Mathematical Modeling of Scoring Rates in Tennis Matches

Yawen Li*

Beijing No'4 High School International Campus, Beijing, China

*Corresponding author: amyliyawen@ldy.edu.rs

Abstract. This paper delves into the profound implications of scoring rates in tennis matches through the application of mathematical modeling. By gathering data from diverse sources, it thoroughly scrutinizes the intricate interplay between scoring rates, including first-serve percentage, second-serve percentage, and double faults, and their effects on match outcomes. The study uncovers striking performance differentials among top-tier players. Notably, the analysis reveals that first-serve percentage exhibits a relatively weak correlation with overall match performance. Conversely, the percentage of first-serve points won significantly and positively influences match outcomes, highlighting its pivotal role. Conversely, higher percentages of second-serve points won and an increased incidence of double faults are associated with decreased success rates in matches. Through the utilization of a multiple linear regression model, this research can elucidate 46% of the variance in match outcomes, underscoring the model's substantive value. The implications of this study extend far and wide within the tennis community, enriching our comprehension of the intricate dynamics of player performance, strategic decision-making, and the practical applications of sports science. The integration of mathematical modeling into tennis promises to be a transformative force, with the potential to reshape the way to understand and practice the sport at all levels, paving the way for future research and advancements in the field.

Keywords: Tennis; mathematical modeling; scoring rates; performance analysis; sports analytics.

1. Introduction

Tennis, as a widely popular sport, attracts a substantial audience and participants. Therefore, a comprehensive study of scoring rates in tennis matches holds practical significance. In traditional tennis matches, scoring rates are influenced by factors such as player performance and skill levels. However, mathematical modeling offers a deeper understanding, aiding in the explanation of variations and trends in scoring rates [1]. The use of mathematical models allows for better prediction of match outcomes and provides players, coaches, and spectators with more information to enhance their understanding of match progress and tactical choices. Research in this field can also provide valuable insights for improving tennis matches, including refining game rules, event organization, and player training [2]. Recent advancements in big data and computational capabilities have expanded opportunities and resources for mathematical modeling, enabling researchers to conduct more comprehensive analyses of scoring rates in tennis matches [3].

Furthermore, the impact of training methods and hand-eye coordination on the accuracy of forehand sidespin table tennis service was revealed, showing that distributed practice was more effective for individuals with high coordination, while those with low coordination could benefit from both training methods [4]. This can also be analyzed in parallel with the technical aspects of tennis training.

According to the stochastic model using a discrete-time Markov chain to analyze a dataset of professional tennis matches, it was found that while the assumption of identical point-winning probabilities for the server is not valid, independence is generally valid, except for specific score situations, providing insights into the importance of each point in winning a game based on court surfaces and player groups [5]. The model offers insights into game-winning probabilities and point importance in tennis matches. In this paper, the relationship between the success rate of serves and scoring rates can be further explored [6].

To clarify a player's serving advantage, you can calculate the overall serve-win rate. According to O'Donoghue's algorithm [7], the serving win probability can be calculated using the formula $p1 *$

$q1 + (1 - p1) * p2 * q2$, where p represents the first-serve success rate, and q is the conditional probability of winning points on a serve after a successful serve. 1 and 2 correspond to the first serve and second serve, respectively. Moreover, to analyze tennis serving and serve-return strategies on slow surfaces, providing insights into the performance and tactics of elite tennis players under these conditions [8], it's also important for the article to promote the analysis of the scoring rate in tennis matches.

2. Methods

2.1. Data Source

The data for this study is sourced from various outlets, including professional tennis match records, official match websites, and historical match data archives. Sample selection is based on several criteria, including the match level (e.g., Grand Slam, ATP Tours, WTA Tours), match type (singles or doubles), data availability of match performance, and the time frame of matches (historical data or recent years).

2.2. Sample Selection

Data collection primarily encompasses two aspects of information, namely the scoring rates in serves and statistical data related to match outcomes. Scoring rates in serves include statistics for different scoring states within a service game, such as first-serve percentage, second-serve percentage, and double faults. Statistical data regarding match outcomes include player wins and losses in matches, including match scores and match duration [9].

2.3. Method Introduction

Multiple Linear Regression: To establish the relationship between scoring rates in serves and match outcomes, multiple linear regression analysis is employed. In the regression model, scoring rates in serves are considered independent variables, while match outcomes are treated as the dependent variable. Through regression analysis, researchers can estimate the impact of different scoring rates on match outcomes and develop mathematical models for predicting match results [10].

Statistical Significance Testing: In the context of multiple linear regression analysis, statistical significance tests are conducted to determine whether different scoring rates significantly affect match outcomes.

3. Results and Discussio

First, in Table 1 analyze each player's performance by calculating the number of matches they have won. Players who have won fewer than ten matches have been excluded, leaving data for the remaining players.

In Table 1, it can be seen that the top 15 players have won at least 31 matches, while the top 10 players have won at least 36 matches. Additionally, the top five players among the Top 15 have won 50% more matches than the bottom five players.

Table 1. The rank of the winner's name and count

Rank	winner_name	Count of winner_name
1	Carlos Alcaraz	58
2	Daniil Medvedev	56
3	Taylor Fritz	49
4	Jannik Sinner	45
5	Novak Djokovic	45
6	Andrey Rublev	45
7	Alexander Zverev	41
8	Stefanos Tsitsipas	39
9	Frances Tiafoe	37
10	Alex De Minaur	36
11	Hloger Hurkacz	35
12	Tommy Paul	34
13	Francisco Cerundolo	34
14	Cameron Norrie	33
15	Casper Ruud	31

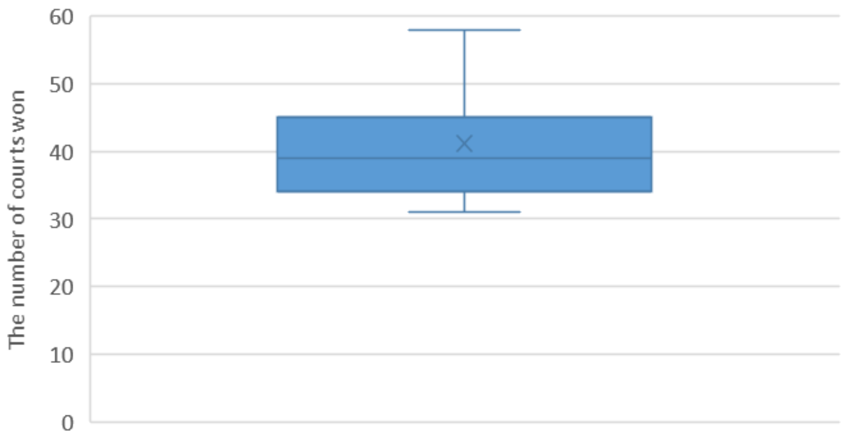

Fig. 1 ATP 2023 Map of the number of counts won by top 15 players

In Figure 1, to analyze the distribution of the number of courts won by the top 15 players, the performance of the top players is highly dispersed, indicating that their performance significantly surpasses that of other players. The median value is also relatively lower and not positioned in the middle. As a result, the field is dominated by a minority of players.

Based on the performance of the players presented in the previous charts, it can continue to study their serving. After calculations, it can be determined the success rates of the first and second serves for each player (serving success means getting the serve into the opponent's service box). Simultaneously, to calculate the scoring rates based on the success of the serve. Below are the data for the Top 60 players.

Using this data, it can begin to analyze the relationship between tennis service and match performance. The first correlation under investigation is the relationship between the 1stIn success rate and match performance.

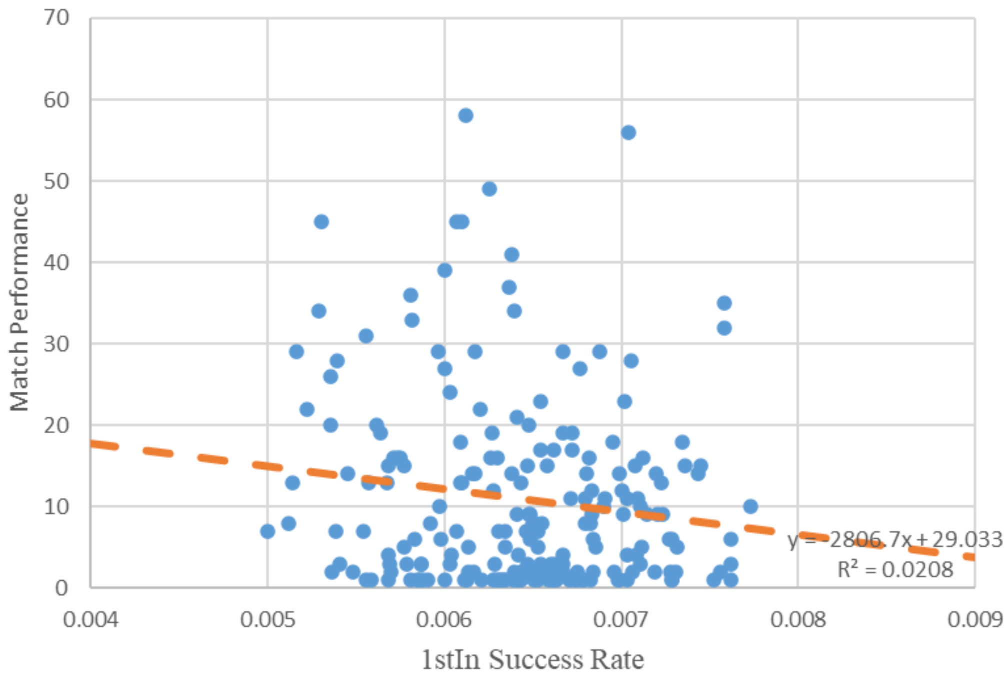

Fig. 2 The relationship between 1stIn success rate and match performance

From the Figure 2, it can be observed that the 1stIn success rate has little correlation with match performance. This paper will continue to study the second correlation, analyzing the relationship between the 1stWon success rate and match performance.

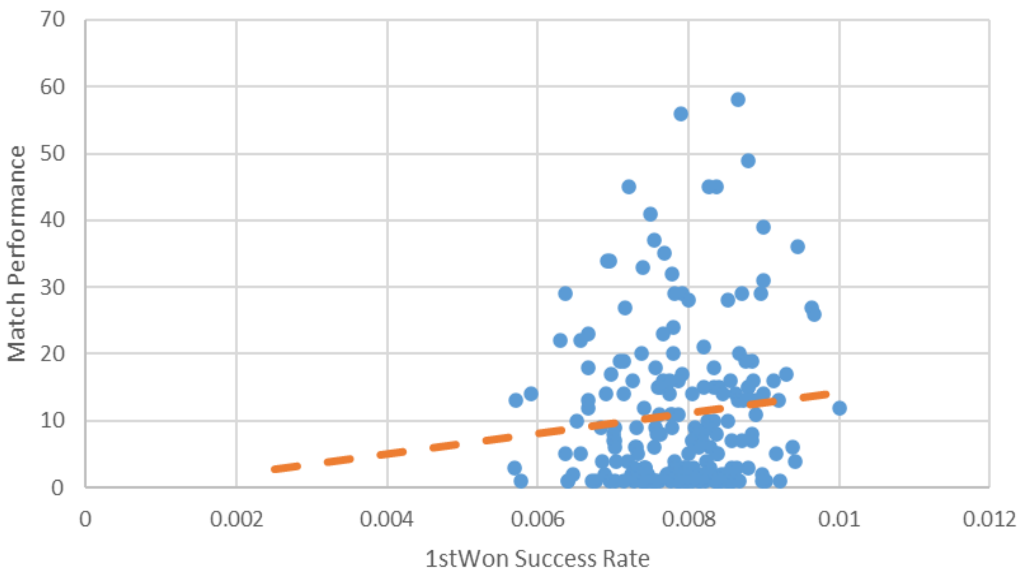

Fig. 3 The relationship between 1stWon success rate and match performance

In the Figure 3, it shows that the relationship between these two variables is not evident on the chart. This could be attributed to two potential reasons: First, the data is limited to matches of high-level players, who generally have high-quality serving. Second, it might be worthwhile to change the analytical perspective and further explore the probabilistic aspect to potentially gain more insights.

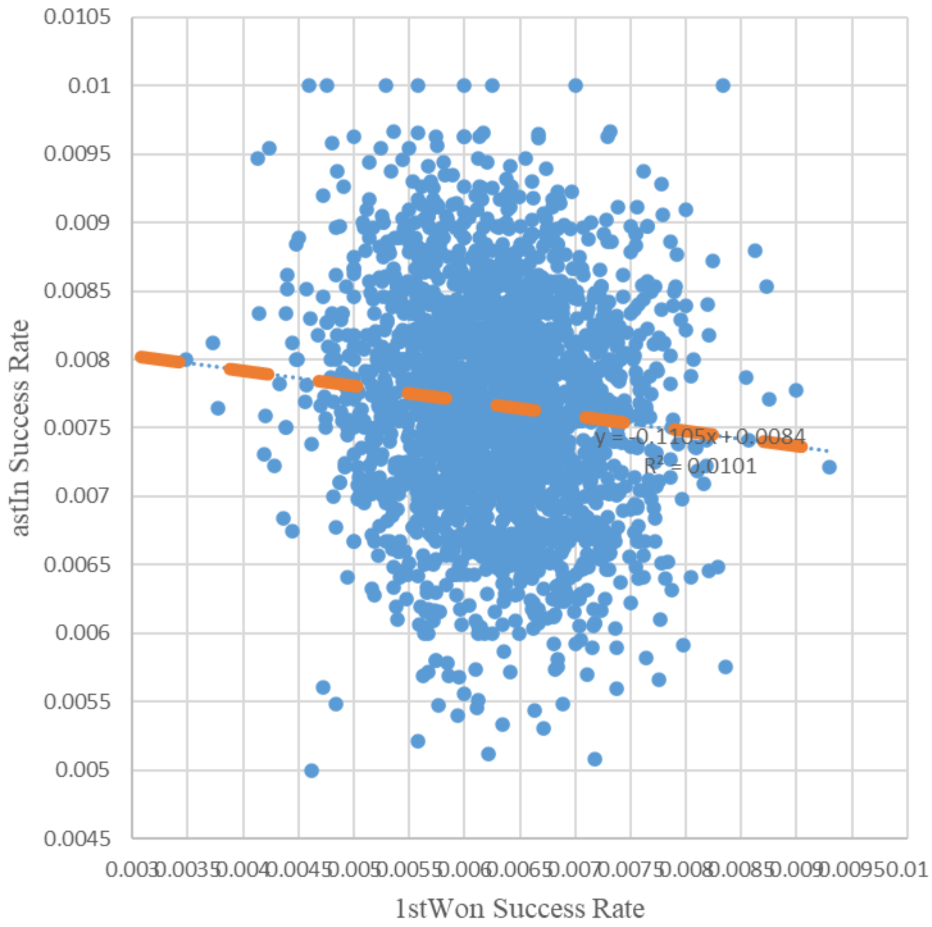

Fig. 4 The relationship between 1stWon success rate and 1stIn success rate

In Figure 4, these two variables show a negative correlation, meaning that as the 1stIn success rate increases, the 1stWon success rate decreases.

Table 2. The OSL model of the data

data name	coef	std err	t	P>|t|	[0.025	0.975]
const	1.161	0.008	149.871	0	1.146	1.176
w_1stIn	-0.006	0	-12.669	0	-0.007	-0.005
w_1stWon	0.0031	0.001	4.72	0	0.002	0.004
w_2ndWon	-0.0056	0	-11.699	0	-0.007	-0.005
w_df	-0.0047	0.001	-3.792	0	-0.007	-0.002

In Table 2, the R-squared measures the goodness of fit of the model. It indicates the proportion of the variance in the dependent variable (win_ratio) that can be explained by the independent variables (w_1stIn, w_1stWon, w_2ndWon, w_df). In this model, the R-squared is 0.460, suggesting that the model can explain approximately 46% of the variance in the dependent variable. This indicates a relatively good fit for the model.

The F-statistic serves as a pivotal tool for assessing the overarching significance of the model. In the context of this model, the F-statistic assumes a value of 483.4, and the associated p-value is observed to be exceedingly close to zero, measuring at 1.33e-301. This statistical information collectively implies the overarching significance of the model. Specifically, it suggests that there exists at least one independent variable with a substantial and statistically significant influence on the dependent variable.

Moving on to the examination of the coefficients associated with each independent variable, these values provide insight into the magnitude of the effects exerted by these variables on the dependent variable. Within this model, the coefficients for the individual independent variables are as follows:

Firstly, the coefficient for 'w_1stIn' is -0.0060. This signifies that for every unit increase in the first serve in percentage (w_1stIn), the win ratio (win_ratio) experiences an average decrease of 0.0060 units. Secondly, the coefficient for 'w_1stWon' is 0.0031. This denotes that for each incremental unit added to the first serve won percentage (w_1stWon), the win ratio exhibits an average increase of 0.0031 units. Thirdly, the coefficient for 'w_2ndWon' stands at -0.0056, implying that for every unit increase in the second serve won percentage (w_2ndWon), the win ratio endures an average decrease of 0.0056 units. Lastly, the coefficient for 'w_df' is -0.0047, indicating that for each unit increase in the double faults percentage (w_df), the win ratio encounters an average decrease of 0.0047 units.

Turning attention to the significance of these coefficients, as assessed by the p-value denoted as 'P>|t|,' it is crucial to note that within this model, all the independent variables exhibit p-values that closely approach zero. This signifies that the effects of these variables on the win ratio are statistically significant. In summary, this model yields several insightful conclusions:

Firstly, an increase in the first serve in percentage (w_1stIn) and second serve won percentage (w_2ndWon) is correlated with a negative impact on the win ratio, indicating a negative correlation between these variables. Conversely, an increase in the first serve won percentage (w_1stWon) is associated with a positive impact on the win ratio, signifying a positive correlation. Lastly, an increase in the double faults percentage (w_df) is also linked to a negative impact on the win ratio, revealing a negative correlation between these variables.

4. Conclusion

The research presented in this essay highlights the significance of mathematical modeling in understanding and analyzing scoring rates in tennis matches. Tennis, as a popular sport with a substantial following, benefits from mathematical models that provide deeper insights into the dynamics of the game. The research focused on the relationship between scoring rates and match outcomes and utilized multiple linear regression analysis to achieve its objectives.

The findings from the multiple linear regression analysis revealed valuable insights into the impact of scoring rates on match outcomes. The model's coefficients provided specific information about each scoring rate's effect on the win_ratio (a measure of match success). Notably, the intercept coefficient (const) indicated that when all independent variables are zero, the estimated value of win_ratio is 1.1610. The coefficient for w_1stIn (First Serve In Percentage) revealed that for every 1-unit increase in first serve in percentage, win_ratio decreases by 0.0060 units. On the other hand, w_1stWon (First Serve Points Won Percentage) had a positive effect, with a 1-unit increase resulting in an increase of 0.0031 units in win_ratio. In contrast, w_2ndWon (Second Serve Points Won Percentage) had a negative impact, with a 1-unit increase leading to a decrease of 0.0056 units in win_ratio. Similarly, w_df (Double Fault Percentage) had a negative influence, with a 1-unit increase causing a decrease of 0.0047 units in win_ratio.

The statistical analysis also provided insights into the distribution of residuals and potential issues within the model. The Omnibus statistic indicated that the distribution of residuals was not normal, which is essential to consider in further model development. The Durbin-Watson statistic suggested low autocorrelation among residuals, and the Jarque-Bera statistic confirmed the non-normality of residuals. Additionally, skewness and kurtosis values provided further insights into the residual distribution, and the condition number hinted at potential multicollinearity issues in the model.

The research conducted in this essay carries significant implications for various stakeholders in the world of tennis, including players, coaches, analysts, spectators, and even those involved in sports betting. It enhances the understanding of tennis dynamics, performance analysis, strategy optimization, injury prediction, spectator engagement, and tennis education. Furthermore, the research contributes to the broader field of sports science, potentially leading to insights that can be applied to other sports as well.

In summary, the mathematical modeling of scoring rates in tennis matches is a promising avenue for further research and development, with the potential to revolutionize the way tennis is understood and played at all levels.

References

[1] Safari Indra, Saptani Entan. Metode latihan dan koordinasi mata tangan meningkatkan akurasi forehand sidespin service tenis meja. Jurnal Keolahragaan, 2019, 7.

[2] Joseph Beerman, Lawrence Sher. Improve Tennis Service through Mathematics. Journal of Physical Education and Recreation, 2013, 52: 55.

[3] Stephanie Ann Kovalchik. Searching for the GOAT of tennis win prediction. Journal of Quantitative Analysis in Sports, 2016, 12: 127-138.

[4] Zainan A W, Ruslan N A S. Exploring the Importance of Players' Characteristics and Performance on Serve and Return of Serve in Winning the Women's Singles Grand Slam Tennis Tournaments. Journal of Physics Conference Series, 2020, 14: 12008.

[5] Sim Min Kyu, Choi Dong Gu. The Winning Probability of a Game and the Importance of Points in Tennis Matches. Research quarterly for exercise and sport, 2019, 91: 1-12.

[6] Cooper Curtis, Kennedy Robert. Expected length and probability of winning a tennis game. The Mathematical Gazette, 2021, 105: 490-500.

[7] Gillet E, et al. A Notational Analysis of Elite Tennis Serve and Serve-Return Strategies on Slow Surface. Journal of Strength and Conditioning Research, 2009, 532–539.

[8] Donoghue P, Ballantyne A. The impact of speed of service in Grand Slam singles tennis. Science and racket sports, 2004.

[9] Jin Shengzhen, Liu Zhanfeng, Guo Zhiguang. A study on the differences in competitive performance characteristics between young elite tennis players and professional tennis players. Journal of Wuhan Institute of Physical Education, 2023, 57(2): 91-100.

[10] Chen Jingcheng. Research on the Development Path of Tennis Industrialization in China in the New Era. Journal of Guangzhou Institute of Physical Education, 2021, 41(5): 3.