faith predicting all 10 wins correctly (worth noting however they consistently favoured Kerber during her down season also). This contrast in predictions indicate we may be overfitting to form, that being we must also consider a player's class or pedigree.

We propose maintaining an additional ability measure for all players, this being their all-time max ability. What was previously considered a player's current ability $\beta_{i,t}$ will now be

referred to as their "form" ability. As usual a player's form ability is generated using equation 10 taking a weight of their overall and surface-specific ability. New, however, is that this process is done twice, once for their form ability and once for their all-time ability. Their new current ability is now a weight of their form ability and their all-time max ability:

Let w = weight assigned to all-time ability, β^{max} = a player's all-time max ability.

$$\beta'_{i,t} = \beta_{i,t}(1-w) + \beta_{i,t}^{max} w \qquad (18)$$

This new ability $\beta'_{i,t}$ is only used to generate a prior probability and score, it does not get used when updating abilities in equation 12. Form abilities ($\beta^{o}_{i,t}, \beta^{s}_{i,t}$) are updated using equation 12 and max abilities are only updated should a player's current form ability exceed their max.

4.2.5 - Trend

ELO is excellent reactively. It quickly increases abilities for over-performance and decreases them for under-performance. It fails however to proactively adjust predictions in anticipation of upsets. When making a prediction, ELO does not consider a player's recent trend, just assuming that the trend is captured in the current ability.

Consider two players: player A has an ability of 2,150 and player B 2,100. Under the current model we would favor player A assigning them a prior probability of 0.57. Now let's say that player A was rated 2,300 3 games ago and player B in contrast was rated 1900. It is likely we would adjust our predictions slightly with the knowledge that player A is in free fall, struggling for form while player B is quickly ascending.

To do this we must create a metric to evaluate how a player is performing relative to their current ability:

Let $\delta_{i,t}$ = player i's trend at time t, f = weight assigned to most recent match (0,1)

$$\delta_{i,0} = 0$$

$$\delta_{i,t+1} = \delta_{i,t}(1-f) + s_{i,t} f \qquad (19)$$

Equation 19 shows trend being updated to include a players most recent match. When you think about it $s_{i,t}$ (equation 17) effectively measures how a player is performing relative to their ability, if their performance is greater than their prior probability $s_{i,t}$ is positive and the opposite is true if their performance is less. Thus when $\delta_{i,t}$ is positive it means a player has recently been outperforming their current ability.

Let b = max percentage of a player's ability that can be added or subtracted to better reflect their recent trend.

$$\beta'_{i,t}(1 + \delta_{i,t} b) \qquad (20)$$

Equation 20 shows the temporary boosting or decreasing of current ability. $\beta'_{i,t}$ is calculated as before using equation 18 but is now either increased or decreased temporarily prior to the match to better reflect a player's current trend. If they are under-performing their current ability is decreased and it is increased if they are over-performing.

4.3 - Also considered

4.3.1 - Time away

Common in sport is the theory of "match fitness" - this being no matter how much you train, the only way to truly be at one's best is by playing competitive games regularly. When a player is away for a prolonged period of time it is likely that when they return, they will return at a lower level than when they left. Due to ELO ratings being outcome-based, we must create a new metric to allow players' abilities to decrease the longer they are away.

Something to consider however is that every player's history prior to their break is different. For example, it can be the case that a prodigious 15/16 year old may qualify for an event but then not play again on the senior tour for a year and thus it wouldn't be fair to depreciate their ability to the same extent as an established player who suffered a severe injury and was absent for a year. As a workaround, instead of flat-out depreciating a player's ability we converge towards their worst ability recorded.

Two depreciation metrics were considered, one being the standard $(1-r)^n$ used in general accounting and the other being a sigmoid function. Both functions were fitted to include a "grace period", this being a period of time where if a player was absent for less than it their abilities would not depreciate. In both attempts however, it was the case that maximum likelihood fitted the parameters such that the grace period was larger than any observed absence, and therefore no player had their ability depreciated.

It would be interesting to retry this should we obtain a dataset flagging injury instead of trying to effectively infer problems based on time away.

4.3.2 - Age

In equation 8 when describing τ, the shape parameter was mentioned and described as ensuring τ doesn't get too small as a player plays more games. What shape is really modeling is certainty of a player's ability (the larger the shape the more certain we are). This means that even as a player plays more games a smaller shape keeps τ to a high enough value such that any unexpected increase or decrease of form is still quickly captured in a player's ability.

It is generally expected that over the course of a player's career they will ascend in ability while young, reach a peak performance and then begin to decline as they age further. We believed this rise and fall could be tied into a player's shape. This would mean that certainty would be low while a player is young in an attempt to capture large increases in ability, heighten as a player reaches their peak where we are most certain and then decrease again as a player ages in an attempt to capture their decline.

Figure 6 illustrates this showing the MLE best fit:

$$shape = 0.22(0.109(age) - 0.00196(age^2)) \qquad (21)$$

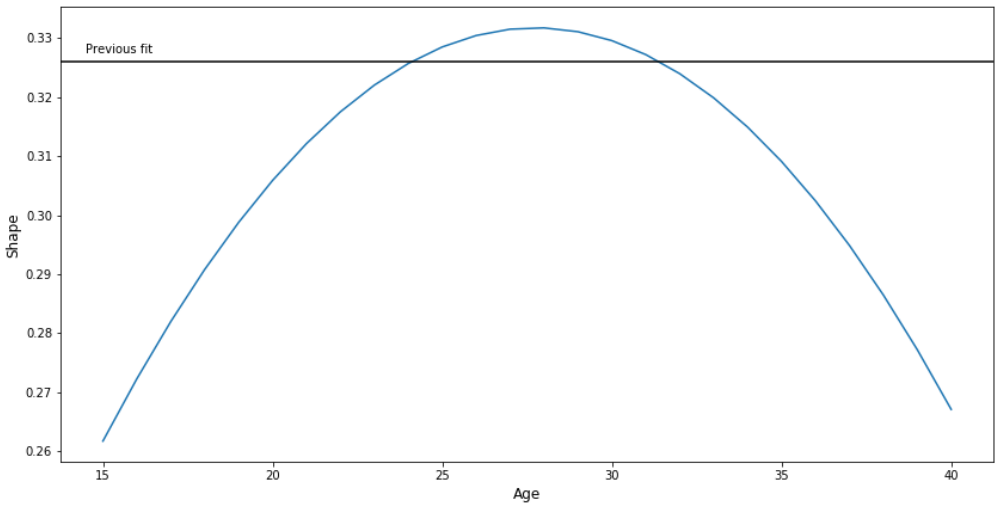

Figure 6: Dynamic shape charting a player's uncertainty parameter as they age

While the model converged on a fit that makes intrinsic sense to our original hypothesis with a player reaching their peak at around 27, the log likelihood ratio test (as discussed in equation 14) failed to achieve a significant value and was thus discarded.

This approach could be re-tried on a train dataset over a larger period of time where more players were observed ascending and falling over the course of their careers.

4.3.3 - Grand Slams

One of tennis fans' biggest criticism of the ELO system is treating each game the same irrespective of round or tournament. It could be argued the "round" criticism is unfounded as the ELO system is designed to reward players for winning more matches and thus its cumulative effect results in players who reach later rounds in tournaments generally earning more points than those who didn't. The "tournament" criticism however seems warranted.

In tennis, the 4 Grand Slam tournaments are the pinnacle of the sport; failing to win one almost certainly guarantees a player won't be remembered as an all-time great. Tennis fans argue that these tournaments should have a higher weight or impact than games from other tournaments.

Giorgi et al in the High Dimensional Model proposed temporarily boosting player difference in Grand Slam tournaments due to a "Grand Slam effect", the idea being that great players are at their best in Grand Slams and thus the difference between a good and a great player is even larger.

Consider two players i and j playing in a Grand Slam, let γ denote the Grand Slam effect:

$$\delta_{ij,t} = \gamma(\beta_{i,t} - \beta_{j,t}) \qquad \gamma => 1 \qquad (22)$$

They propose should γ be found to be significantly greater than 1 it would suggest a Grand Slam effect to be present.

It could be argued that this approach does not solve the problem proposed by tennis fans. By boosting the ability difference between two players, one does not assign a higher weight to a specific match but instead reduces the better player's upside while increasing their downside.

Take a Grand Slam game where the ability difference $(\beta_{i,t} - \beta_{j,t})$ = 300 and let γ = 1.2. The ability difference is now said to be 360. The better player's prior probability using equation 7 is now 0.888 instead of 0.849. This reduces the winning player's upside by 0.039 using equation 4 while increasing their downside by the same amount. In essence equation 22 just boosts the prior probabilities' certainty.

The squared errors (the squared difference between the prior probability and the outcome) (Fig 7) show how the model including trend (model 7) fared across different tournament levels within the test dataset. It can be seen with the exception of Davies Cup; it was most effective in Grand Slams with the lowest mean and median squared errors across all other tournament levels. Applying the approach suggested could artificially improve the model by making our strong predictions more certain.

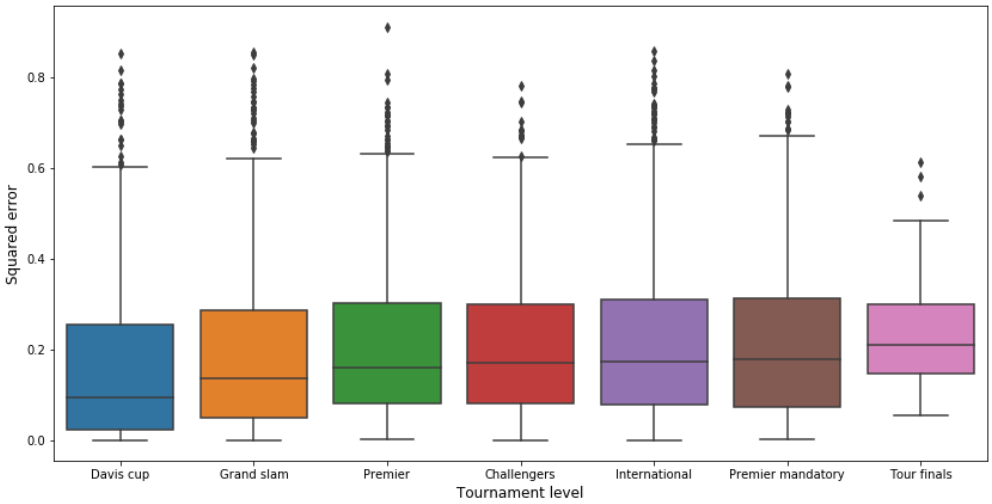

Figure 7: Trend model (model 7) brier scores across various tournament levels

Instead what tennis fans are really proposing is heightening τ for Grand Slams thus increasing the impact these matches have on a player's ability:

$$\tau' = \gamma \times \tau \qquad \gamma => 1 \qquad (23)$$

Equation 23 was tested and resulted in a small log likelihood increase, however failed to obtain a significant value.

5 – Results

Table 1 – Models

Model	Description	Number of parameters
1	**Basic model:** *Section: 4.1.1*	3
2	**Surface model - eta constant:** *Section: 4.1.2*	4
3	**Surface model - eta vector:** *Section: 4.1.2*	6
4	**ITF:** *Section: 4.2.2*	5
5	**Performance model:** *Section: 4.2.3*	7
6	**All time model:** *Section: 4.2.4*	8
7	**Trend model:** *Section: 4.2.5*	10

Table 2 – Fitted parameters

Model	K	Offset	Shape	Surface Weight	ITF deduction	Binom – p	Straight sets boost	All time	Trend rate	Trend weight	Log Lik
1	248.58	18.1	0.44	-	-	-	-	-	-	-	-13,869
2	247.96	20.62	0.417	0.189	-	-	-	-	-	-	-13,822
3	247.96	20.62	0.417	[0.11 (Hard), 0.227 (Clay), 0.206 (Grass)]	-	-	-	-	-	-	-13,820
4	270	1	0.389	0.127	0.329	-	-	-	-	-	-13,366
5	270.05	1	0.326	0.131	0.327	0.528	0.079	-	-	-	-13,226
6	270.05	1	0.26	0.125	0.318	0.53	0.064	0.386	-	-	-13,205
7	269.95	1	0.26	0.121	0.339	0.53	0.072	0.598	0.285	0.074	-13,163

One of the more surprising results from Table 2 is surface weight being a constant rather than 3 individual weights. The low surface weight for Hard in model 3 is likely down to the fact a large proportion of games were played on hard surfaces (62.7% of the WTA train data) and thus the overall ability was largely a reflection of hard ability.

It would be interesting to see someone investigate surface weight further - looking at time dependancies. The WTA season is broken into mini surface seasons where players start on hard surfaces move to Clay and Grass and then back to Hard. It could be the case whereby surface weight is greater when players just move to a particular surface and tails off the longer, they play on it as overall ability begins to reflect surface ability more.

Model 4 shows just how sorely our model was missing ITF data. It is important to reaffirm model 4 does not attempt to predict ITF games, but just considers the effect they have on WTA matches. It is possible model 4 could be improved should the MLE include ITF matchups.

Model 5 includes the main goal of this dissertation which was to find a way to incorporate match score in tennis ELO. It is now the case where a player can win a game and it still result in a decrease of their ELO ability should they perform to a lower level than expected. The performance function in equation 16 is relatively simplistic and could potentially be improved further to include more granular details such as serving, returning and break-point performance.

Model 7 shows the effect recent trend can have on a player's current ability allowing for it to be inflated or deflated by up to 7.4%. From the data seen, trend is generally between ± .6 meaning in reality current ability is generally adjusted ± 4.4% (0.6 * 0.074).

5.1 – Prediction

Table 3 –Model performance on test data

Model	Accuracy	Brier score
1	64.20%	0.2198
2	64.57%	0.2185
3	64.67%	0.2186
4	66.73%	0.2078
5	67.20%	0.2068
6	67.07%	0.2063
7	67.57%	0.2050

Table 4 –Model performance vs WTA and bookmakers

Model	Accuracy	Brier score
WTA – ranking*	63.52%	-
Bet365 – normalised	65.86%	0.2056
Model 7	66.89%	0.2086

*17 games not included as missing winner or loser rank

Table 4 compares our predictions to the bookmakers' data (as discussed during section 3.2.2) as well as the WTA rankings. Although on the surface focusing plainly on accuracy it would appear our model has outperformed the bookies, the brier score (mean squared error of all predictions) tells a different story.

Currently a prediction is considered correct should the prediction be > 0.5 for the winning player. Interestingly if we adjust that to ≥ 0.5 the Bet365 model improves to 68.92% while ours remains the same. It would appear that to mitigate risk Bet365 consider the certainty in their prediction and converge towards 0.5 if they are unsure.

Table 5 –Model calibration

Model	50-59%	60-69%	70-79%	80-89%	90-100%	Calibration
Model 7	0.555 (0.551)	0.649 (0.651)	0.733 (0.747)	0.838 (0.844)	0.949 (0.926)	0.00016
Bet365	0.547 (0.556)	0.652 (0.644)	0.777 (0.747)	0.886 (0.841)	0.98 (0.915)	0.00147

Table 5 compares our model's calibration with that of Bet365's model (in brackets is the mean predicted value for both models within the specified bin). When creating this table, games in which Bet365 were 50/50 split between players were removed as they were neither correct nor incorrect.

Calibration looks at how close a model's forecasted probabilities matchup with that of the true probabilities, meaning ideally, we want forecasts with a 50% probability to be correct 50% of the time. Calibration is also vitality important for the purpose of ranking players; if 51% probabilities are right 65% of the time there is an unfair advantage towards the better player as they will receive a larger upside when it comes to updating abilities than they should, and the opposite is true if they are only right 40% of the time. The calibration value at the end shows the mean squared difference between the fraction correct within the bin and the mean predicted value (the lower the better).

Figure 8 visualises the model calibration comparison. It can be seen for the most part model 7 toes the perfectly calibrated line relatively well, slightly overestimating the certainty in 70-79% range while slightly underestimating the 90-100%. In contrast, the bookies consistently

over-perform their probabilities, meaning their fraction of positives within each bin is generally larger than the mean predicted value. This is likely down to the normalising of the bookies' probabilities which removed some certainty from their predictions.

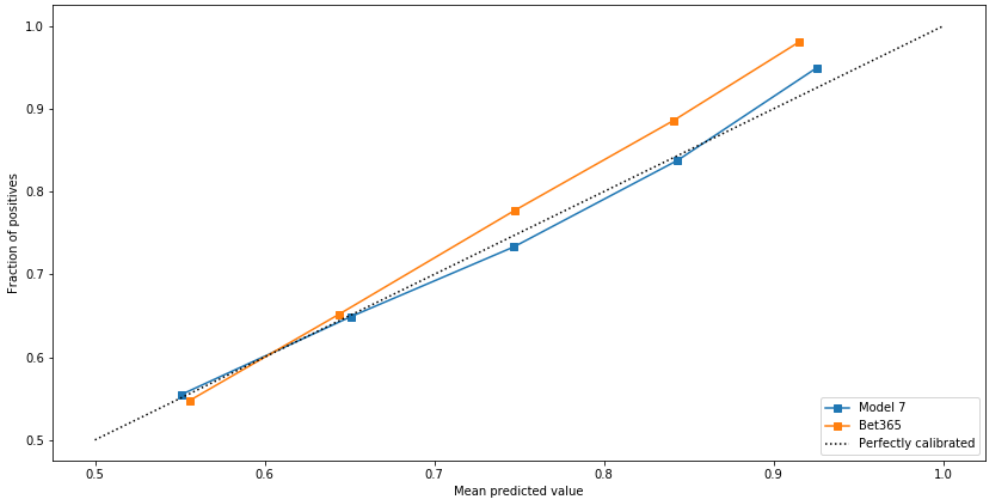

Figure 8: Model calibration comparison

5.2 – Ranking

Table 6 –WTA ranking vs Model 7, year-end 2019

Position	WTA	Model 7 - Overall	Model 7 - Hard	Model 7 - Clay	Model 7 - Grass
1	Ashleigh Barty	Serena Williams	Serena Williams	Serena Williams	Serena Williams
2	Karolina Pliskova	Ashleigh Barty	Ashleigh Barty	Ashleigh Barty	Ashleigh Barty
3	Naomi Osaka	Bianca Andreescu	Bianca Andreescu	Kiki Bertens	Petra Kvitova
4	Simona Halep	Petra Kvitova	Naomi Osaka	Simona Halep	Bianca Andreescu
5	Bianca Andreescu	Naomi Osaka	Karolina Pliskova	Bianca Andreescu	Karolina Pliskova
6	Elina Svitolina	Simona Halep	Aryna Sabalenka	Petra Kvitova	Simona Halep
7	Petra Kvitova	Karolina Pliskova	Petra Kvitova	Naomi Osaka	Naomi Osaka
8	Belinda Bencic	Kiki Bertens	Simona Halep	Petra Martic	Johanna Konta
9	Kiki Bertens	Aryna Sabalenka	Elina Svitolina	Karolina Pliskova	Aryna Sabalenka
10	Serena Williams	Karolina Muchova	Caroline Wozniacki	Karolina Muchova	Kiki Bertens

Table 6 compares the year-end WTA ranking with our model's (model 7 - the trend model) overall and surface-specific rankings. It is important to note "Model 7 - Overall" is no one player's true ability as in our ELO model we always consider the surface being played on. The rankings for each surface, however, are the true current rankings of each players surface specific ELO ability.

The standout difference between the two approaches is Serena Williams. During the 2019 season Williams was hampered with injuries resulting in her only competing in 10 tournaments within the 52-week window thus only earning 62.5% (10/16) of her available points. She did however reach the later two Grand Slam finals (Wimbledon and the US Open) seeded 11 and 8 respectively. This once again displays the WTA ranking problem that is cumulative ability ≠ current ability.

Caroline Wozniacki being ranked top 10 on the hard surface was highly unexpected and likely highlights a flaw in the model. While her 2018 season was strong an Australian Open win the highlight resulting in a short-lasted return to WTA number 1. From that point on however, she has continuously struggled with injuries, playing irregularly with long breaks in between. So hampered was her career since that Open win, she announced her retirement in 2020 at the same event.

A combination of things are likely playing a part in her inflated ranking: 1. ELO's inability to decrease players abilities while they don't compete, 2. Wozniacki had an incredibly high peak in her career which is still being considered by the all-time parameter. While time away penalties were discussed earlier, it could also be the case where all-time ability should also consider how long it has been since a player's peak.

6 - Discussion and conclusion

Within this dissertation, we have shown a simple and explainable ranking model that drastically outperforms the current WTA approach and is competitive with state-of-the-art bookmakers' models. We have shown time and time again how cumulative ability is in no way reflective of current ability and thus does not align with the purpose of tournament seeding.

It must be noted however that it is likely the current WTA model is commercially driven. It is within the WTA's financial interests to have the world's best players at as many tournaments as possible in order to appease sponsors and heighten TV rights interest. By taking cumulative ability as the ranking metric they ensure players are motivated to play at more tournaments and thus reward them for doing so.

If it is the case whereby the ranking are commercially driven the number of players seeded in tournaments should be reduced such that the advantage earned on an uneven playing ground is removed in as much as possible. This would vary tournament to tournament due to (as stated earlier) different tournaments seeding a different number of players, however, the number of players seeded in any tournament under the WTA ranking system should be minimised in as much as possible.

7 – References

1. Women's Tennis Association. 2020. Rankings. [online] Available at: https://www.wtatennis.com/news/1312140/rankings.
2. Bradley, R. A. and Terry, M. E. (1952) Rank analysis of incomplete block designs: I, The method of paired comparisons. Biometrika, 39, 324–345.
3. Cattelan, M., Varin, C. and Firth, D., 2012. Dynamic Bradley-Terry modelling of sports tournaments. Journal of the Royal Statistical Society: Series C (Applied Statistics), 62(1), pp.135-150.
4. Gorgi, P., Koopman, S. and Lit, R., 2019. The analysis and forecasting of tennis matches by using a high dimensional dynamic model. Journal of the Royal Statistical Society: Series A (Statistics in Society), 182(4), pp.1393-1409.
5. Morris, B. and Bialik, C., 2015. Serena Williams And The Difference Between All-Time Great And Greatest Of All Time. [online] FiveThirtyEight. Available at: https://fivethirtyeight.com/features/serena-williams-and-the-difference-between-all-time-great-and-greatest-of-all-time/.
6. Kovalchik, Stephanie. (2016). Searching for the GOAT of tennis win prediction. Journal of Quantitative Analysis in Sports. 12. 10.1515/jqas-2015-0059.
7. Morris, B., Bialik, C. and Boice, J., 2016. How We'Re Forecasting The 2016 U.S. Open. [online] FiveThirtyEight. Available at: https://fivethirtyeight.com/features/how-were-forecasting-the-2016-us-open/.
8. Sackman, J., 2019. Rethinking Match Results As Probabilities. [online] Heavy Topspin. Available at: http://www.tennisabstract.com/blog/2019/12/28/rethinking-match-results-as-probabilities/.