Patterns of Play: Predicting tennis match outcomes and player styles

Hunter Hobbs
University of Colorado, Boulder

Nivetha Kesavan
University of Colorado, Boulder

Pratik Revankar
University of Colorado, Boulder

Dmitri Tarasov
University of Colorado, Boulder

ABSTRACT

Who would be favored to win in a tennis match between Ivan Lendl and Novak Djokovic, two of the best player to ever compete, yet decades apart? In a game that some claim has been around since the 12th century, there have been countless greats to compete at the highest levels. What similarities exist between top-competitors from different eras? What styles can be inferred from analyzing historical tennis match data? What attributes would the ideal player have under certain conditions? These are some of the questions that this paper aims to answer through data mining techniques. First we integrate several data sets, and apply a range of data mining techniques to search for clusters and determine player advantages. Through these various mining techniques we highlight key rivalries among professionals and search for interesting correlations between players. Finally we see if we can successfully predict a winner between two chosen opponents. We show that one can pull dense analytics from tennis data by using data mining techniques.

ACM Reference Format:
Hunter Hobbs, Nivetha Kesavan, Pratik Revankar, and Dmitri Tarasov. 2020. Patterns of Play: Predicting tennis match outcomes and player styles. In *Proceedings of* . ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The introduction of technology has revolutionized the way we consume and enjoy sports. Instant replays, "Hawk-eye" verdicts, match predictions, forecasts, and player analyses have, in many ways, improved the sport, and the decision making involved. Tennis has recently seen an increase in the use of technology, with major tech players, like IBM and Accenture investing and creating partnerships with ATP, WTA, and all major Grand Slams. With huge investments and use of technology and real-time match statistics, it has become imperative that this data is analysed, understood fully, and made available to the appropriate audiences.

Technology has created a disruption in the sport, and several top ranking players have invested and begun to include analytics to their team, to better understand their game and their opponents. Current world leader, Novak Djokovic is known to leverage match statistics, and his team of physiotherapists, coach and strategy analysts used AI algorithms to improve his game, which played a part in helping him win the 2019 Wimbledon Grand Slam. Judy Murray, coach and mother of Andy Murray, is also a supporter of technology and has heavily relied on a data-driven approach to her coaching.

With an evolving digital audience, data and technology have been leveraged to provide a better experience and give key insights. Data mining and analysis have shown great promise in improving various aspects of the sport, such as decision making by the chair umpire, available data for real-time commentary, pre-match predictions for marketing, creating traction, media hype, and post-match player analysis.

Having the domain knowledge and being huge fans of the sport, we use data mining techniques and the data available to highlight some key insight and showcase the disruptive potential that data and technology has in

tennis. We think that if we can predict match outcomes for two, or four in the case of doubles, opponents; this information can be leveraged in training or match analysis. One could use such knowledge to train in a specific way for an upcoming opponent.
Applying clustering techniques on a subset of features could lead to interesting groupings of players, or potential outliers to highlight some skillful advantages. This would be valuable information to understand why some players are extraordinary on the professional level.

2 RELATED WORK

2.1 Player rankings for match prediction

Clarke and Dyte [2] used ATP rankings to predict the player's chance of winning in a head to head match. The player rankings are derived from a set of rating points, the rating points of two players prior to the tournament is collected and using the Logistic Regression model they predict the winner of the head to head based on the difference in points rating.

McHale and Morton[7] developed an alternative player ranking based on the past match results, number of points won by each player in a match, how recent the match was played, etc. This ranking was used to predict the winner of a match and it performed better than using ATP rankings. This new ranking focused on ease of win, opponent quality, how recent the match was and the absolute performance of the player whereas ATP ranking gives higher weight-age to how frequently a player participated in different events rather than absolute match performance.

2.2 Points won in a serve

Klassen and Magnus[4] used rankings and points won by players on service to predict the match outcome at the beginning of the match as well as during the match using a graph technique based on the current score and current server.

Liu[6], Newton[8] modelled the probabilities of a player winning a set and match based on the probability of a player winning a point while serving.

2.3 Common Opponent Strategy

Knottenbelt et al. [5] used a hierarchical Markov model to predict the probability of each player winning a match. The model uses the match statistics of an opponent that has been encountered by both players and using the statistics, it computes the probability of each player winning a serve and the match.

2.4 Tactical Analyses in Professional Tennis

Leeuw et. al [3] performed tactical analyses for a specific player using Subgroup Discovery. They used Subset Discovery for the data-set of point,match, and stroke characteristics with 4,13, and 27 features respectively. Using this method, the authors identified the characteristics of a successful serve point of a specific player and they used it to identify the difference between generic successful serve point and serve point of the specific player.

The best subgroup with the highest ROC curve with an area of 0.79.

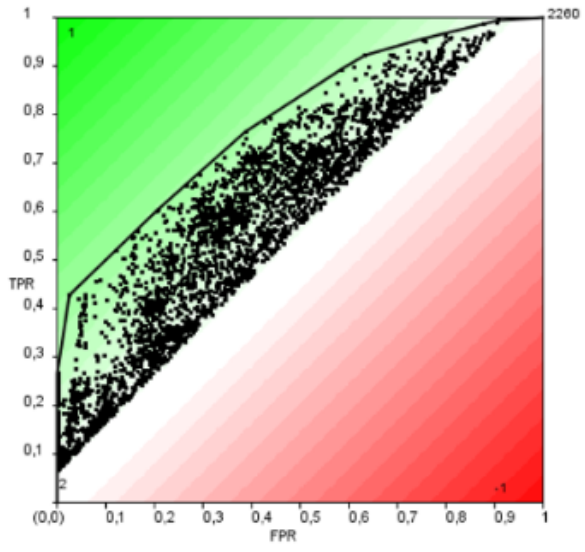

Figure 1: ROC Curve search depth 2

2.5 Modelling ATP Tennis as a Network

Zhang [11] created a network model of ATP matches based on parameters such as win,forehand,backhand, and serve networks. The paper discusses finding common groups with Louvain's Algorithm. Based on the network,the authors discovered the structural roles which revealed the prominent attribute in any individual player's game.

The first experiment was to find the ranking groups based on player ranks which had uninformative results.

2.6 Finding Maximal Non–Redundant Association Rules in Tennis Data

Weidner et.al[10] presents a method of using maximal association rules for tennis data-set. The diagram of the process is described as:

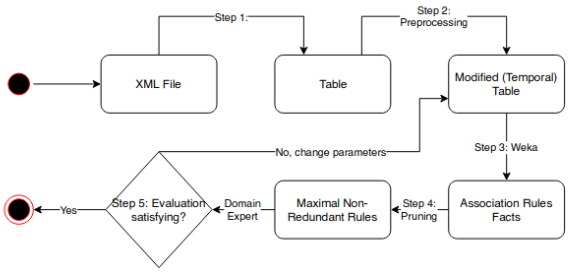

Figure 2: maximal non-redundant association rules process

The first two steps is processing the raw data into a time based table of entries. The third step is split into two parts which are the generation of association rules and the creation of facts. The process of association rules is generated through an algorithm like apriori algorithm. The other part is to use the association rules to establish Prolog facts. The third step is to transform the Prolog facts to maximal non-redundant association rules which might not be present from the apriori algorithm.

3 DATA SET

The data sets that were used for this paper were compiled from several different sources. Mainly the data can be divided into three classes: player data, match data, and tournament data. The player data schema consists of player meta-data, e.g., age, height, weight, handedness. The match data schema consists of statistical match data, e.g., surface type, tournament, date, winner, winner seed. The tournament data has similar meta-data to the match data but at the tournament level. The data sets that are used for the project are the following hyperlinks:

- ATP singles and doubles matches
- Jeff Sackmann data set
- ATP serve and Volley data set
- WTA matches
- ATP matches data-set
- Tennis Betting Data
- Open Data Soft
- Jeff Sackmann Tennis Match Charting Project

4 MAIN TECHNIQUES APPLIED

As mentioned in Section 1, Tennis data is analyzed by different entities for different use cases. Given the vast amount of Tennis data available over the years, we have identified four different interesting tasks that we hope to explore in this project.

4.1 Match results prediction

For the task of outcome prediction, ATP and WTA data from various datasets were integrated into two bigger datasets. After integration, both datasets included 49 features about the match. Data preprocessing involved handling noisy and missing data, data cleaning to avoid duplicate entries and encoded all the categorical data into numerical data. To have data consistency, matches with 3 sets were chosen and scaled. Some of the earlier works in outcome prediction mentioned that using Markov Chain achieved an accuracy of 0.66 for ATP data[1] and 0.63 for WTA data [9]. Machine learning models are used to predict the match winner while other machine learning models achieved higher scores than

Markov Chain. Once a holdout set for training was identified, we trained Support Vector Machines, Decision Trees, K-Nearest Neighbors, XGBoost and AdaBoost on trh training set. We used K fold cross validation as well. From the validation results, it was inferred that the statistical data for each player, like number of aces, service points, double faults, breakpoints, etc. were the main features involved in predicting the outcome of the match.

4.2 Clustering for player styles

Used clustering techniques, to mine player style groupings from the ATP data (e.g., offensive, defensive, baseline, volley).

The data of the play by play actions in The Jeff Sackmann charting project data a match were condensed to a player summary. The information of each player was aggregated to be the summation statistic of each of the player move characteristics. Data without a clear individual player mapping was thrown out which made 21 of the 37 files in the charting project being included in the aggregation. Some of the 64 parameters for each players performance were:

- aces
- bk_pts
- bp_saved
- crosscourt
- deep
- dfs
- down_middle
- down_the_line
- first_in
- first_pts
- first_pts_won
- first_unret
- inside_in
- inside_out
- serve_return
- snv_pts

The two main methods used were an exploratory data analysis using a PCA dimension reduction with a parameter heat map and a DBSCAN to cluster groups. .

The exploratory data analysis was used to see if an individual feature could be useful in distinguishing between players for a particular style.

The DBSCAN is a method to cluster the players based on a distance metric of the normalized player parameters. The DBSCAN was used on a subset of the parameter data to analyze if a subset of the data would provide unique clusters. Then each of the clusters would of been evaluated to see if there is a relation between of a certain player style in a cluster.

4.3 Matching player rivals

Player match data was aggregated to highlight top rivalries in the ATP tour for the years - 2000 to 2019 (2020 has been excluded since most tournaments after March were suspended due to the coronavirus pandemic). The top 10 rival pairs were identified, and individual player metrics were further explored like - number of tournament finals reached, number of titles won, Grand Slam (GS) wins and evolution of player ranking from 2000 to 2019. Python's Pandas and NumPy libraries were used to conduct the analysis and visualize the results, as shown in Section 7 - Visualizations.

4.4 Evaluation Metrics

The evaluation metrics for each of the four tasks are different which is listed as:

- Predict Match outcomes: have a holdover set for evaluation
- Player Styles: Inspect generated clusters and outliers
- Player Rivals: Manual Evaluation

5 KEY RESULTS

5.1 Match results prediction

To predict the winner of the ATP matches, Jeff Sackmann's data for the years 2000 to 2020 was used to predict the winner of WTA matches, Kaggle's WTA data for the years 2000 to 2016 were used. Some of the earlier works in outcome prediction using Markov Chain achieved an accuracy of 0.66 for ATP data[1] and 0.63 for WTA data [9]. In our model, features like number

of aces, service points, double faults, break points, first serves etc.by each player were used for the prediction. Both ATP and WTA data have been trained and tested on four different machine learning models - Support Vector Machines, Decision Trees, K Nearest Neighbors, AdaBoost, XGBoost. Grid Search CV was used for Hyperparameter Tuning of the above models and SVM predicts the outcome better than other models.
Performance of the different models for winner prediction on the ATP test data is given below:

Table 1: Model performance on ATP test set

Model	Accuracy	Recall	Precision
Decision Tree	0.779690	0.765328	0.785620
SVM	**0.795179**	0.814896	0.784703
AdaBoosting	0.790982	0.789474	0.792641
KNN	0.778474	0.788930	0.772862
XGB	0.792934	0.794636	0.792008

Performance of the different models for winner prediction on the WTA test data is given below:

Table 2: Model performance on WTA test set

Model	Accuracy	Recall	Precision
Decision Tree	0.660168	0.894630	0.608588
SVM	**0.668226**	0.439530	0.817882
AdaBoosting	0.668075	0.441115	0.806697
KNN	0.666115	0.905006	0.611625
XGB	0.667221	0.897548	0.619231

As we can see from the tables above, Support Vector Machines has higher accuracy than the other three model.

5.2 Matching player rivals

Player rival pairs from the ATP circuit, for the years, 2000 to 2019 were identified, and tabulated. Data from the *Tennis Betting Data* (http://www.tennis-data.co.uk/alldata.php) was used for this analysis.

Figure 3 shows the top 10 rivalries, mostly in the *2010* decade, till the year 2019, which has Djokovic-Nadal as the leading rival pair, having faced each other 52 times. Djokovic is leading 28-24 against Nadal. It's interesting to note that the top 3 rivalries are shared between three players in the top 5 seeds, namely, Novak Djokovic(1), Rafael Nadal(2) and Roger Federer(5). It's also interesting to note, that at least one of these players are in every top 10 pair.

Player1	Player2	Players	H2H	Total	Year	Decade
Djokovic N.	Nadal R.	Djokovic-Nadal	28 / 24	52	2019	2010s
Djokovic N.	Federer R.	Djokovic-Federer	26 / 21	47	2019	2010s
Federer R.	Nadal R.	Federer-Nadal	17 / 24	41	2019	2010s
Djokovic N.	Murray A.	Djokovic-Murray	25 / 10	35	2017	2010s
Ferrer D.	Nadal R.	Ferrer-Nadal	6 / 26	32	2019	2010s
Federer R.	Wawrinka S.	Federer-Wawrinka	24 / 3	27	2019	2010s
Berdych T.	Djokovic N.	Berdych-Djokovic	3 / 23	26	2017	2010s
Murray A.	Nadal R.	Murray-Nadal	8 / 17	25	2016	2010s
Federer R.	Murray A.	Federer-Murray	14 / 10	24	2015	2010s
Federer R.	Roddick A.	Federer-Roddick	21 / 3	24	2012	2010s

Figure 3: Top 10 rival pairs from 2000 to 2019

Initial exploratory data analysis (EDA) for the *Tennis Betting Data*, revealed the following observations, with regards to the data fields, as shown in Figure 4

Figure 4 shows the distribution of column values for the *Tennis Betting Data*. This shows fairly populated data across all fields, with sparse data for ***W3, L3, W4, L4, W5***, and ***L5***. These columns represent matches that went on to the third, fourth or fifth set and the data suggests that most matches were completed in two sets.

5.3 Player Styles Prediction

The results of the heat maps shows that the correlations of each of the attributes is highly correlated to a players skill for any single attribute as shown in Figure 18.

Thus the result of the DBSCAN focus on the player skill is an expected result for the each of the player attributes are highly correlated with player skill as shown in Figure 19 which was reduced to two dimension after the clustering.

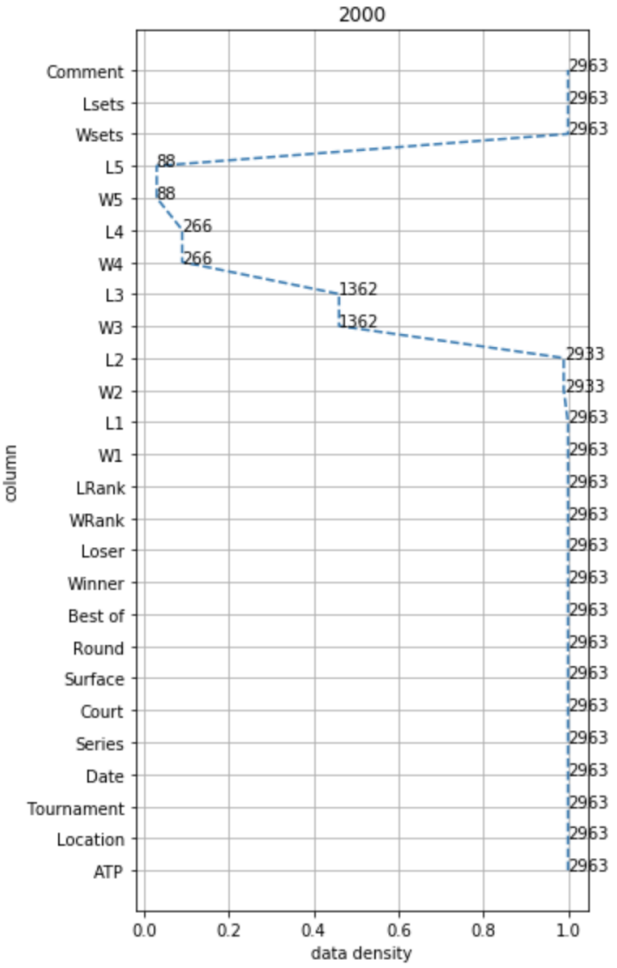

Figure 4: EDA for 2000

The data shows only three particular groups of people in the specific clustering which were The average players, a small cluster of advanced players which includes Novak Djokovik, and the outlier group. The relation of the advanced player group seems rather spurious given of the low minimum cluster size used and the inherent instability of the small cluster.

The dimensional reduction of using UMAP to generate non-linear manifolds shows that the data is still highly focused on player skill as there doesn't appear to be any divergence from a particular manifold as shown in Figure 19.

Thus the results is the cluster is the capture of ordinary and extraordinary players as outliers. The max point distance appears to be a cutoff point on a players skill for the normalized data parameters. Thus the system makes for an adequate outlier detector to remove the exceptional players contextual outlier. Based on domain knowledge, Each of the extraordinary outlier players all have an exceptional result in a particular skill. An example would be Rafael Nadal's exceptional aggregated statistic of unforced forehand.

6 APPLICATIONS

Gathering and analyzing tennis player and match data can give us plenty of valuable insights into deep understanding the skill set of professional competitors. When we clustered the data on different combinations of features, we observed that the outlying data points correlated to the outstanding professionals. Those players that have cemented themselves at the greatest-of-all-timers.

We successfully trained an ensemble of machine learning models to accurately predict the outcome of a match between two hypothetical players. This information would be valuable for online betting platforms or casinos in determining money lines for betting on matches. The gambling industry relies heavily on such information in determining who is favored in any sports betting situation.

Through our exploration and mining of the ATP tennis data, we discovered rivalries among some of the more notable professionals. This information could be leveraged to target marketing for upcoming matches. TV providers and streaming services could use these rivalry scores to inform customers of matches that will be more exciting to watch. This information could also be used to promote content for purchase and thus increase overall revenue for the content providers.

Overall, we've observed that applying data mining and machine learning techniques to historical sports data can lead to valuable inferences used from players to gamblers to content providers. What we've found is only a limited view into the available information hiding in sports statistics. As sports continues to advance in the technologies used, one can only assume that more and more data will continue to be collected. There are

many more patterns of play to be found as the gathering of sports data continues to get more granular.

7 VISUALIZATIONS

Figure 5 shows the players ranked by number of tournaments played and the total number of finals reached from the year 2000 to 2019. Roger Federer leads with a total of 20 unique tournaments played and having reached 116 finals.

Figure 6 shows the players ranked by number of tournaments won from the year 2000 to 2019. Roger Federer again leads with a total of 73 unique tournament wins.

Figure 8 shows the players ranked by most wins at an individual tournament, with Rafael Nadal having won 12 times at the Monte Carlo Masters (*Masters 1000*) and Roger Federer also having 12 titles at the Gerry Weber Open and Wimbledon (*Grand Slam*).

Figure 9, 10, 11 shows the number of tournaments won by Roger Federer, Rafael Nadal and Novak Djokovic at various tournament series from the years 2000 - 2019, respectively.

On analysing *number of wins at Grand Slams*, most Grand Slam tournaments are held by Roger Federer*20*, Rafael Nadal *19* and Novak Djokovic *16*, and Figure 7 shows individual records held by the top players in the ATP men's circuit at the *Grand Slams*, with Rafael Nadal at the French Open(*12*), Roger Federer at the Wimbledon Championship(*8*) and the US Open(*5*) and Novak Djokovic at the Australian Open(*7*).

On analysing the *evolution of player ranks* at the tournaments (Grand Slams and Masters 1000), Figure 12 shows Federer's ATP rank at the time he won the Grand Slam tournament and Figure 13 shows his rank at the Masters 1000 tournaments. Similar analyses was done for Nadal and Djokovic at Figure 14, 15 and Figure 16, 17 respectively.

Player	Tournament	Total_Finals
Federer R.	20	116
Djokovic N.	12	69
Nadal R.	13	60
Murray A.	10	28
Roddick A.	9	27
Ferrer D.	6	17
Tsonga J.W.	4	14
Hewitt L.	5	13
Monfils G.	4	11
Del Potro J.M.	5	11

Figure 5: Most tournament finals reached

Player	Win_Count
Federer R.	73
Djokovic N.	41
Nadal R.	38
Roddick A.	17
Murray A.	16
Ferrer D.	11
Tsonga J.W.	9
Hewitt L.	7
Isner J.	7
Moya C.	7

Figure 6: Most tournaments won

Player	Grand Slam	Count
Nadal R.	French Open	12
Federer R.	Wimbledon	8
Djokovic N.	Australian Open	7
Federer R.	Australian Open	6
Federer R.	US Open	5
Djokovic N.	Wimbledon	5
Nadal R.	US Open	4
Agassi A.	Australian Open	3
Djokovic N.	US Open	3
Murray A.	Wimbledon	2

Figure 7: Most Wins at a Single Grand Slam event

Player	Tournament	Win_Count	Lost_Count	Total_Finals
Nadal R.	Monte Carlo Masters	11	1	12
Federer R.	Gerry Weber Open	9	3	12
Federer R.	Wimbledon	8	4	12
Nadal R.	Internazionali BNL d'Italia	7	2	9
Federer R.	Swiss Indoors	7	4	11
Federer R.	Western & Southern Financial Group Masters	7	1	8
Federer R.	Australian Open	6	1	7
Federer R.	Masters Cup	6	4	10
Djokovic N.	Sony Ericsson Open	6	1	7
Djokovic N.	Wimbledon	5	1	6

Figure 8: Most wins at a tournament

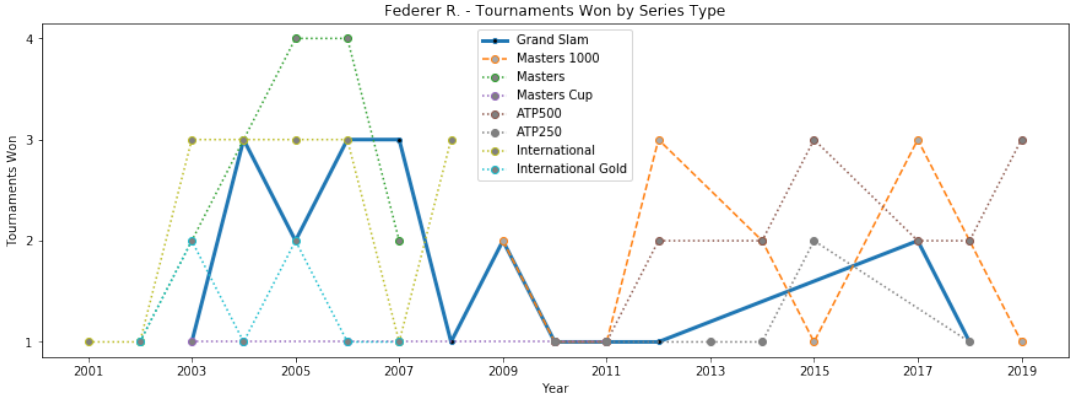

Figure 9: Tournaments Won by Federer (2001 - 2019)

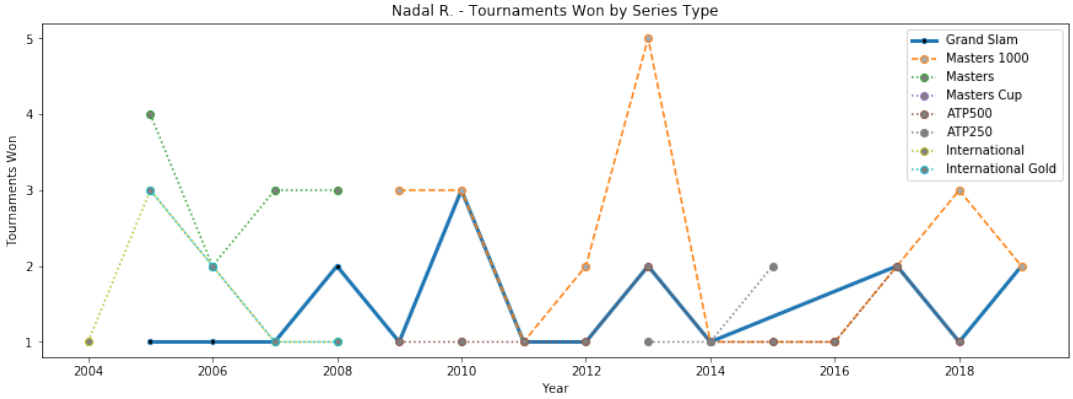

Figure 10: Tournaments Won by Nadal (2004 - 2019)

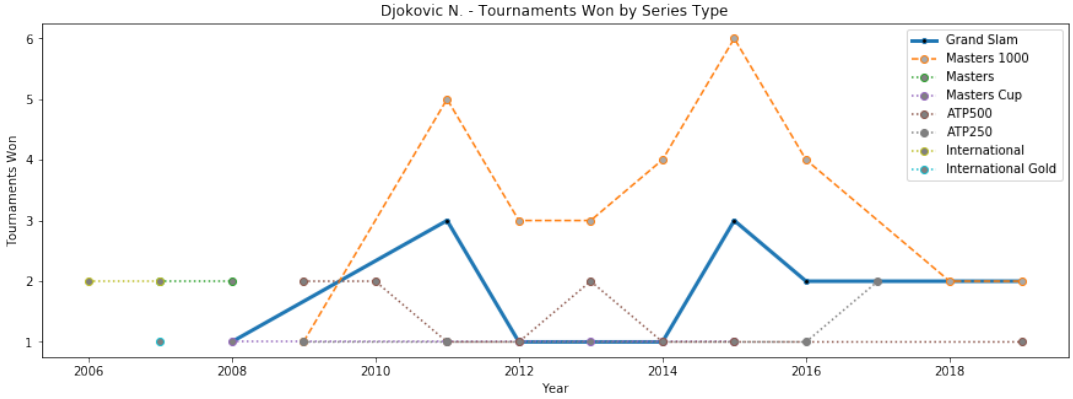

Figure 11: Tournaments Won by Djokovic (2006 - 2019)

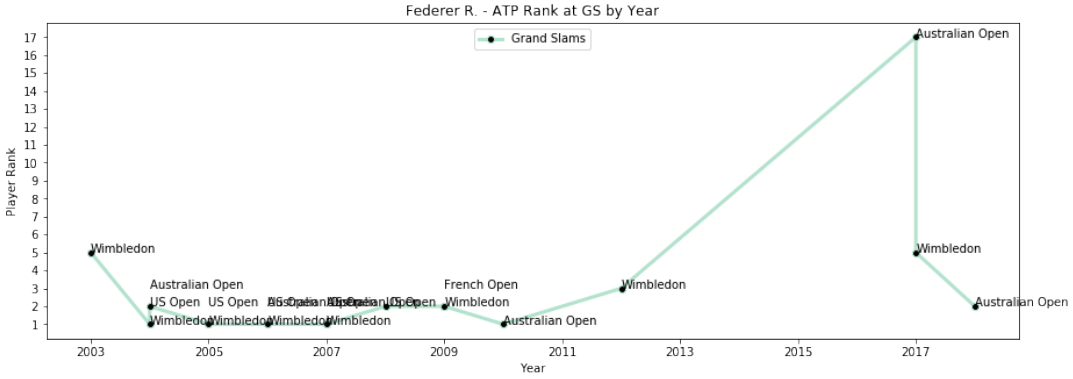

Figure 12: Federer's ATP Ranking at Grand Slams (2003 - 2019)

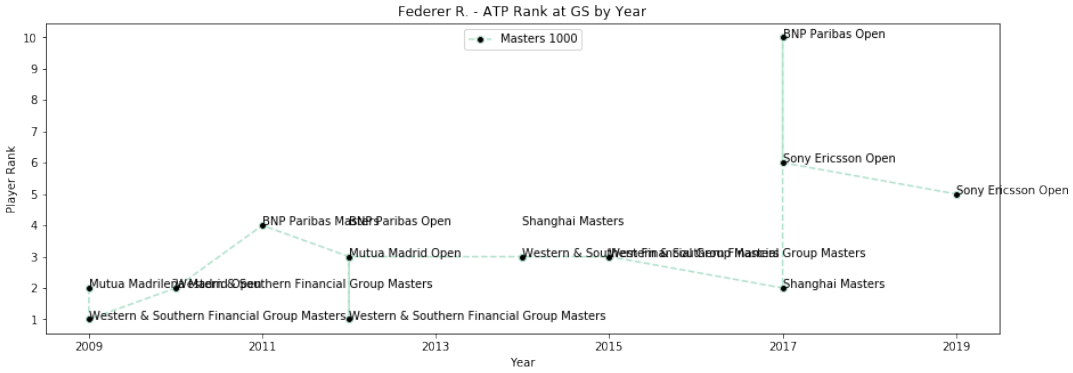

Figure 13: Federer's ATP Ranking at ATP Masters1000 (2009 - 2019)

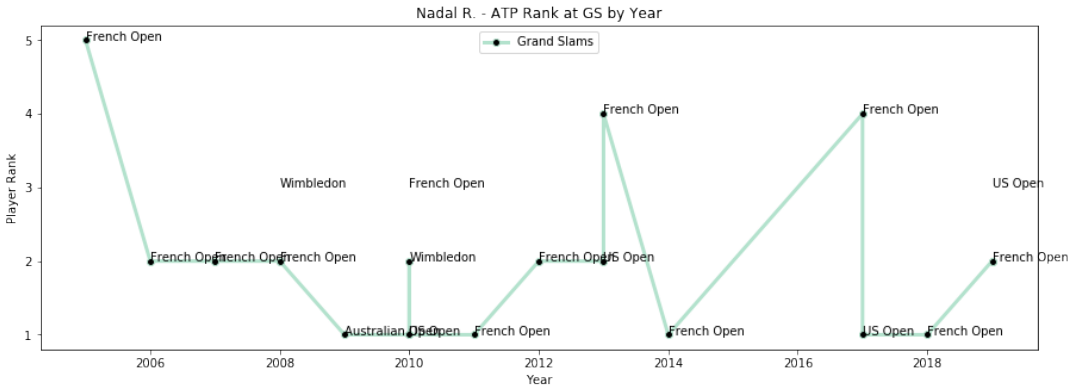

Figure 14: Nadal's ATP Ranking at Grand Slams (2005 - 2019)

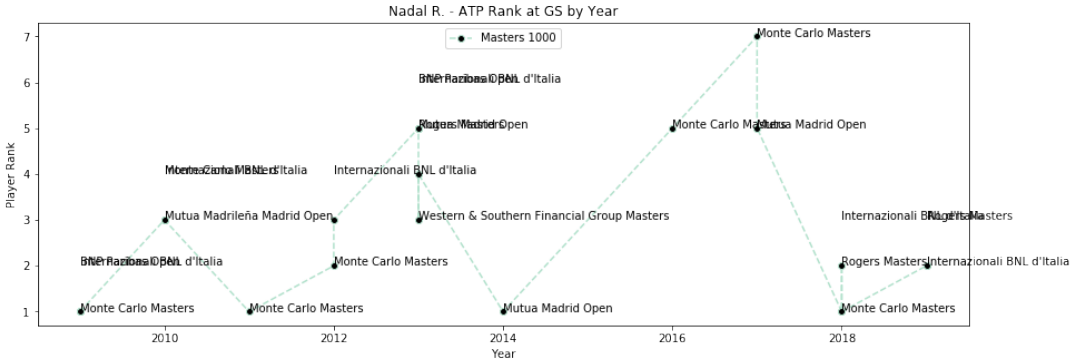

Figure 15: Nadal's ATP Ranking at ATP Masters1000 (2009 - 2019)

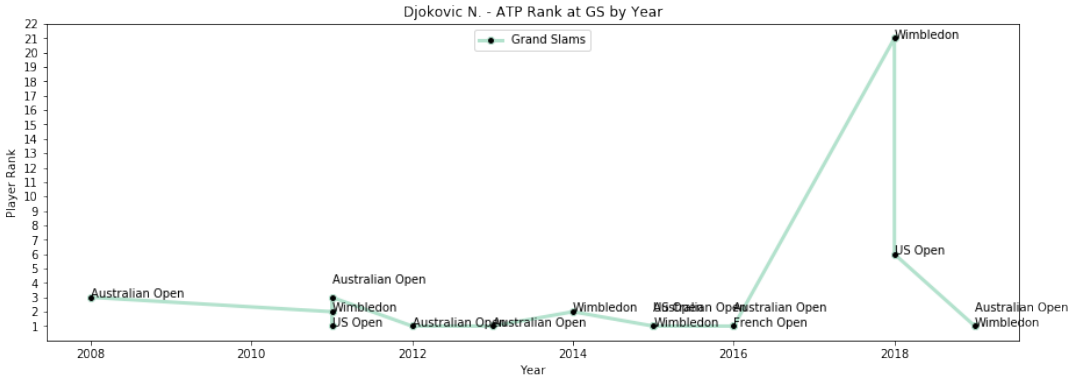

Figure 16: Djokovic's ATP Ranking at Grand Slams (2008 - 2019)

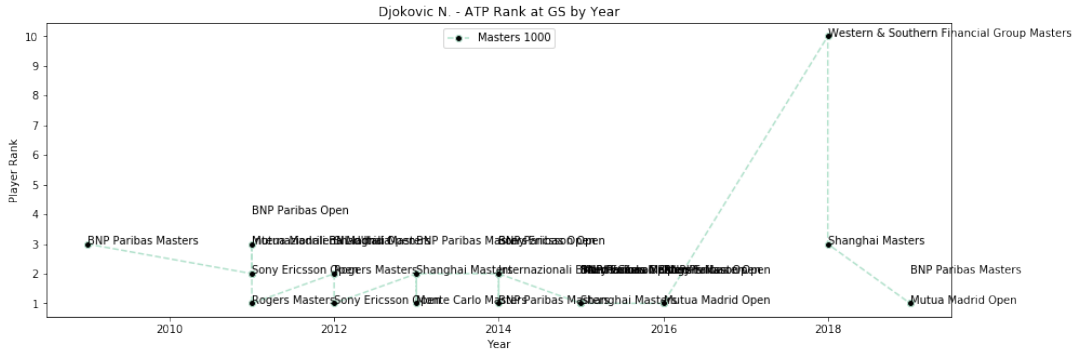

Figure 17: Djokovic's ATP Ranking at ATP Masters1000 (2009 - 2019)

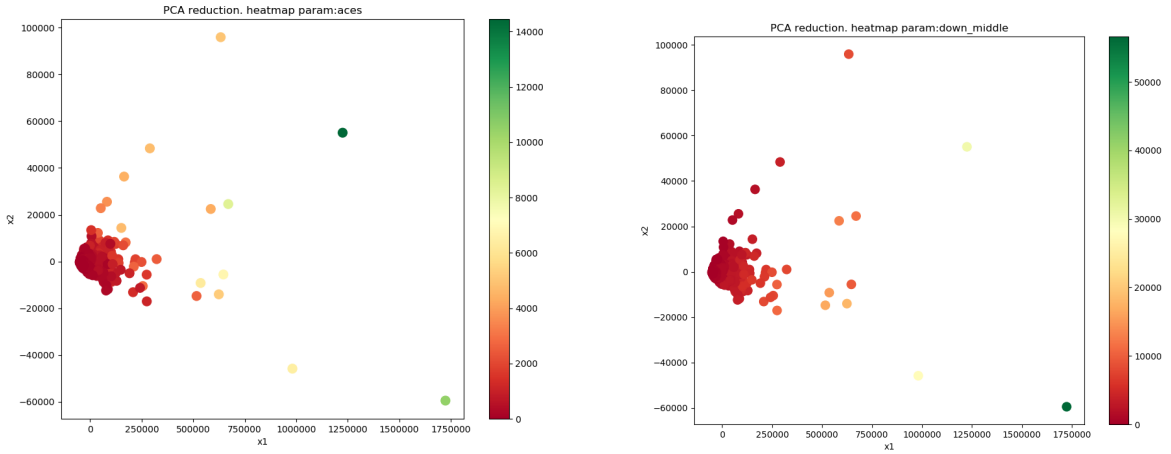

Figure 18: Heat maps of individual params: (left) aces, (right) down middle

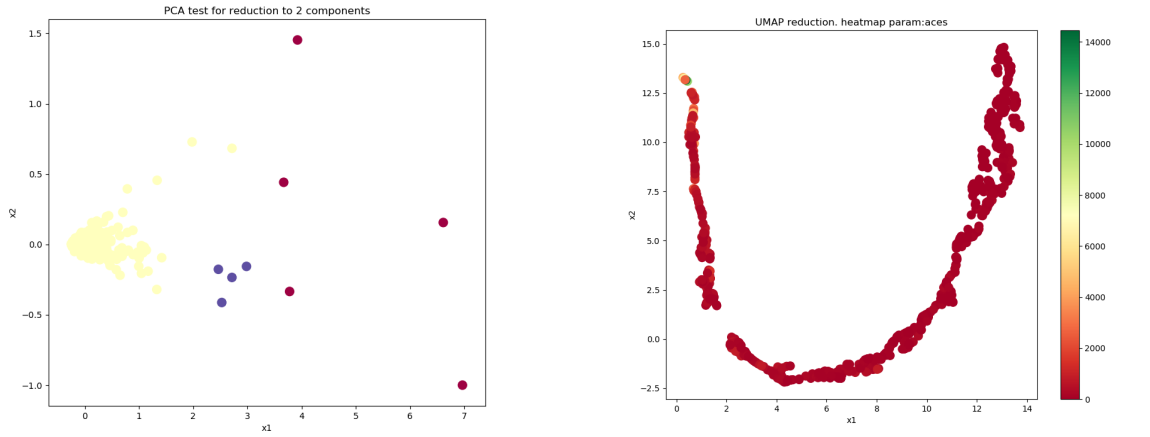

Figure 19: Left: PCA projection after DBSCAN clustering eps=0.2, min points=3, outlier players: Stefan Edberg, Andre Agassi, Rafael Nadal, Andy Murray, and Roger Federer; Right: UMAP projection for aces

REFERENCES

[1] Brown A. Barnett T. and Clarke S. [n.d.]. *Developing a model that reflects outcomes of tennis matches.* Swinburne University. http://strategicgames.com.au/8mcs.pdf.

[2] S.R. Clarke and D. Dyte. 2000. Using official ratings to simulate major tennis tournaments. *International Transactions in Operational Research* 7, 6 (2000), 585–594. https://doi.org/10.1111/j.1475-3995.2000.tb00218.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1475-3995.2000.tb00218.x

[3] Arie-Willem de Leeuw, Aldo Hoekstra, Laurentius Meerhoff, , and Arno Knobbe. 2019. *Tactical Analyses in Professional Tennis.* Stanford University. https://www.researchgate.net/profile/Rens_Meerhoff/publication/335716472_Tactical_Analyses_in_Professional_Tennis/links/5d7790d092851cacdb2e2f10/Tactical-Analyses-in-Professional-Tennis.pdf.

[4] Franc J G M Klaassen and Jan R Magnus. 2001. Are Points in Tennis Independent and Identically Distributed? Evidence From a Dynamic Binary Panel Data Model. *J. Amer. Statist. Assoc.* 96, 454 (2001), 500–509. https://doi.org/10.1198/016214501753168217 arXiv:https://doi.org/10.1198/016214501753168217

[5] William J. Knottenbelt, Demetris Spanias, and Agnieszka M. Madurska. 2012. A common-opponent stochastic model for predicting the outcome of professional tennis matches. *Computers & Mathematics with Applications* 64, 12 (2012), 3820 – 3827. https://doi.org/10.1016/j.camwa.2012.03.005 Theory and Practice of Stochastic Modeling.

[6] Y. Liu. 2001. *Random walks in tennis.* Missouri Journal of Mathematical Sciences, vol.13, no. 3.

[7] Ian McHale and Alex Morton. 2011. A Bradley-Terry type model for forecasting tennis match results. *International Journal of Forecasting* 27, 2 (2011), 619 – 630. https://doi.org/10.1016/j.ijforecast.2010.04.004

[8] Paul Newton and Joseph Keller. 2005. Probability of Winning at Tennis I. Theory and Data. *Studies in Applied Mathematics* 114 (04 2005), 241 – 269. https://doi.org/10.1111/j.0022-2526.2005.01547.x

[9] Joss Peters. 2017. *Predicting the Outcomes of Professional Tennis Matches.* University of Edinburgh. https://project-archive.inf.ed.ac.uk/msc/20172425/msc_proj.pdff.

[10] Daniel Weidner, Martin Atzmueller, and Dietmar Seipel. 2019. *Finding Maximal Non–RedundantAssociation Rules in Tennis Data.* University of Würzburg and Tilburg University. https://arxiv.org/pdf/1909.00985.pdf.

[11] Steven Zheng. 2019. *Modelling ATP Tennis as a Network.* Stanford University. http://web.stanford.edu/class/cs224w/project/26380753.pdf.