# Elo Ratings and the ATP

Justine Huang

Stat 157 Final Project

Fall 2014

# 1 The Elo rating system

## 1.1 History of the Elo system

The Elo rating system was designed by Arpad Elo, an amateur chess player and physicist. He was appointed the chairman of the United States Chess Federation rating committee, and in that role, he would implement the now-famous Elo rating and ranking[1] system for chess players. Now, it is used by the World Chess Federation (abbreviated to FIDE) to rank its players, and although the equations have been refined since its initial US introduction in 1960 and international introduction in 1970, the rating system has stayed remarkably similar to the Elo's original ideas.

Before the Elo system was implemented for chess ratings, the Harkness system was used to rate and rank players. Although the Harkness system was initially heralded because it was the first time chess players had a way of quantifying his or her abilities, there were fringe cases that made the system less accurate than some preferred. The class system in chess that remains today (i.e., Senior Master, Master, Expert, Class A) is a vestige of the Harkness system that Elo found to be an accurate portrayal of the standard deviation in terms of strength of performance over a series of games (Ross).

The Elo system has plenty of merits, which has contributed to its longevity of use in chess and application to everything from NFL teams (notably by FiveThirtyEight) to the rating individual attractiveness ("Facemash", the precursor to Facebook in the movie "The Social Network") to rating the quality of players in the video game World of Warcraft. One reason the Elo system is widely implemented is because Elo ratings are simple to calculate and are fairly transparent. In addition, even though the spread of the resulting numbers are somewhat arbitrary depending on the implementation and cannot be directly applied to the games at hand, they are accurately reflective of the relative performance of a player against the included field. Elo ratings are a good benchmark of the quality of a player; it does a much better job of accounting for quality of the opposition than many other rankings systems, including the ranking system the ATP uses.

## 1.2 Elo system assumptions

Before jumping into the mechanics of how the Elo rating system works, there are a few assumptions that are important to keep in mind when looking at Elo ratings:

1. The distribution of individual performances is approximately normal. In the context of chess ratings, the standard deviation is approximately one class.
2. Player ratings, which are reflective of skill, are normally distributed.

[1] A quick digression: although ranking and rating may seem to mean the same thing, they are slightly different. A *rating* is a quantified measure of the strength of a player or team. A *ranking* is a list giving the relative strength of players or teams. Rankings are usually designed by ordering a particular list of players or teams by their ratings.

3. When Elo was originally designing the rating system, he assumed that in chess, performance could only be inferred from wins, losses, and draws. This assumption has been loosened in more recent applications of the rating; for example, in various NFL team Elo ratings, the points spread of the game is also included in the rating calculation.

The first two assumptions regarding the normal distribution tend to be the most controversial assumptions and tend to be the source of mathematical concerns towards the model. The USCF found the use of the normal distribution to be potentially inaccurate, particularly for lower rated players, just given the observed performance of lower ranked players against higher ranked players. Because of this concern, the USCF now uses a logistic distribution model for its ratings, while the FIDE still uses the normal distribution for its player rating calculations.

### 1.3 Mechanics

The mechanics around the Elo rating system are very simple. The movement of one's rating is calculated using two equations. Call $R_{A,}$ $R_B$ the initial ratings, $E_{A,}$ $E_B$ the expected scores, $S_{A,}$ $S_B$ the actual scores, and $R'_{A,}$ $R'_B$ the new ratings of players A and B, respectively. When Player A competes in a match against Player B, Player A has an expected score ($E_A$) that can be calculated by the formula below:

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

The same is done for Player B, but with $R_A$ and $R_B$ switched in the previous equation, resulting in $E_B$. By algebra, $E_A + E_B = 1$. Once the match is played and $S_{A,}$ and $S_B$ are determined, $R'_A$ and $R'_B$ are calculated by the below formula:

$$R'_{A\,=}\, R_A + K\,(S_{A\,-}\, E_A)$$

The same is done for Player B.

Note, that in a pure Elo rating system with equal transactions (where the winner earns N rating points and the loser drops N rating points), the average number of points stays the same, and $R_A + R_B = R'_A + R'_B$.

The K in the above equation is called the k-factor, which helps determine how much a player's ranking can fluctuate after a given match. Practically speaking, if the K-factor is too high, the sensitivity of the model will be high when there are few events and many points will be exchanged per game. In chess, the K-factor used in calculation is based on how many matches a player has already played and their existing rating. As of July 2014, the FIDE uses the following ranges (FIDE):

- K = 40 for players until they have played 30 games, for all players until their 18th birthday, as long as their rating remains under 2300
- K = 20 as long as a player's rating remains under 2400

- K = 10 once a player’s published rating has reached 2400 and remains at that level subsequently, even if the rating drops below 2400

The rationale behind this multi-tiered system is that higher rated players tend to have more stable abilities and hence, their ratings should not fluctuate as much. Ratings deflation can occur over the long run in a ratings system with equal transactions because players tend to enter the system with a low rating and leave the system with a higher rating (Wikipedia). The multi-tiered system used by the FIDE described above can combat this issue by essentially injecting points into a system. Furthermore, other Elo-based chess ratings systems will feed rating points into the system by giving bonus points to improving players while others use a rating floor. In ratings systems where there is a rating floor, where if one’s rating drops below the floor they are stricken from the list, there can be ratings inflation. The actual average rating number chosen is fairly arbitrary; in many applications, and in the following analysis, it is chosen to be 1500.

### 1.4 Other concerns around use of Elo ratings

One particular issue around the use of Elo ratings is the conflict of interest for a player who wishes to protect his or her rating. Since Elo ratings do not fluctuate if one does not play a match, if one is theoretically happy with her or her rating, he or she can maintain it by not playing. Therefore, one downfall of Elo ratings is that players have to be active for the ratings to be accurate. Even if a player decides to play, he or she can manipulate his potential outcomes; selective pairing is an issue in certain Elo ratings systems. A player can try to only compete against opponents with a higher Elo, given his or her less risk of losing ratings points. This is not an issue in most Elo ratings systems, though; for example, in the tennis implementation talked about in below, given the random nature of a draw, a player will have to play lower ranked players in order to have the chance to play opponents with a higher Elo rating. Another question that has been raised has been around the accounting of intangibles in Elo ratings. It can be argued that the Elo ratings do not account for intangibles such as a psychological advantage one player has over another particular player. On the other hand, it can be argued that intangibles are essentially wrapped up in the ratings already, given that Elo ratings are calculated based on actual performances.

## 2 The ATP ranking system

### 2.1 ATP background

The ATP is the “governing body of the men’s professional tennis circuits – the ATP World Tour, the Challenger Tour and the ATP Champions Tour” (ATP). The ATP World Tour encompasses 61 tournaments in 30 countries and is widely considered to be the competitive circuit for the best men’s tennis players in the world. Tournaments in the ATP World Tour are classified into a few tiers, which are: Grand Slams, ATP World Tour Masters 1000, ATP World Tour Masters 500, and ATP World Tour Masters 250. The ATP Challenger Tour is essentially one step below the ATP World Tour, and it is often used as a training ground for younger players to refine their skills and help grow their rankings in hopes of playing in higher level tournaments.

There are four Grand Slams: the Australian Open, the French Open, Wimbledon, and the US Open. The Australian Open is held in Melbourne, Australia in January of each year on hard courts. The French Open is held in Paris, France between May and June each year on clay courts. Wimbledon is held in London, United Kingdom between June and July each year on grass courts. The last grand slam of each year, the US Open, is held in Flushing, New York, between August and September on hard courts. Grand Slams have long been considered the most prestigious tennis tournaments in the world, and are often used as a measure of a player's success or greatness. To win a Grand Slam, a player must win seven rounds of matches; the total draw of each tournament consists of 128 players.

ATP World Tour Masters 1000 tournaments are the second most prestigious tier of tournaments on the ATP World Tour. There are nine of these tournaments, three on clay (Monte-Carlo, Madrid, and Rome) and six on hard courts (Indian Wells, Miami, Canada, Cincinnati, Shanghai, and Paris) (ATP). The draw size of these tournaments ranges from 48 players (Paris) to 96 players (Indian Wells and Miami) and feature first round byes for highly seeded players.

### 2.1 ATP ranking criteria

The year-end Emirates ATP Rankings are based on each player's performance in the aforementioned 61 tournaments. It does not, however, count every tournament a player participates in. For 2014, the rankings are based on his total points from the four Grand Slams (where the winner earns 2000 points), eight of the nine ATP Masters 1000 tournaments (where the winner earns 1000 points), and his best six results from all other ATP tournaments played. For players outside the top 30, who are not automatically accepted into Masters 1000 level and Grand Slam tournaments, if they do not play a Grand Slam or Masters 1000 tournament, they can increase the number of "other" tournaments counted in his ranking by one.

At the end of the season, there is the ATP World Tour Finals, which were held in London, England in November 2014. The top seven in the Emirates ATP Rankings as of November 3, 2014 automatically qualified for this round robin tournament. In the case where one or more current-year Grand Slam champions did not make the top seven but fell between eighth and twentieth, the highest ranking champion qualifies as the eighth player. If that does not occur, the player in eighth place on November 3, 2014 qualifies as the eighth player.

### 2.2 Criticisms of the ATP rankings

There are a few issues with the ATP rankings and have been well documented in the wake of complaints from players such as Rafael Nadal (Bryant). In the current ATP rankings, points are measured on a rolling 52-week basis, which can make the rankings seem somewhat inconsistent week to week. For example, in the hypothetical case where a player won the Monte-Carlo Masters 1000, he will have those 1000 points 51 weeks after the day he won the tournament final, but if he does not win the tournament again, in 53 weeks after he won the final he will not have those 1000 points. The potential fluctuation between two very close rankings dates is large, and is not necessarily reflective of the true rating and ranking of tennis players.

ATP rankings also do not take strength of schedule and opponent into account. Any tournament in the same tier is given the same amount of points, regardless of the strength of the players in the draw. Because six of the non-Grand Slam and Masters 1000 tournaments count in the rankings, if a player can play as many of these "other tournaments" as he can, then his chance at a favorable draw increases, and his ranking inflates. Finally, the rankings system punishes those who draw unfortunate draws in Grand Slams; if a player happens to always draw a top four ranked player in the first round of a tournament, his strength of schedule hurts him immensely in the ATP points system.

## 3 Data Set / Objective

The data set used in the following analysis is comprised of the 2013 and 2014 Emirates ATP Rankings, and the match results from all the Grand Slam and ATP Masters 1000 level matches. All of the data was gathered through the ATP's website.

There were 201 players in this analysis; any player that played either a Grand Slam or ATP Masters 1000 level match was included. These players ranged from Noah Rubin (ranked 770 on November 4, 2013) and Rafael Nadal (ranked 1 on November 4, 2013).

There were 1203 matches in the data set, 695 in the Masters 1000 tournaments, and 508 in the Grand Slams. First-round byes were treated as matches against a player with an Elo rating of 0 and wins by walkover were treated as straight sets wins (a Grand Slam match is best-of-five sets and a Masters 1000 match is best-of-three sets).

The objective of this analysis is to determine whether the ATP rankings are providing a list of the best players in the ATP in a reasonable order. Another objective is to see whether the field that played at the ATP World Tour Finals lines up with the Elo rankings' predicted field.

## 4 Applying Elo ratings

The analysis below was conducted using the aforementioned data set and Elo equations. The expected score calculation was done using each player's rating from before the tournament through all the matches in a tournament, and his new, post-tournament rating was calculated by comparing the expected number of his wins in the tournament to his actual number of ratings. This was then repeated for the next tournament. In total, this process was repeated 13 times for all the tournaments in the data set.

### 4.1 The base case

The base case used in this analysis is as follows:

- Initial ratings are based on a logistic function applied onto the final 2013 rankings, with the average Elo rating set at 1500
- The k-factor for Grand Slams is 30 and the k-factor for Masters 1000 tournaments is 20
- For Grand Slams: a three set win is worth 1, a four set win is worth 0.8, and a five set win is worth 0.6. Similarly, a three set loss is worth 0, a four set loss is worth 0.2, and a five set loss is worth 0.4.
- For Masters 1000: a two set win is worth 1 and a three set win is worth 2/3. Similarly, a two set loss is worth 0 and a three set loss is worth 1/3.

The decision to use a logistic function to set the initial ratings versus the traditional Elo assumption that ratings are approximately normally distributed was based on similar reasoning to the USCF's reasoning that the normal distribution is not reflective of the actual skill distribution of players. Similarly, tennis player skill is likely not normally distributed; given the result of many past tournaments, it is likely that the skill level of lower ranked players is closer than the skill level of higher ranked players (where the gap between players is larger). The logistic function is a good representation of this type of spread; the equation used here was:

$$\text{Logistic function} = \frac{1}{1+1.05^{-(2013\ final\ ranking-5)}}$$

The parameters were estimated such that the ratings difference decreases at the inflection point, approximately where the 2013 final ranking = 5. Then, those results were scaled to the chosen 1500 average by taking the average of the logistic function results, multiplying by 1500, and subtracting that from 1500 + 1377 (the average of the logistic function results). In Figure 1 below, the ratings are graphed against the resulting Elo ranking of the player.

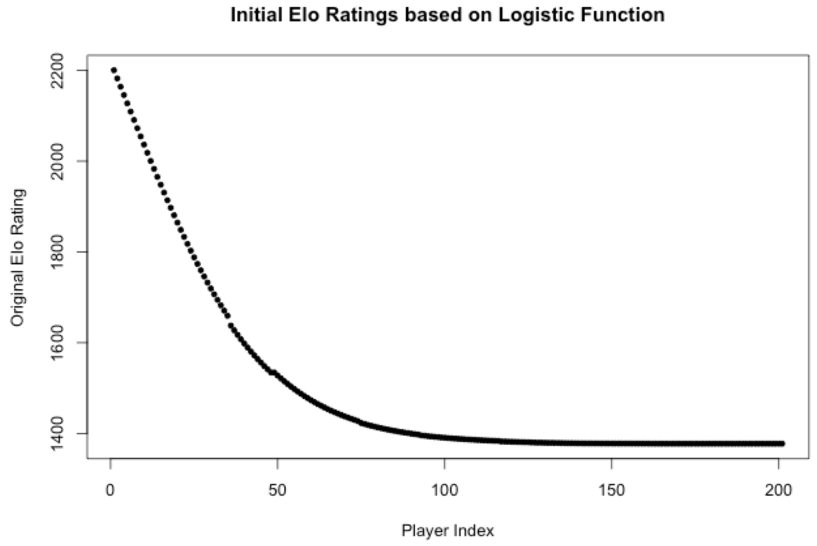


Figure 1: Initial Elo ratings used in the base case

The decision to use a two-tiered k-factor based not on player ratings and number of matches played but on the tournament tier is based on the limited size of the data set. The maximum number of matches a player can play in the data set used is 84 matches, which was highly unlikely. In addition, many of the players in the data set only played one or two matches. To eliminate the possibility for ratings inflation, to keep it a pure Elo rating system, and to reflect the increased prestige and importance of the Grand Slams, the k-factor of 30 for Grand Slams and 20 for Masters 1000 tournaments was chosen. 20 and 30 were chosen using the two-thirds rule that chess initially used for their k-factor tiers.

The actual win values assigned to Grand Slams and Masters 1000 matches differ because Grand Slam matches are best of five sets and Masters 1000 matches are best of three sets. For straight sets wins in either type of match, the value assigned is a full 1. For non-straight sets wins, the actual win value is assigned ratably by the potential number of sets played; for example, a four-set win is worth 0.8 because the two players played 80% of the possible sets (four of five). Instead of just assigning a 1 to all wins and a 0 to all losses, the ratable win values were used because of the nature of tennis matches in hopes of increasing the precision of the ratings, and as an analog to the points spread used by FiveThirtyEight in their Elo ranking of NFL teams.

| Player | Elo | ATP | Difference |
|---|---|---|---|
| Nadal | 1 | 3 | -2 |
| Djokovic | 2 | 1 | 1 |
| Del Potro | 3 | 138 | -135 |
| Federer | 4 | 2 | 2 |
| Ferrer | 5 | 10 | -5 |
| Murray | 6 | 6 | 0 |
| Wawrinka | 7 | 4 | 3 |
| Tsonga | 8 | 12 | -4 |
| Berdych | 9 | 7 | 2 |
| Raonic | 10 | 8 | 2 |
| Almagro | 11 | 72 | -61 |
| Haas | 12 | 76 | -64 |
| Nishikori | 13 | 5 | 8 |
| Gasquet | 14 | 27 | -13 |
| Isner | 15 | 18 | -3 |
| Robredo | 16 | 17 | -1 |
| Simon | 17 | 21 | -4 |
| Cilic | 18 | 9 | 9 |
| Youzhny | 19 | 47 | -28 |

Figure 2: Base case Elo rankings versus ATP rankings at 2014 season end

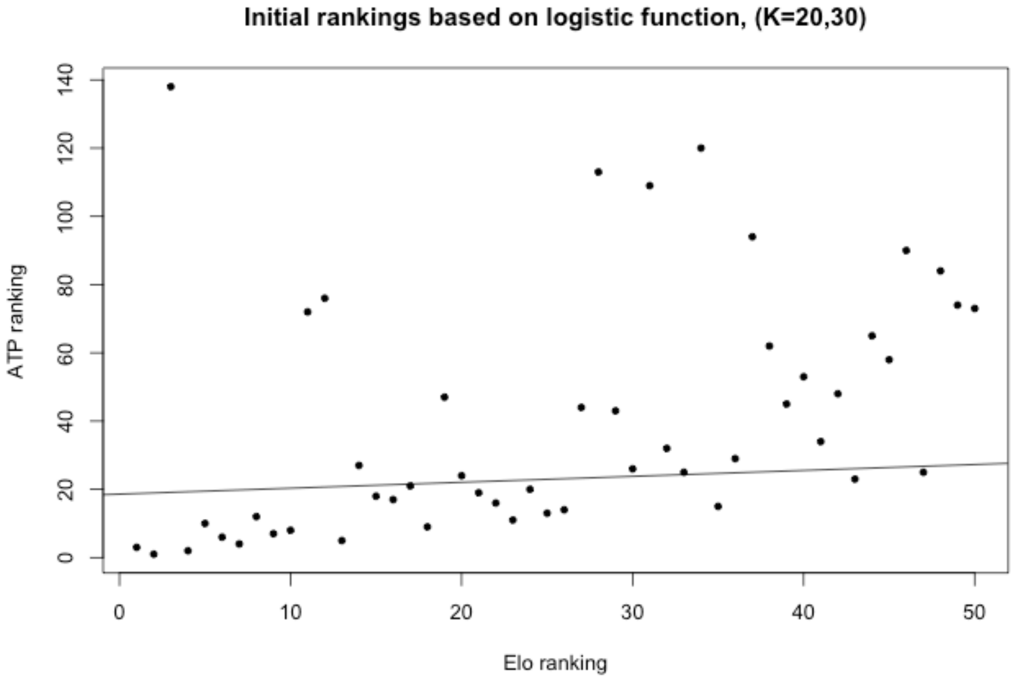


Figure 3: Comparison of base case and ATP

| ATP Year-End Championship - Logistic Initial Ratings | | |
|---|---|---|
| **ATP Final 8** | **Elo Final 8** | |
| Djokovic | 🟨 Nadal | Djokovic |
| Federer | 🟨 Del Potro | Federer |
| Wawrinka | 🟥 Ferrer | |
| 🟥 Nishikori | Murray | |
| Murray | Wawrinka | |
| Berdych | 🟥 Tsonga | |
| Raonic | Berdych | |
| 🟥 Cilic | Raonic | |

🟥 Difference between Elo ranking and ATP ranking

🟨 Player injured

Figure 4: Base case hypothetical 2014 ATP Year-End Championship roster

As shown in Figure 4, using the base case, the eight players that qualify for the hypothetical year-end championships based on the Elo rating are fairly different than the actual field that played in London in November. In fact, the adjusted R-squared when regressing the Elo rankings on the ATP rankings is only 0.1565. Juan Martin Del Potro benefitted the most from the Elo system versus the ATP system, as seen above, where he finished 138 in the ATP rankings but third in the Elo ratings. This is due to the fact that Del Potro got injured very early on in the season, and was helped by his high ranking in the beginning of the year (he was ranked fifth at the end of 2013). Because Elo ratings cannot change if a player does not play, he maintained a high rating. This problem may imply that Elo ratings used in a tennis context must require a certain number of matches played within a given season for the rating and ranking to be valid.

Kei Nishikori and Marin Cilic provide interesting case studies into how ATP rankings differ from a pure Elo rating method.

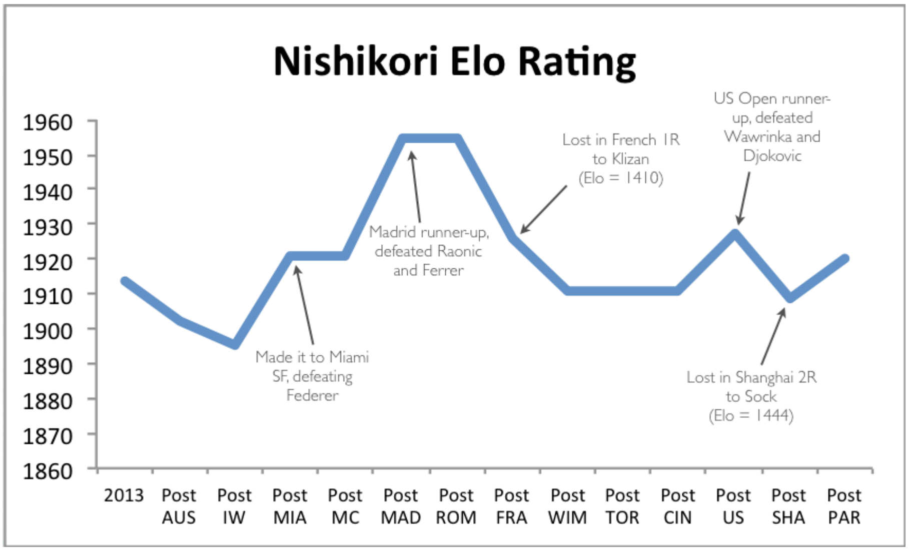


Figure 4: Kei Nishikori's Elo rating movement throughout the season

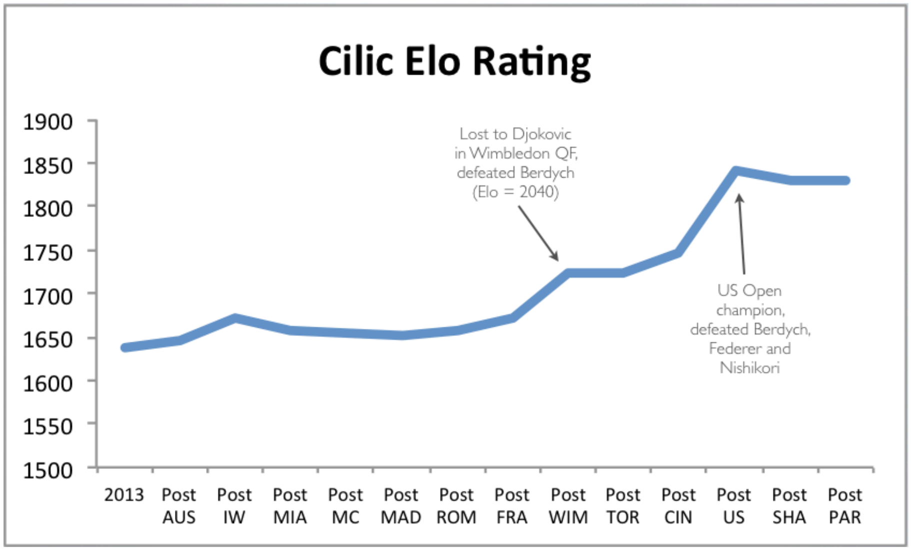


Figure 5: Marin Cilic's Elo rating movement throughout the season

Nishikori's rating fluctuated wildly through the 2014 season, as he defeated substantially higher ranked players, such as Roger Federer, but would also lose to significantly lower ranked players, such as Jack Sock. His inconsistency was not rewarded by the Elo system, but the ATP ranking system rewarded him handsomely for his runner-up performance in the US Open. In addition, his Elo rating did not increase as much as one might have expected post the US Open; this may be due to the fact that he had an easier draw than thought. The ATP rankings are strength of schedule-neutral, versus how Elo ratings take strength of schedule directly into consideration.

Cilic's rating was fairly flat throughout the first half of the season, which is consistent with analysts talking about his disappointing to expected level of performance. In the first half, Cilic essentially beat who he was supposed to, and lost to stronger players (ATP). In the second half of the season, Cilic improved his ranking, and received a bump after winning the US Open and defeating three quality players for the title (Tomas Berdych, Roger Federer, and Kei Nishikori). That improvement, however, was not enough to compensate for his low ranking at the beginning of the year (37) following his doping ban at the end of the 2013 season.

### 4.2 Sensitivity analysis – use of logistic function

Given the large discrepancy in Cilic's Elo ranking and Cilic's ATP ranking, the Elo rating was performed again with all players given an initial flat Elo rating of 1500. This is more consistent with the way the Emirates ATP rankings have rankings points roll off 52-weeks after each tournament, making the ranking effectively only based on this year's performance and with very little weight given to the previous year's rankings. The logistic function results in an Elo rating system that puts weight on the previous year.

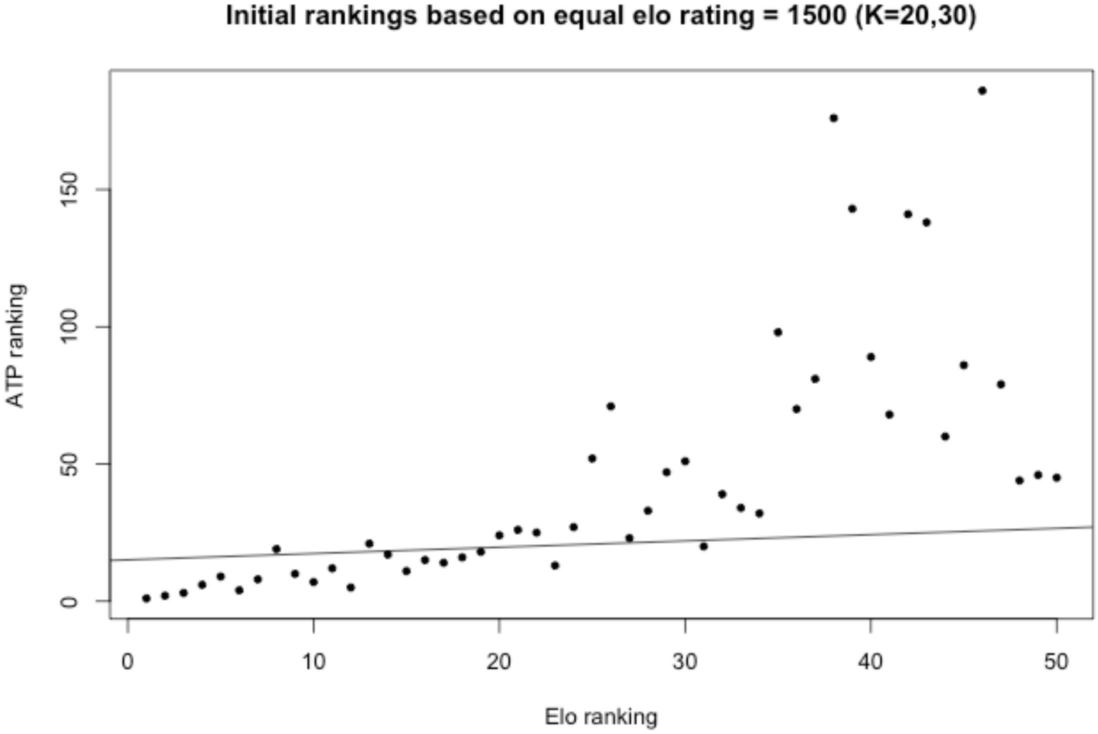
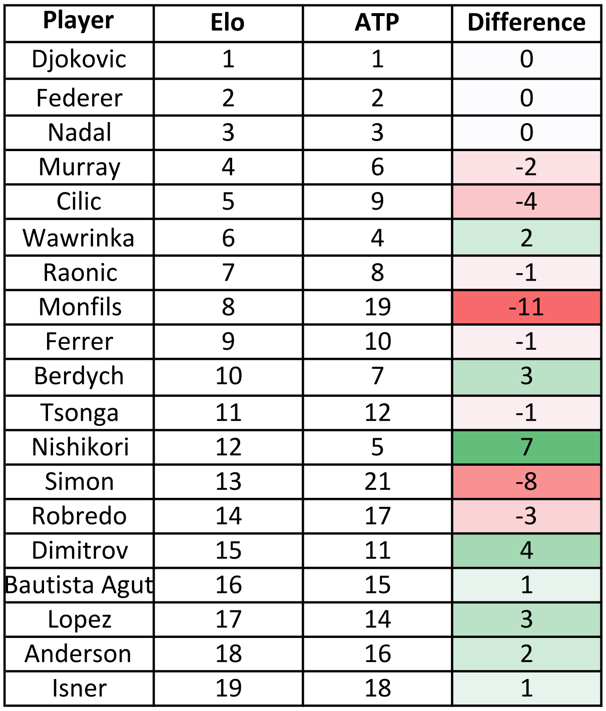


Figure 6: Comparison of using flat initial Elo rating on rankings and ATP ranking



| Player | Elo | ATP | Difference |
|---|---|---|---|
| Djokovic | 1 | 1 | 0 |
| Federer | 2 | 2 | 0 |
| Nadal | 3 | 3 | 0 |
| Murray | 4 | 6 | -2 |
| Cilic | 5 | 9 | -4 |
| Wawrinka | 6 | 4 | 2 |
| Raonic | 7 | 8 | -1 |
| Monfils | 8 | 19 | -11 |
| Ferrer | 9 | 10 | -1 |
| Berdych | 10 | 7 | 3 |
| Tsonga | 11 | 12 | -1 |
| Nishikori | 12 | 5 | 7 |
| Simon | 13 | 21 | -8 |
| Robredo | 14 | 17 | -3 |
| Dimitrov | 15 | 11 | 4 |
| Bautista Agut | 16 | 15 | 1 |
| Lopez | 17 | 14 | 3 |
| Anderson | 18 | 16 | 2 |
| Isner | 19 | 18 | 1 |

Figure 7: Elo rankings versus ATP rankings, using flat initial ratings

As seen in Figures 6 and 7, this method results in a final Elo ranking closer to the ATP's results, especially for the highest ranked players; the adjusted R-squared increases to 0.5495 in this case versus the 0.1565 when using the initial ratings based on the logistic function. The correlation between these two different Elo rankings is only 0.230. One big part of this difference is probably due to players like Del Potro; he did not play many matches during the year so he did not have the chance to move his ranking past the initial 1500.

This analysis suggests that the logistic function based Elo rankings may be answering a different question than the Emirates ATP rankings are aiming to answer. The base case Elo ranking is answering the question of who is the absolute best player, regardless of injury. The flat initial Elo rankings and the ATP system are answering the question of who the best player is during the current season. Juan Martin Del Potro is likely one of the best players in the world when he is not injured, hence his high base case rating; however, when he is hurt for the majority of the season, no one considers him one of the best players that season, hence the modified Elo and ATP rankings.

### 4.3 Sensitivity analysis – selection of k-factor

The k-factor plays such a large role in Elo ratings that to check for the sensitivity of the model to adjustments in k-factor, the ranking method was performed again using the same base case, but instead of a tiered k-factor, a flat k-factor of 25 was used for all tournaments. This essentially eliminates the "prestige" and "importance" factor given to Grand Slams in the base case. The results are shown in Figure 8 below.

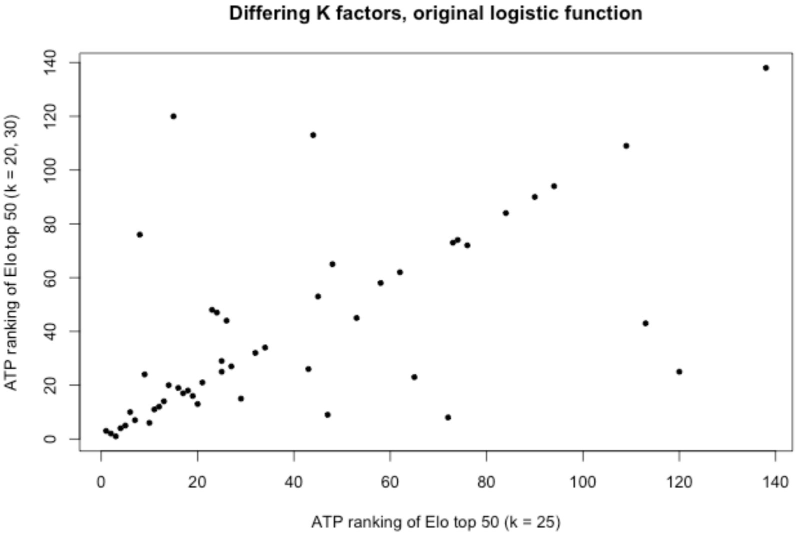


Figure 8: Changing the K-factor used from 20 and 30 to K = 25

This plot implies that the k-factor change to a flat k-factor did not alter the rankings remarkably for the top 50; most of the points in the plot lie on the straight diagonal line or close to it, suggesting the rankings are the same or close regardless of which k-factor was used. The correlation between the two rankings is 0.6285.

**4.4 Sensitivity analysis – selection of win values**

The win value for a four-set win in a Grand Slam match was 0.8 in the base case because four of the five possible sets were played. Another way to look at this actual win value assignment is to give a four-set win the weight of 0.75, because the winner won three of the four sets actually played.

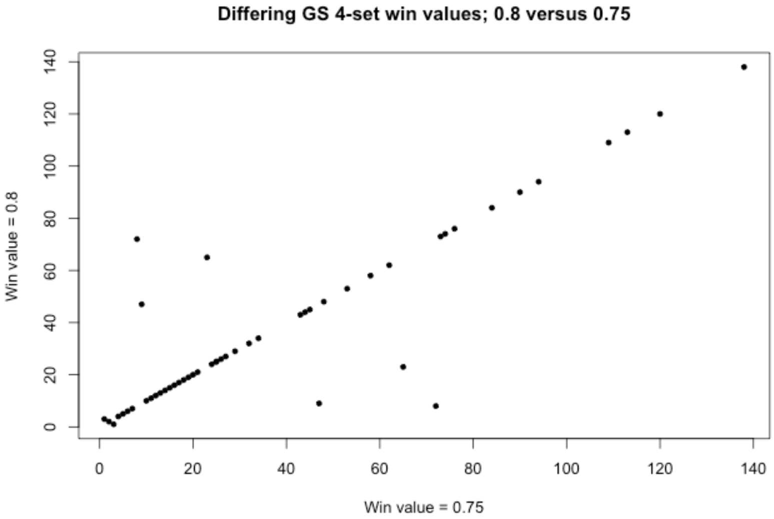


Figure 9: Changing the four-set win value from 0.8 to 0.75

Changing the 4-set win value retains the base case rankings more than altering the k-factor, which is expected since there are not a lot of four-set matches in the data set. The correlation between rankings in the base case and altering the 4-set win to 0.75 is very high, at 0.8777.

## 5 Conclusion

Using Elo ratings to rank tennis players has its pros and cons versus the traditional ATP ranking system. The Elo rating system is more fair to players, in that it does not punish players for unfortunate draws since they do not schedule their own matches. For example, Ryan Harrison, a promising young American player, would not have been punished for drawing Novak Djokovic or Rafael Nadal in the first or second rounds of various tournaments given their large advantage in rating. In the ATP system, he is punished for losing early in Grand Slams. All in all, the Elo rating system is a better head-to-head rating system for tennis given that it counts directly for opponent skill. Although the Elo rankings in this analysis are fairly incomplete given the limited nature of the data set, as more matches are added to the Elo model used, they will become more accurate.

There are some cons of using the Elo rankings, however. There needs to be a way of accounting for injuries in the system, instead of just brushing them aside as in this analysis. One problem that may be solved with the addition of more matches is the outsized effect of initial rankings in the base case.

In conclusion, this analysis has shown that the ATP rankings may not be an entirely accurate representation of the best tennis players at any given time, but to create a much improved ranking system, there are many factors that need to be taken into consideration.

# Works Cited

"2014 Calendar." ATP World Tour. ATP, n.d. Web. 08 Dec. 2014.

"About the ATP Challenger Tour." ATP World Tour. ATP, n.d. Web. 08 Dec. 2014.

"Emirates ATP Rankings FAQ." ATP World Tour. ATP, n.d. Web. 08 Dec. 2014.

"FIDE Rating Regulations Effective from 1 July 2014." World Chess Federation. FIDE, 1 July 2014. Web. 08 Dec. 2014.

Glickman, Mark, and Albyn Jones. "Rating the Chess Rating System." Chance 12.2 (1999): 21-28. Web.

|, Howard Bryant. "How to Fix Tennis' Big problems." ESPN. ESPN Internet Ventures, 22 June 2013. Web. 08 Dec. 2014.

Ross, Daniel. "Arpad Elo and the Elo Rating System." Chess News. ChessBase, 16 Dec. 2007. Web. 08 Dec. 2014.

Silver, Nate. "Introducing NFL Elo Ratings." DataLab. FIveThirtyEight, 4 Sept. 2014. Web. 08 Dec. 2014.

"Singles Rankings." ATP World Tour. ATP, 08 Dec. 2014. Web. 08 Dec. 2014.

Sonas, Jeff. "Rating Inflation - Its Causes and Possible Cures." Chess News. ChessBase, 27 July 2009. Web. 08 Dec. 2014.

"Elo Rating System." *Wikipedia*. Wikimedia Foundation, 12 June 2014. Web. 08 Dec. 2014.