Analyzing and Predicting Peak Performance Age of Professional Tennis Players

Miha Mlakar
Jožef Stefan Institute
Jamova cesta 39
SI-1000 Ljubljana, Slovenia
miha.mlakar@ijs.si

Tea Tušar
Jožef Stefan Institute
Jamova cesta 39
SI-1000 Ljubljana, Slovenia
tea.tusar@ijs.si

ABSTRACT

Tennis is a sport where physical ability, match experience and mental toughness all have to be at the peak to achieve best results. To determine when this typically happens, we analyzed aging trends and calculated the average aging curve for professional players from 1974 to 2014. We showed that 25 is the age when it is most likely for a player to reach his career peak. In addition to analyzing performance, we also developed an algorithm for predicting how a player will perform in the next years. The algorithm finds players that had similar career paths and uses these players to predict the performance for the next years.

1. INTRODUCTION

Tennis is a very popular sport played all over the world by thousands of players of different ages. In recent years we are witnessing older players like Federer (33), Ferrer (33), Lopez (33) and Karlovič (36) still winning ATP tournaments and playing important roles on the ATP World Tour. This contradicts what we think and poses an interesting question: what is the age at which a tennis player is most likely to be at the peak of his career?

Of course the answer is not the same for all players. If we draw a line presenting the performance of a player depending on his age, we get his aging curve [5]. Players have different aging curves, due to various reasons such as different constitutions, styles of play, mental preparedness and other factors. Also some players tend to develop faster physically and some players get injured more frequently when they get older. All these facts influence the player careers and consecutively their aging curves.

Measuring players' peak performances and drawing their aging curves is an integral part of player analysis in many sports. These analyses first began in baseball, where the database of all statistics is very detailed and covers players from major and minor leagues [4]. By trying to determine the player aging curves the clubs and scouts are trying to find out which players are worth buying. We draw our inspiration for this paper from one of the most well-known systems for determining aging curves and also predicting career peaks in baseball called PECOTA [4].

In tennis, there were some analyses measuring how age influences performance in tennis, but to our knowledge none of them did it thoroughly on a great amount of data. In [3] authors identified the age of peak performance in a broad range of sports including tennis and associated the performance peak with how much explosiveness and how much stamina is needed in this sport. In [1], the authors showed how tennis players in Grand Slam tournaments are now older then they used to be in nineties. In addition, they analyzed peak performances for players of Wimbledon in 2014.

In this paper we analyzed aging trends and presented average aging curves for professional tennis players. In addition, we are interested in predicting how well a specific player will perform in the next years. The predictions are based on determining the similarities between the players that already had similar career paths and have similar characteristics.

Predicting a player's performance or his aging curve can be very useful when trying to estimate, for example, if and when a young prospect player will break the top 10 or how long can a player over 30 keep his current level.

The structure of this paper is as follows. In Chapter 2 we describe the preprocessing of data, including normalization and determining which players should be included in the analysis and which shouldn't. Chapter 3 presents the average aging curve for professional tennis players. We also detected and presented other facts about player performances in dependance of the age or time playing. Chapter 4 describes the algorithm for predicting the performance and presents the obtained results of predictions. The final chapter concludes the findings and names the possibilities for future work.

2. DATA PREPARATION

The data used in this paper includes rankings and player characteristics for men ranked on the ATP rankings from the beginning of 1974 to the end of year 2014. We included all professional players, not just the best, because we wanted to obtain a general tennis aging curve and also we wanted to be able to predict the aging curve for players of different levels.

2.1 Data filtering

Since we were analyzing players aging trends, we had to be careful not to analyze players with missing data. So, we removed all players that (i) were in the middle of their careers in 1974, because maybe the career peak has already passed and (ii) players that still had active careers at the end of 2014, because their peak may still be coming. Including these players with inaccurate career performances would impair the results.

The further inspection of data showed that some players have rankings just for a few years. In order to find the appropriate aging curve, we had to determine, if we should limit the career lengths, or use all players.

Let say we have a player that is ranked only for one year. The analysis would show that this year is his performance peak and the best year for playing tennis. Because his aging curve would have been different if he would play more years, his example would influence the shape of the aging curve in an incorrect way.

To see how limiting the career lengths would influence the distribution of players, we draw two graphs representing player rankings depending on their ages. On the first one (Figure 1) we draw lines for player with at least two years on the rankings. The second one (Figure 2) shows players with at least ten years long careers.

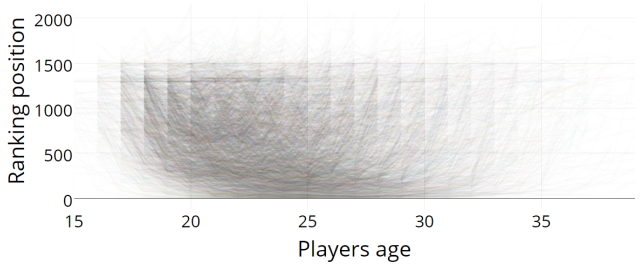

Figure 1: Player ranking for players who were on rankings at least 2 years

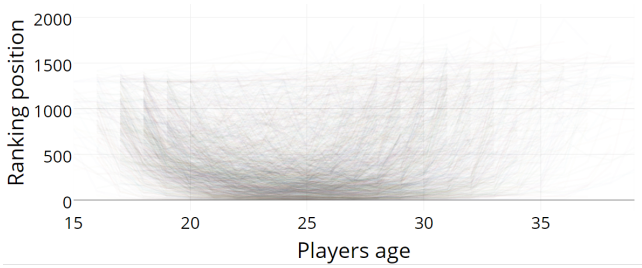

Figure 2: Player ranking for players who were on rankings at least 10 years

We can see that the main difference between graphs is with the players aged between 18 to 25 that are not ranked very high. This is normal because in tennis, if you are ranked lower than 350 you have more expenses than incomes [2]. This results in the fact that promising players might get discouraged and drop their rackets even before they reach their best years. Due to the fact their career endings are not a result of physical nature, keeping these players in the analysis would wrongly alter its results. So, we decided to keep only players with their careers lasting at least ten years. With this rule, we also ensured that phase of getting experience and reaching the top and the phase of physical declination because of age gets included.

2.2 Data normalization

In order to be able to compare and calculated aging curves from players of different quality we had to normalize the performance of each player. We could normalize the player positions or the number of points obtained, but both measures are not very appropriate for normalization. Problem with the position is that it is not equally hard to get 10 positions if you are ranked in top 20 or if you are ranked in top 500. So to include that the normalization would have to be very complex. The problem with gained points is that in 2009 the ATP decided to change the number of point gained on tournaments. So after that date the points are incomparable, so not appropriate to normalize.

To overcome this issue we created surrogate points. For every position on the ranking we calculated average number of points that were needed for that position. The surrogate points obtained in this way are not influenced by the rule changes or by the missing data for number of points, thus we used them for normalization.

3. TENNIS AGING CURVE

As already mentioned in the introduction, there is a lot of talk nowadays about how in modern tennis experienced players are more dominant and better than they used to be and also that the younger players don't have enough quality and experience to win big tournaments.

To determine if this is true, we took the last 30 years of rankings data and for every year divided players into four groups; players younger than 20 years, players between 20 and 24 years, players between 24 and 30 years old and players older than 30 years. In this way we measured young prospect players, players reaching their career peak, experienced players and players that are considered old(er). For every group we summed all the point that the group obtained throughout the year. The results are presented in Figure 3.

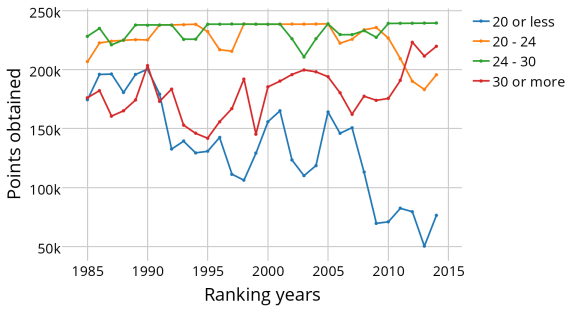

Figure 3: Performance of different age groups.

We can see that performance of younger players really got

worse in the last 10 years. On the contrary the players over 30 years are performing better in the last 10 years. Since this differences in performances could be the reason for shifting the aging curve, we decided to test this and calculate two aging curves. The first one is for players born before 1975 and the second one is for players born after that year. The curves are presented in Figure 4.

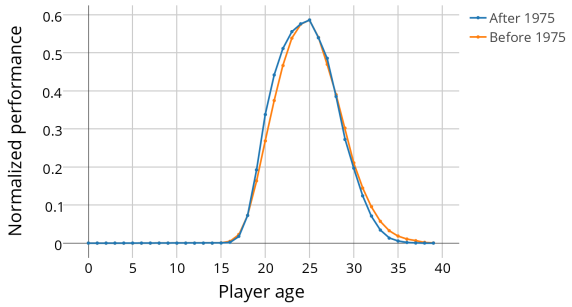

Figure 4: Aging curves for players born before and after year 1975.

We can see a small difference between the curves, but both curves clearly show that a player is most likely to reach his career peak at the age of 25. We can also see that the curve is less steep before the peak and more steep after it. This indicates that players are more gradually approaching their peak performance and that after 25 years the chances of reaching career peak performance are decreasing quickly.

In addition to combining all normalized performances to see what is the average performance over the years, we also wanted to see when players reach their best ranking. The distribution where for every year we counted how many players reached their best ranking is presented in Figure 5.

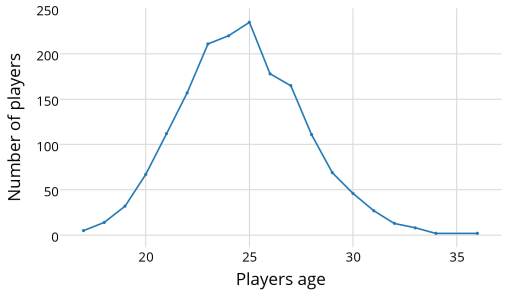

Figure 5: When players reach their best position.

We can see that most players reach their best position when they are 25 years old which coincides with the aging curve.

Since we know that the man's body is reaching its maximum physical capabilities before 25 years, we wanted to know, what is the reason that the peak performance age is only at 25. The reason is experience. Tennis is as much a mental game as it is a physical one. And to be able to handle the pressure and learn to adapt your game to the various opponent styles, you need to play on professional tour for some time. Figure 6 presents how many years after getting a first ATP ranking, players reach their best career position.

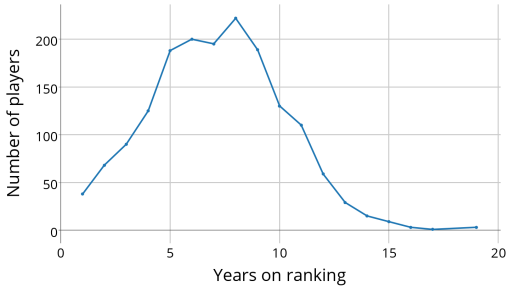

Figure 6: Years needed to reach the best position.

We can see that most player need 9 years on the ATP tour to reach their best position. The distribution presenting players' first appearance on the rankings (Figure 7) shows that most prayers get their first points around the time, when they are 17 or 18 years old. By combining age of first ranking and the years needed to get all the experience, the players are typically 25 years old at their career peak.

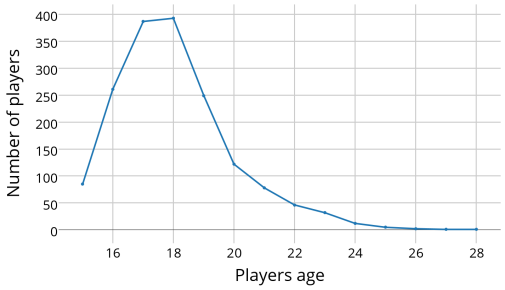

Figure 7: Age of first ranking.

4. PREDICTING PERFORMANCE

Predicting a player's performance can be very interesting, if we want to estimate how high on the rankings a player can get in the next few years. To make such predictions, we designed an algorithm based on similarity measures that finds other players that already had similar career paths. The prediction is then obtained by combining the performances these similar players had in predicted years.

4.1 Prediction procedure

When comparing two players we took the beginning (age of first ranking) of their careers and for every (next) year compare their highest number of obtained points for this year. We summed the absolute differences for every year to get an overall similarity difference between two players for a specific number of years.

After comparing a specific player to all other players, we rank players by their overall similarity difference. The players with smallest differences are picked to contribute to

predictions. The prediction is the average number of points obtained by these player for a specific year.

4.2 Testing predictions

In order to determine the accuracy of the predictions, we decided to use leave-one-out methodology to predict the future number of points for all players in our database.

There are three main parameters that could be tuned: (i) the number of years taken for learning the similarities, (ii) the number of similar players taken for making the predictions and (iii) for how many years in the future we make predictions.

For determining the similarities between players it is true that increasing the number of years used, will improve the similarities and thus also the predictions. We opted for the 5 years long period, since this should be long enough to determine the career trend and also short enough not to miss the expected career peak.

For the number of similar players we did some preliminary tests with 5 and 10 players and the obtained results were similar (Figure 8). For the predictions we choose to use 5 similar players, since we wanted the predictions to include only players with very similar career paths.

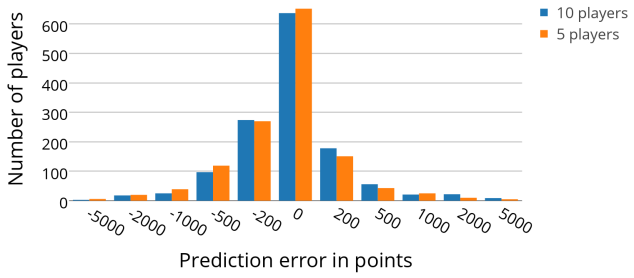

Figure 8: Prediction error for using 5 or 10 similar players for predicting the number of points

To test the prediction accuracy of the algorithm, we predicted players' points for 2, 3 and 4 years in advance. The histogram of prediction errors is shown in Figure 9.

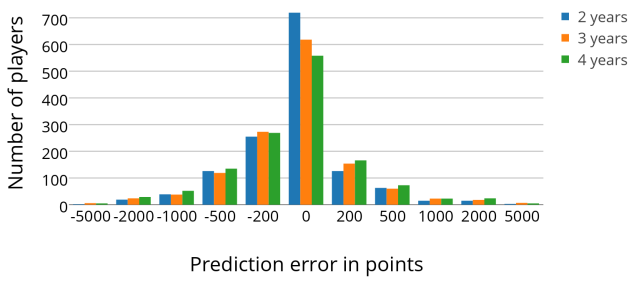

Figure 9: Prediction error for predicting players' points for 2, 3 and 4 years in advance.

The x axis shows the classes of errors. For example, the one around zero represents all predictions that were between -100 and 100 surrogate points away from the actual number of points. We can see that most of the predictions are within the 100 points radius. The medians were 93 for 2 year predictions, 117 for 3 year predictions and 129 for 4 year predictions. There are some players with larger prediction errors, but of course some players get injured or suddenly improve dramatically, so their predictions will be off for a big margin. This events are almost impossible to predict so no matter the algorithm, there will always be some players with wrong predictions.

5. CONCLUSIONS AND FUTURE WORK

In this paper we analyzed the data from professional tennis players ranked on the ATP World Tour rankings between 1974 and 2014. We defined our own measure for comparing players' performances throughout the years disregarding the changes in ranking system.

We presented aging trends and calculated average aging curve for professional tennis players. We showed that a player is most likely to be in his career peak when he is 25 years old.

In addition, we designed algorithm for predicting how many points will a specific player get in the next years. The algorithm is based on finding players with similar careers and for most players prediction error is less then 100 points.

In order to make predictions more accurate, additional player characteristics would have to be used. We could use player rankings data from junior (under 18 years) rankings, detailed statistics obtained for all matches and include additional custom-defined parameters like style of play or some other player specifics.

6. ACKNOWLEDGMENTS

Thanks to Jeff Sackmann for making the data publicly available and to Jernej Zupančič for long and productive discussions.

7. REFERENCES

[1] C. Bialik. Get off my tennis lawn! http://fivethirtyeight.com/features/get-off-my-tennis-lawn. Accessed:2014-07-04.

[2] C. Bialik. Tennis has an income inequality problem. http://fivethirtyeight.com/features/tennis-has-an-income-inequality-problem Accessed:2014-07-04.

[3] R. Schulz and C. Curnow. Peak performance and age among superathletes: track and field, swimming, baseball, tennis, and golf. *Journal of Gerontology*, 43(5):113–120, 1988.

[4] N. Silver. *The signal and the noise: Why so many predictions fail-but some don't*. Penguin, 2012.

[5] K. J. Stewart. Physical activity and aging. *Annals of the New York Academy of Sciences*, 1055(1):193–206, 2005.