Modeling Tennis Matches Using Monte Carlo Simulations Incorporating Dynamic Parameters

Submitted to

PROF. ANIRUDH SINGH RANA

(Department of Mathematics, BITS Pilani, Pilani campus.)

BY

PRAKHAR AGRAWAL	2021B4A70817P
AKSHAJ GUPTA	2021B4A31737P
JAYANT AGGARWAL	2021B4AA2324P

1. Introduction

Tennis features an action-based scoring as well as a mental aspect that makes it difficult to predict match results. Our study uses Monte Carlo simulation for modelling tennis matches while taking into account specific scenarios such as player fatigue along with the momentum shifts generated from a service break against an opponent. This paper applies some basic statistical techniques including data preprocessing, and simulations to analyse match prediction for Roger Federer and Novak Djokovic, two of the greatest tennis players in history.

Python was used for preprocessing while R was used for the simulation. Two R scripts (basic and improved versions) were employed for simulating match outcomes. The simulations include constant and variable probabilities derived from historical tennis data sets. The results of the simulation were interpreted in the context of the paper and checked against actual events.

2. Tennis Scoring System

The hierarchical scoring system in tennis progresses from points to games, sets, and matches. Players win games by scoring at least four points with a two-point margin. To win a set, a player must win at least six games and lead by at least two. If the set reaches a 6–6 tie, a tiebreak is played to determine the winner. Matches are typically played as best-of-three or best-of-five sets. Dynamic factors such as fatigue, which affects player performance in prolonged matches, and momentum, which can shift after critical points like a break of serve, add complexity to predicting outcomes. These factors were integral to the simulations conducted in this study.

3. The In-Play Dataset

The in-play dataset, which records point-by-point details for tennis matches, was used for this project. This dataset includes information on server identity, score progression, point winners, and key events such as break points and their conversions. It is ideal for dynamic modeling because it provides the granularity needed to calculate probabilities for serve win, return win, and break conversion rates. Compared to play-by-play datasets, which capture even finer details such as shot type and ball placement, the in-play dataset is sufficient for modeling the discrete scoring structure of tennis. It is particularly useful for implementing dynamic adjustments related to fatigue and momentum, making it the backbone of this study.

The project consists of 3 main stages:

- Data cleaning and extraction of relevant player statistics from the dataset
- Implementing the calculated statistics in custom versions of the Monte Carlo simulation to predict the winner of tennis matches that already took place in real life
- Analyzing the results.

4. Data Preprocessing

Data preprocessing was conducted in Python to prepare the dataset for simulations. Match and point-level data for Grand Slam tournaments from 2011 to 2024 were loaded. The data was filtered to include only singles matches involving Roger Federer and Novak Djokovic. Cleaning steps included standardizing player names to resolve non-ASCII issues and removing incomplete or irrelevant entries, such as doubles matches. The cleaned dataset was then used to calculate serve win probabilities, segmented into non-fatigue (sets 1–2) and fatigue (sets >2) conditions. Similarly, return win probabilities and break conversion rates were computed. Bar plots were generated to visualize performance differences under fatigue and non-fatigue conditions, providing insights into the dynamics of both players and forming the foundation for simulations. The probability calculation is shown below

P(Serve Win)= Points Won on Serve / Total Points Served

P(Return Win)= Points Won While Returning / Total Points Faced While Returning /

P(Break Conversion)= Break Points Won / Break Points Attempted

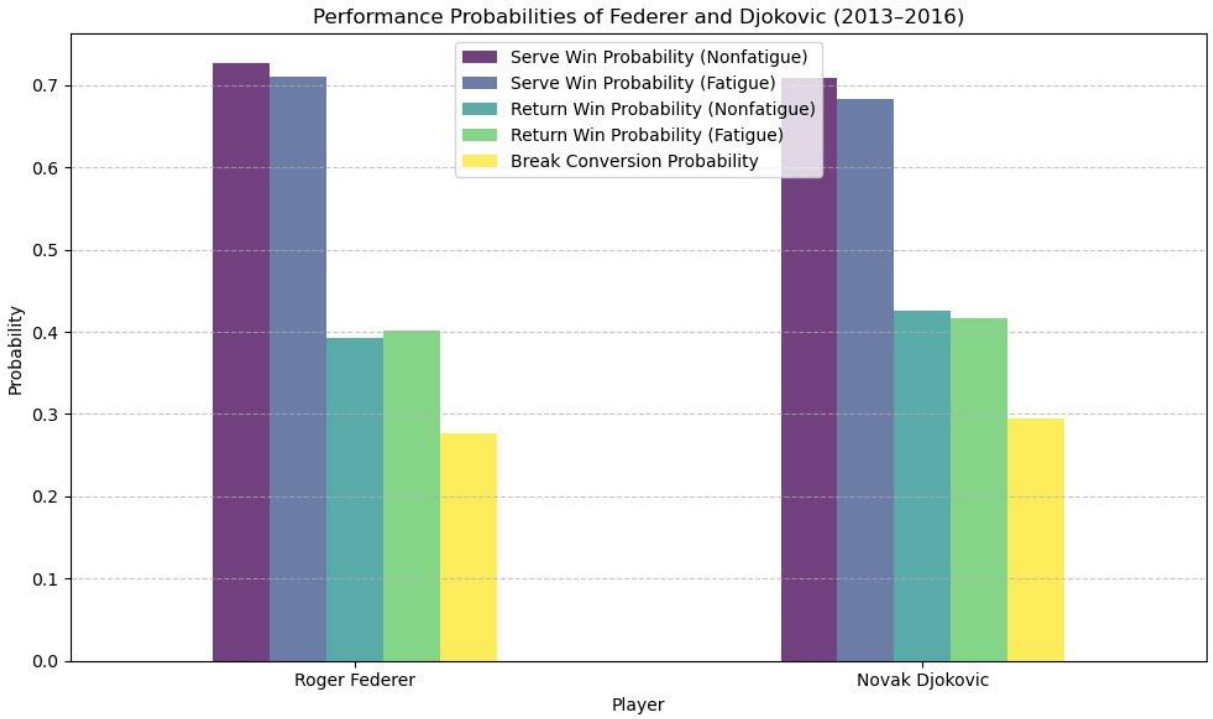
P(Fatigue Serve Win)=P(Serve Win in Sets > 2)=

Points Won on Serve in Later Sets / Total Points Served in Later Sets

P(Fatigue Return Win)= P(Return Win in Sets > 2)=

Points Won While Returning in Later Sets / Total Points Faced While Returning in Later Sets

Bar graph showing the probabilities for the 2 players




```
Final Results (Formatted Table):
              Serve Win Probability (Nonfatigue)  Serve Win Probability (Fatigue)  Return Win Probability (Nonfatigue)  Return Win Probability (Fatigue)  Break Conversion Probabi
ity
Player

Roger Federer                           0.726004                         0.710186                            0.392872                          0.401150                0.27
193
Novak Djokovic                          0.708051                         0.682646                            0.425059                          0.416773                0.29
302
```

5. Simulation Processes

The implementation of the simulations was done via two R scripts. A rudimentary scheme (tennis_simulation.R) operated, and an improved simulation scheme (tennis_simulation_improved.R) ran subsequently. Both simulated matches point by point, the probabilities determined from historical data, with the improved scheme allowing for dynamic adjustments so as to appear more realistic.

The two simulations paid particular attention to the modeling of serve probabilities. Serve win probabilities were based on the initial calculations for tiny and large servers. If the first serve failed to give the point to the server, then the opponent's return win probability was built in for the second serve. All these factors acknowledge that the returner would thus increase his influence in a rally. Such intricacies explain why the improved script applies greater realism by accounting for the respective tactical differences between Alpha-type and Beta-type serves. In contrast, the basic simulation applied serve win probabilities statically to every point: demonstrating ignorance toward these variations.

In the basic simulation, static serve and return probabilities were used without consideration of fatigue or momentum changes. Matches were simulated on a point-by-point basis, determining the outcome by means of random sampling; this was performed 400 times for a given match to estimate a likely winner. Although quick, such practice lacks the dynamic adaptability that would be required to replicate true match dynamics.

In the improved model, as the improvements to some extent already involved fatigue, it decreased the server's probability of winning the serve after the second set to reflect the decline in performance once a match gets longer in duration. It also modeled momentum effects, increasing probabilities for players who broke their opponent’s serve to simulate the psychological advantage gained. Combined with the differentiation between first and second serves, these adjustments allowed probabilities to evolve during the match, making the improved model more reflective of real-world tennis scenarios.

6. Results and Analysis

The output from the simulations drew critical insights into the match dynamics between Federer and Djokovic. The primary R-script gave two sets of results. When fatigue was incorporated into the models, Djokovic was suggested to have a win probability of 72.98% while Federer had a mere 27.02%. The statistics attested that he has superior stamina, similarly supportive of his convincing historical performance in long and intensely exhausting matches. Without including fatigue factors, it was found that Federer had a superior winning percentage of 53.29%, while the chances of winning for Djokovic were put at 46.71%. It further depicts Federer ahead in matches with regards to fatigue considerations being low.

The other optimized R-script, which incorporated fatigue and momentum, brought two models back on a closer basis. The exit generally suggests the battle for match union, with Djokovic winning 50.72% of matches and Federer winning 49.28%. Noteworthy is that this balancing act of results suggests the importance of momentum against fatigue. Federer was able to show amazing persistence and leverage breaks and psychological pressure after key points; because of this, he made a comeback against Djokovic, replicating live scenarios when both players adjusted their tactics based on the momentum of the match. The results confirm the theoretical framework of the paper and emphasized the need for dynamic parameters to operate a more true simulation.

The simulation results agree well with the real-world performance of Federer and Djokovic. With superior fitness and stamina, Djokovic has an upper hand over Federer in long rallies. However, in a physical contest, using momentum helps him close the gap against Djokovic when he plays aggressively and commands a break at defining times. The results demonstrate that while static models can provide baseline predictions, dynamic parameters such as fatigue and momentum are essential for capturing the complexities of tennis matches.

Result of basic simulation

```
result                  List of  2
  $ With_Fatigue   :List of 4
   ..$ Djokovic_Wins          : int 7298
   ..$ Federer_Wins           : int 2702
   ..$ Djokovic_Win_Percentage: num 73
   ..$ Federer_Win_Percentage : num 27
  $ Without_Fatigue:List of 4
   ..$ Djokovic_Wins          : int 4671
   ..$ Federer_Wins           : int 5329
   ..$ Djokovic_Win_Percentage: num 46.7
   ..$ Federer_Win_Percentage : num 53.3
```

Result of improved simulation

```
result                  List of  4
  $ Djokovic_Wins          : int 5072
  $ Federer_Wins           : int 4928
  $ Djokovic_Win_Percentage: num 50.7
  $ Federer_Win_Percentage : num 49.3
```

7. Conclusion

The project extensively established Monte Carlo simulations that modeled the outcomes of tennis matches in the presence of dynamic factors that include fatigue and momentum, with discrete differences and constraints. A Python preprocessing script calculated serve, return, and break probabilities efficiently from the roll-on dataset whilst R simulations modeled matches occurring under static and dynamic conditions. The new matron provided much more realistic match-play considering both mental and physical aspects. This underscores the importance of not losing sight of both the static and dynamic parameters to enable accurate predictions, even though tennis by its nature faces significant jumps in unpredictability. Future work could include any number of additional factors, such as player surface preferences and environmental conditions, to identify improvements to the models. The project demonstrates the promise of data-driven methods toward sports analytics and offers the same opportunities to researchers, coaches, and tennis fans.

8. Reference

- J. Krčadinac *et al*., "Modeling Tennis Matches Using Monte Carlo Simulations Incorporating Dynamic Parameters," *2023 46th MIPRO ICT and Electronics Convention (MIPRO)*, Opatija, Croatia, 2023, pp. 1281-1286, doi: 10.23919/MIPRO57284.2023.10159731.