Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 3
CECS 2023

Quantifying Tennis Player Performance: A Linear Regression Approach

Yuxi Zeng[1,*], Siwei Zhong[2]

[1] School of Information Management, Jiangxi University of Finance and Economics, Nanchang, Jiangxi, 330006, China

[2] International College, Jiangxi University of Finance and Economics, Nanchang, Jiangxi, 330006, China

* Corresponding Author

Abstracts. This paper uses linear regression to quantitatively analyse the performance of players in the men's singles competition at Wimbledon 2023. Firstly, the data is processed by observationally analysing the match data to ensure compliance with the tournament standards and regulations. Next, key metrics were extracted, including short-term and long-term metrics, as well as the introduction of Serve Indicator to consider the impact of serve advantage on player performance. Then, the most important independent variables were identified through Random Forest feature analysis and parameters were calculated using least squares to construct performance indicators for use in linear regression. Finally, through data visualisation and analysis, it was found that player 1 usually performs better at critical moments, showing greater stability and consistency, while player 2 shows greater variability and unpredictability. Overall, the linear regression method in this paper is valuable and practical for quantifying tennis players' performance, and can provide a reference for players and coaches to help them better analyse and improve their performance.

Keywords: Quantifying Tennis Player Performance; Linear Regression; Random Forest Feature Analysis; Least Squares Method; Athlete Performance Evaluation.

1. Introduction

In this study, we quantitatively analysed the performance of tennis players through linear regression methods. Firstly, we carried out an observational analysis of the data from the featured matches in the men's singles at Wimbledon 2023 and performed a series of pre-processing on the data, including standardising the form of the data of interest, filling in missing values, correcting for outliers, and classifying the data [1]. We then extracted and summarised the data for each player in each match based on short and long term metrics using Excel, and then calculated several key indicators of 'performance' such as consecutive wins, points lead, error rate and serve indicator.

In the model construction section, we firstly identified the most important independent variables through the Random Forest analysis, and then used the least squares method to calculate the parameters to obtain a linear regression model of the athletes' performance. Through the Random Forest, we selected eight variables to test, and finally chose the six most important variables to construct the athlete's "performance" index [2]. The data were standardised to eliminate the effect of dimensionality between characteristics.

The analysis of the results is based on the data from the Wimbledon 1301 tournament in 2023, where the performance of the athletes is visualised in terms of points and board units. The results show that player 1 outperforms player 2 at the beginning and end of the match, but there are a few situations during the course where player 2 outperforms player 1. This is consistent with the problem description and suggests that our model is appropriate. Player 1 usually performs better at key moments, and although player 2 outperforms player 1 in five phases, this outperformance is not sustained [3]. Whenever the two performances were close, Player 1 usually reacted quickly and regained the lead by a significant margin, demonstrating Player 1's dominance and resilience in the game. In addition, the relatively small range of fluctuations in Player 1's curve suggests greater stability and consistency

in the race. In contrast, Player 2, while performing well at certain moments, demonstrated greater variability and unpredictability over the course of the race.

2. Related Work

In studies related to the quantification of tennis players' performance, a variety of methods and models have been proposed to assess and compare athletes' performance. Traditionally, the assessment of athletes' performance often relies on the final score or ranking of a match; however, these single metrics cannot fully reflect the details and dynamics of an athlete's performance during a match. As a result, researchers have begun to explore more detailed and comprehensive assessment methods [4].

Some studies have used statistical methods to assess athletes' performance by analysing various statistics (e.g., serve points, serve-receive points, match-winning points, and unforced errors, etc.) during a match. These methods are able to provide more specific performance metrics, but still lack dynamic and time-series considerations [5].

In recent years, with the development of data science and machine learning techniques, more and more studies have begun to utilise these techniques to quantify athlete performance. For example, by constructing regression models or classification models to analyse the influencing factors of athlete performance and predict the outcome of a competition or athlete performance rating [6]. These methods are able to integrate the effects of multiple factors on athlete performance and provide a more accurate and comprehensive assessment.

The quantitative model based on linear regression proposed in this study combines Random Forest Feature Analysis and Least Squares, which not only takes into account the dynamics of athletes' performances in matches, but also introduces factors such as serve advantage, which is a useful supplement and extension of the existing research [7]. Through careful analysis of match data and model construction, our method is able to assess the performance of tennis players in a more comprehensive and accurate way.

3. Model: Linear Regression of Player Performance

This section quantifies player performance with existing data, enabling a visual comparison of their relative performance levels in graphs. Recognizing the limitations of using single scores to assess performance, we quantified short-term and long-term indicators based on individual points and entire games.

The paper initially conducts an observational analysis of the C-Wimbledon_featured_matches data file for the 2023 Wimbledon Men's Singles based on an understanding of the tournament format. It verifies that all data conforms to match standards and regulations [8]. However, for subsequent data processing and modeling convenience, the following adjustments were made to irregular values:

Standardized related data forms. Due to the existence of the "tie-break" rule, the "p_score" column has two scoring methods: "0 15 30 40 AD" and the decisive set "1 2 3…". Therefore, in accordance with match regulations, these data are unified.

Filled missing values. Missing attributes in serve_width, serve_depth, and return_depth stem from either unforced errors or the opponent’s unreturnable shots; gaps in speed_mph are mostly observed during net touches and double faults [9]. For ease of further data handling, all missing values are uniformly filled with 'U'.

Corrected outliers. In instances where "rally_count" is not zero, there are 21 entries (0.3% of the data) where "speed_mph" is NA. These 21 instances of "rally_count" are corrected by replacement with zero.

Data categorization. Due to the diverse nature of the columns in the attachment, the data from column A to AT were sequentially divided into four categories in Table 1.

Table 1. Categories of data

Type	Data
ID	match_id, player1, …, point_no
Short-term advantages	p1_set, p2_set, …, set_victor
Long-term advantages	p1_ace, p2_ace, …, p2_break_pt_missed
Surface condition	p1_distance, p2_distance, …,return_depth

Subsequently, embracing the notion that points are linked to games, which in turn relate to sets and ultimately to winning the match, we utilized Excel to extract data from the 'C-Wimbledon_featured_matches' file, summarizing it at the granularity of a game. Based on the extracted data, we further calculated several indicators crucial to "performance" for each player in each match, as shown in Table 2.

Moreover, in consideration of the notes provided in the problem statement, we specifically introduced the Serve Indicator to incorporate the advantage of serving first into the factors affecting player performance.

Table 2. New variables created from existing data

	Term	Definition
CW	Consecutive Wins	Number of sequential points won in a game
LS	Lead Score	Point lead in a game
SFR	Single Fault Rate	Frequency of first serve faults
SI	Serve Indicator	Indicates the player with the serving advantage

Among them, the formulas of LS, SFR are as follows:

$$LS = point_{j1} - point_{j2} \tag{1}$$

$$SFR = {}_{\text{ser}}2_{i/\ \text{ser}i} \tag{2}$$

$point_{j1}$ and $point_{j2}$ are points gained by $player_1$ and $player_2$ when game $n_3 = j$, respectively. ${}_{\text{ser}}2_i$ is the number of times $player_i$ serve twice, while ser is the total number of servings made by $_{\text{player}_i}$.

The Consecutive Wins indicate the player's sustained advantage in the match, distance run reflects their physical exertion, and Single Fault Rate provides information on the player's average error level [10]. Overall, Player 1 performed better than Player 2, consistent with actual match outcomes, thus validating the relevance and value of the chosen indicators.

Figure 1 indicates that, on average, Player 1's successive wins and a lower rate of faults significantly surpass Player 2, with relatively stable energy expenditure, demonstrating adaptability and consistency throughout the match. The data distribution reveals a convergence of Player 1's median and third quartile, suggesting a maintained advantage in roughly half of the games played; in contrast, Player 2 exhibits a higher median error rate [11]. This suggests that streaks of wins and losses may influence player performance, a hypothesis that will be further analyzed subsequently. Additionally, the variability aspect shows that Player 2 experiences greater performance fluctuations, indicating less stability in match play. Anomalies in energy consumption for Player 1, notably higher than usual, could suggest proactive countermeasures in response to the opponent's offensive plays.

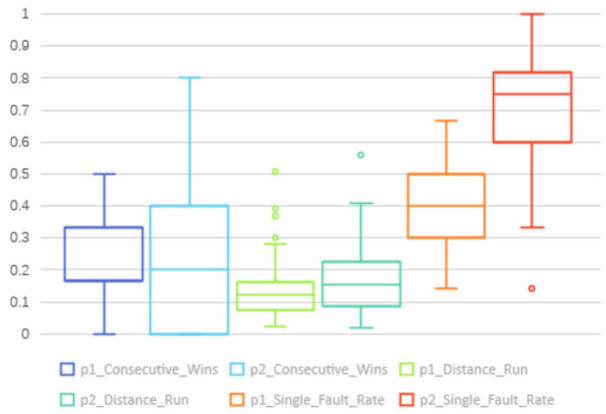

Figure 1. Boxplot of the min-max normalized data for Team 1301

4. Model Development

In this section, we construct the performance indicator using linear fitting, first determining the most significant independent variables through random forest feature analysis, and then calculating parameters using the least squares method. The derived formula is:

$$P_i(n_1, n_2, n_3, n_4) = f\left(x_1, x_2, \cdots, x_j, \theta_1, \theta_2, \cdots, \theta_p\right) \tag{3}$$

Where $P_i(n_1, n_2, n_3, n_4)$ is performance of $player_i$ in the n_1 match, n_2 set, n_3 game, n_4 point ($i = 1\text{or}2; n_1 = 1301,1302 \ldots 1701; n_2, n_3, n_4 = 1,2,3, \ldots$); $x_1, x_2, \cdots, x_j$ is independent variable and $\theta_1, \theta_2, \cdots, \theta_p$ is unknown model parameters of multivariate linear functions;

Random Forests are effective for both regression and classification tasks, and importantly, they can calculate the importance of individual variables (Variable Importance Measure), indicating the significance of each feature in the model. Hence, this study utilizes this Random Forest characteristic for feature extraction from the dataset. After assessing feature importance, eight variables were selected for testing. The six most significant variables were then chosen to construct the "performance" index for athletes [12]. In order to eliminate the influence of inter-feature dimensionality during the test, the data are standardized and forwarded. The results of the feature selection are presented in Table 3.

Table 3. Variable importance measure through Random Forest

Variable	Explanation or Source	Importance
Consecutive Wins 1	Number of sequential points won in a game for player 1	29.00%
Consecutive Wins 2	Number of sequential points won in a game for player 2	27.80%
Serve Indicator	Indicates the player with the serving advantage	16.90%
Single Fault Rate	Frequency of first serve faults	8.40%
Lead Score 1	Point lead in a game for player 1	5.50%
p1_distance_run	'Wimbledon_featured_matches' file	4.90%
p2_distance_run	'Wimbledon_featured_matches' file	4.50%
rally_count_z-score	'Wimbledon_featured_matches' file	3.10%

The Gini Index is employed to assess the importance of each feature in a Random Forest, but it does not directly represent the scoring or prediction outcomes of the samples. Therefore, the subsequent discussion will focus on models related to performance scoring.

The least squares method is a linear fitting model that finds the best function match for the data by minimizing the sum of the squares of the errors, aiming to make the sum of the squares of the errors between the obtained data and the actual data as small as possible. To enhance the analysis and representation of the data, this paper processes two sets of data based on the six performance indices

of athletes previously selected, treating every point as a short particle and every game as a long particle, thereby obtaining expressions for athlete performance at specific times.

$$P_i(n_1, n_2, n_3, n_4) = 0.346 + 0.02X_5 - 0{,}005X_7 + 0.335X_3 \quad (4)$$

$$P_i(n_1, n_2, n_3) = -2.695 + 3.139X_1 + 0.081X_2 + 0.208X_3 - 0.01X_4 + 0.199X_5 + 0.468X_6 \quad (5)$$

Among them, $X_1, X_2, \cdots, X_7$ is Consecutive Wins 1, Consecutive Wins 2, Serve Indicator, Single Fault Rate, Lead Score 1, Distance Run 1 and Current Point respectively.

5. Result Analysis

Taking the data of 2023-wimbledon-1301 as an example, the performance of the athletes is visualized by point and game units, resulting in Figure 2.

In Figure 2, player1's performance is superior to player2's at both the beginning and end. But there are several games where player2 overtakes player1 during the process, which corroborates the description of the problem, indicating that our model is appropriate.

Player1 generally outperforms Player2, showing better control at critical moments. Although Player2 surpasses Player1 in 5 stages, such overtaking is not enduring. Whenever the two performances are close, Player1 often quickly responds and regains the lead with a significant advantage, indicating Player1's dominance and resilience in the match. Additionally, the fluctuation range of Player1's curve is relatively small, implying greater stability and coherence during the match. In contrast, Player2, despite moments of outstanding performance, exhibits greater variability and unpredictability in the match.

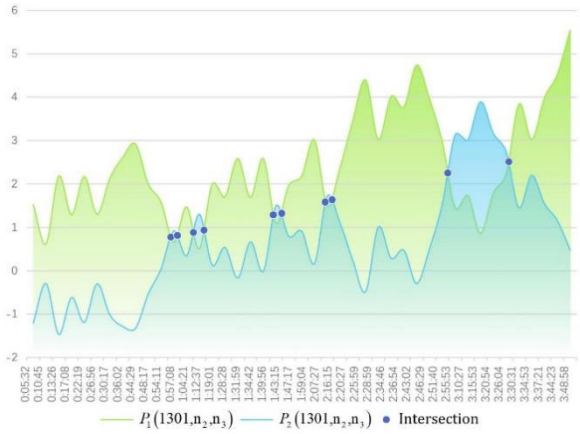

Figure 2. Player Performance Comparison per Game

In Figure 3, P1 (orange bars) usually surpasses P2 (blue bars) in both quantity and height, indicating Player1's overall superior performance in scoring points during the match, not only gaining the advantage in most points but also showing higher consistency and stability. While Player2 does surpass Player1 on certain points, these instances are infrequent and typically not lasting.

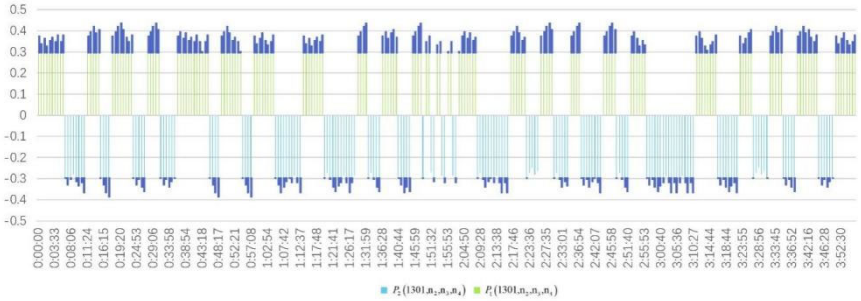

Figure 3. Performance Differential at Each Scoring Point

6. Conclusion

Combining the above discussions, this paper uses linear regression to quantitatively analyse the performance of tennis players. Through observational analyses and data processing of the 2023 Wimbledon men's singles tournament data, we draw some important conclusions.

First, we ensured that the data complied with the tournament standards and regulations by adjusting the form of the data, filling in missing values, correcting outliers and classifying the data, and facilitated the subsequent data processing and modelling.

Second, we extracted key performance indicators based on the tournament data and introduced Serve Indicator to consider the impact of first-mover advantage on players' performance. By comparing the performance indicators, we found that player 1 usually outperforms player 2, which is consistent with the results of the actual tournament and verifies the relevance and value of the indicators.

In terms of model development, we identified the most important independent variables using Random Forest feature analysis and calculated the parameters by least squares to derive performance metrics for use in linear regression.

Finally, by analysing the data from match 1301 at Wimbledon 2023, we found that player 1 usually outperformed player 2 throughout the match, especially at key moments. Although player 2 outperformed player 1 in certain moments, this outperformance was not sustained. Overall, Player 1 shows better control and greater stability, while Player 2 shows greater variability and unpredictability.

In summary, the linear regression method in this paper is valuable and practical for quantifying the performance of tennis players, and can provide a reference for players and coaches to help them better analyse and improve their performance.

7. Discussion

In this study, we quantitatively analysed the performance of tennis players through linear regression methods. Firstly, we carried out an observational analysis of the data from the featured matches in the men's singles at Wimbledon 2023 and performed a series of pre-processing on the data, including standardising the form of the data of interest, filling in missing values, correcting for outliers, and classifying the data. We then extracted and summarised the data for each player in each match based on short and long term metrics using Excel, and then calculated several key indicators of 'performance' such as consecutive wins, points lead, error rate and serve indicator.

In the model construction section, we firstly identified the most important independent variables through the Random Forest analysis, and then used the least squares method to calculate the parameters to obtain a linear regression model of the athletes' performance. Through the Random Forest, we selected eight variables to test, and finally chose the six most important variables to construct the athlete's "performance" index. The data were standardised to eliminate the effect of dimensionality between characteristics.

The analysis of the results is based on the data from the Wimbledon 1301 tournament in 2023, where the performance of the athletes is visualised in terms of points and board units. The results show that player 1 outperforms player 2 at the beginning and end of the match, but there are a few situations during the course where player 2 outperforms player 1. This is consistent with the problem description and suggests that our model is appropriate. Player 1 usually performs better at key moments, and although player 2 outperforms player 1 in five phases, this outperformance is not sustained. Whenever the two performances were close, Player 1 usually reacted quickly and regained the lead by a significant margin, demonstrating Player 1's dominance and resilience in the game. In addition, the relatively small range of fluctuations in Player 1's curve suggests greater stability and consistency in

the race. In contrast, Player 2, while performing well at certain moments, demonstrated greater variability and unpredictability over the course of the race.

References

[1] Kramer, T., Huijgen, B. C., Elferink-Gemser, M. T., & Visscher, C. (2017). Prediction of tennis performance in junior elite tennis players. Journal of sports science & medicine, 16(1), 14.

[2] Kovalchik, S. A. (2016). Searching for the GOAT of tennis win prediction. Journal of Quantitative Analysis in Sports, 12(3), 127-138.

[3] Whiteside, D., Cant, O., Connolly, M., & Reid, M. (2017). Monitoring hitting load in tennis using inertial sensors and machine learning. International journal of sports physiology and performance, 12(9), 1212-1217.

[4] Cui, Y., Gómez, M. Á., Gonçalves, B., & Sampaio, J. (2018). Performance profiles of professional female tennis players in grand slams. PloS one, 13(7), e0200591.

[5] Signorile, J. F., Sandler, D. J., Smith, W. N., Stoutenberg, M., & Perry, A. C. (2005). Correlation analyses and regression modeling between isokinetic testing and on-court performance in competitive adolescent tennis players. The Journal of Strength & Conditioning Research, 19(3), 519-526.

[6] Buszard, T., Reid, M., Krause, L., Kovalchik, S., & Farrow, D. (2017). Quantifying contextual interference and its effect on skill transfer in skilled youth tennis players. Frontiers in psychology, 8, 1931.

[7] Triolet, C., Benguigui, N., Le Runigo, C., & Williams, A. M. (2013). Quantifying the nature of anticipation in professional tennis. Journal of Sports Sciences, 31(8), 820-830.

[8] Cui, Y., Gómez, M. Á., Gonçalves, B., Liu, H., & Sampaio, J. (2017). Effects of experience and relative quality in tennis match performance during four Grand Slams. International Journal of Performance Analysis in Sport, 17(5), 783-801.

[9] Hayes, M. J., Spits, D. R., Watts, D. G., & Kelly, V. G. (2021). Relationship between tennis serve velocity and select performance measures. The Journal of Strength & Conditioning Research, 35(1), 190-197.

[10] Keaney, E. M., & Reid, M. (2018). Quantifying hitting activity in tennis with racket sensors: new dawn or false dawn?. Sports Biomechanics.

[11] Cui, Y., Gómez, M. Á., Gonçalves, B., & Sampaio, J. (2019). Clustering tennis players' anthropometric and individual features helps to reveal performance fingerprints. European journal of sport science, 19(8), 1032-1044.

[12] Fett, J., Ulbricht, A., & Ferrauti, A. (2020). Impact of physical performance and anthropometric characteristics on serve velocity in elite junior tennis players. The Journal of Strength & Conditioning Research, 34(1), 192-202.