Momentum analysis of tennis players based on entropy weight method and personal's correlation coefficient

Cong Wang[1,*], Kairan Wang[2], Yumo Hu[3]

[1]School of Transportation and Logistics Engineering, Wuhan University of Technology, Wuhan, China,430070

[2]School of Economics, Wuhan University of Technology, Wuhan, China, 430070

[3]School of Mechanical and Electronic Engineering, Wuhan University of Technology, Wuhan, China, 430070

*Corresponding author: wcong135@gmail.com

Abstract. In sports, a player or team may feel "powerful" at a certain stage of the game, and the momen-tum of the game may be shifted to a certain side, and the analysis of game momentum and player momentum has been an important topic in the field of sports research. In this paper, we use the data from the final of the 2023 Wimbledon Tennis Championships to first construct a comprehensive model that includes the analysis of players' performance during the match, in order to understand the reasons for the shift in the momentum of the match and its impact on the final outcome of the match. The model utilizes a modified entropy weighting method to quantify the performance of tennis players, and constructs "momentum" characteristics to quantify the cumulative advantage among tennis players and its effect on the momentum of the match. The relationship between the performance of the players and the match result is then analyzed by using the personal correlation coefficient, and a strong positive correlation is found between the two, which indicates that momentum and momentum play a key role in de-termining the winner of the match.

Keywords: Tennis Sport, Momentum Analysis, Entropy Weight Method, Correlation Analysis.

1. Introduction

The men's singles final of the 2023 Wimbledon Open between 20-year-old Spanish star Carlos Alcaraz and 36-year-old Novak Djokovic not only marked the rise of a new star in tennis, but also ended Djokovic's long reign at Wimbledon. The outcome of this match sparked an in-depth discussion on a number of key factors, player performances, match dynamics, and momentum shifts that contributed to the winner's victory. Alcaraz dominated at key moments with outstanding performances and calm coping mechanisms, while Djokovic failed to win despite displaying indomitable spirit and high levels of athleticism. The fluidity of the scoring and the complexity of the momentum transformations between the two sides in the match were the most exciting aspects of the match. The order in which the two players scored points was elusive, and it was clear that the dynamics and uncertainty of the match were more captivating to the audience than the skill and tactics of the players[1]. Another major focus of the match was the frequent changes in momentum, with the experienced Djokovic gaining the upper hand, and the dynamic Alcaraz pulling one back. The constant shifts in momentum made the match all the more thrilling. The dynamic nature of this match and the frequent changes in momentum added to the uncertainty and spectacle of the match, while also highlighting the significant impact of momentum on the outcome. How should one measure a player's performance[2], does momentum exist to allow an already dominant player to further gain an even greater advantage[3], how does the momentum of a match change, and are the turns of the match influenced by the momentum of the player.

Citing previous research, players with excellent serving skills can not only create conditions for the use of various tactics after the serve, but also psychologically suppress their opponents and limit their tactics, so players with good serving skills tend to score more efficiently[4]. Professional tennis matches tend to be mixed rather than pure strategy Nash equilibrium due to the mutual constraints of technical and tactical strategies. In order to break the equilibrium, both sides need to optimize their

strategies and incorporate new combinations of tactical strategies in order to increase the weight of winning, so it is especially important to adjust the tactics according to the situation of the match[5].

In view of the above research, this paper quantitatively evaluates the performance of tennis players by the improved entropy weighting method, and quantifies the cumulative advantage among tennis players and its influence on the momentum of the match through the construction of "momentum" characteristics. Then the relationship between the performance of players and the result of the match was analyzed by using the personal correlation coefficient, and it was found that there was a strong positive correlation between the two.

2. The basic funamental of Entropy weight method Principle and Pearson's correlation coefficient

2.1. Entropy weight method Principle

Entropy Weighting Method (Entropy Weighting Method) is a weight allocation method for multi-feature decision-making problems, and its basic idea is to determine the objective weights of the features based on the magnitude of the variability of the indicators. Information is a measure of the degree of orderliness of a system, and entropy is a measure of the degree of disorder of a system; according to the definition of information entropy, for a certain indicator, the entropy value can be used to determine the degree of dispersion of a certain indicator. If the information entropy of an indicator is smaller, it indicates that the indicator is worth the greater degree of variability, the more information it provides, the greater role it can play in the comprehensive evaluation, and the greater its weight. The smaller the information entropy, the opposite[6].

The steps of the entropy weight method are as follows:

1. Standardized normalization of data:

Need to ensure that the data used have a certain measurement or evaluation standards, and because the unit of measurement of the indicators is not consistent, so in order to ensure that the data of different attributes are comparable, need to calculate the integrated weights before the normalization process, converted to the value of the same size range, that is, the absolute value of the indicators into a relative value, and make:

$$X_{\mathrm{ij}} = \left|X_{\mathrm{ij}}\right| \tag{1}$$

For positive indicators:

$$X_{ij} = 0.998\frac{X_{ij} - \min\left\{X_{1j},\ldots,X_{nj}\right\}}{\max\left\{X_{1j},\ldots,X_{nj}\right\} - \min\left\{X_{1j},\ldots,X_{nj}\right\}} + 0.002 \tag{2}$$

For negative indicators:

$$X_{ij} = 0.998\frac{\max\left\{X_{1j},\ldots,X_{nj}\right\} - X_{ij}}{\max\left\{X_{1j},\ldots,X_{nj}\right\} - \min\left\{X_{1j},\ldots,X_{nj}\right\}} + 0.002 \tag{3}$$

2. Calculate the ratio of each indicator under each program:

Let there be a primary indicator with m secondary indicators and n data have been obtained, denoted as matrixY_{ij} . Under the same indicator, the proportion of values taken in each year to the total value is calculated using the following formula:

$$P_{ij} = \frac{Y_{ij}}{\sum_{i=1}^{n} Y_{ij}}\left(i = 1,2,\ldots,n;\ j = 1,2,...,m\right) \tag{4}$$

3. Calculate the information entropy: For each indicator attribute, its information entropy needs to be calculated. In the entropy weight method, the information entropy indicates the diversity of the attributes as well as the degree of discretization, and the following formula is usually used:

$$E_j = -ln(n)^{-1} \sum_{i=1}^{n} P_{ij} \, ln \, P_{ij} \; (if \; P_{ij} = 0, define \; E_j = 0) \tag{5}$$

4. Calculate the weight of each indicator: first calculate the weight by information entropy:

$$w_j = \frac{1 - E_j}{k - \sum E_j} (k = m) \tag{6}$$

then calculate the weight by calculating information redundancy:

$$D_j = 1 - E_j \tag{7}$$

and then calculate the weight of the indicator:

$$w_j = \frac{D_j}{\sum_{j=1}^{m} D_j} \tag{8}$$

5. Calculate the composite score for each program: Use the weights obtained for each indicator to score, usually higher weighted indicators have a greater impact, using the formula below:

$$H_i = \sum_{j=1}^{m} w_j * X_{ij} \tag{9}$$

The advantage of entropy weighting method is to determine the weights of indicators according to the degree of variability of each indicator value, as an objective assignment method, avoiding the bias brought by human factors. Compared with other subjective assignment methods, the accuracy is higher and the objectivity is stronger, which can better explain the results obtained. And the calculation steps of entropy weight method are simple and easy to operate. However, entropy weighting method also has some shortcomings, such as ignoring the importance of the indicators themselves, sometimes the weights of the determined indicators will be far from the expected results, the data quality requirements are high, while entropy value method can not reduce the number of dimensions of the evaluation indicators, that is, entropy weighting method conforms to the laws of mathematics with a strict mathematical significance, but often ignores the decision maker's subjective intentions.

2.2. Pearson's correlation coefficient

Correlation is a number that describes the degree of correlation between two indicators. The correlation coefficient takes values in the range [-1,+1]. Positive values indicate that as the value of one variable increases so does the other, negative values indicate the opposite, and zero indicates that there is no correlation between the two indicators[7].

Pearson's correlation is the most commonly used correlation calculation, and the Pearson's correlation coefficient, usually denoted by the letter r, is used to measure the linear relationship between two random variables

The Pearson's correlation coefficient for the aggregate (population) between two variables is defined as the normalized covariance between the two variables. The definition is shown in the following equation:

$$\rho_{ab} = \frac{\mathrm{cov}(a,b)}{\sigma_a^{\,2} \sigma_a^{\,2}} = \frac{W\left[(a - \mu_a)(b - \mu_b)\right]}{\sqrt{\sigma_a^{\,2} \sigma_b^{\,2}}} \tag{10}$$

The Pearson's correlation coefficient of the sample can be obtained by estimating the covariance and standard deviation of the sample indicators with the following expression:

$$r = \frac{\sum_{i=1}^{n}(a_i - \overline{A})(b_i - \overline{B})}{\sqrt{\sum_{i=1}^{n}(a_i - \overline{A})^2}\sqrt{\sum_{i=1}^{n}(a_i - \overline{B})^2}}) \tag{11}$$

3. Results

3.1. Data standardization

Data Source: https://www.comap.com

First of all, we need to ensure the completeness of the data of each game, so this paper preprocesses the obtained data. The BP neural network algorithm is used to fill in the gaps, including the time of the game, the scores of the two sides, the serving side, the batting situation, etc., which will not be described in the following. Unify the time format and the representation of the score by the time of the start of the serve in each round. Convert data into processable numerical data. For example, the resulting data in the player score, due to the rules of the game, the score contains AD, which is not conducive to the subsequent data processing, so this paper will be converted to 50.

3.2. Feature extraction

The following extracts each of the performance characteristics of the players during the game and is used later to evaluate the performance of the players.

Since a win by either of the two sides results in a difference in scoring, this paper compares the difference between the wins of the two sides in a set of matches, with a few of the games taken at random to show the wins in the match, as shown in Figure 1. This paper shows visualization plots of the difference in scoring (Player 1 score - Player 2 score) for each scoring point in a few randomly selected matches. Each graph represents one game. The horizontal coordinate is the sequence of scoring points in the match, and the vertical coordinate is the difference in scoring between the two players, where a positive difference means that Player 1 is leading the scoring at this scoring point, and a negative difference means that Player 2 is leading the scoring at this scoring point. A random seed of 42 is fixed to ensure that the subsequent examples use data from the same game.

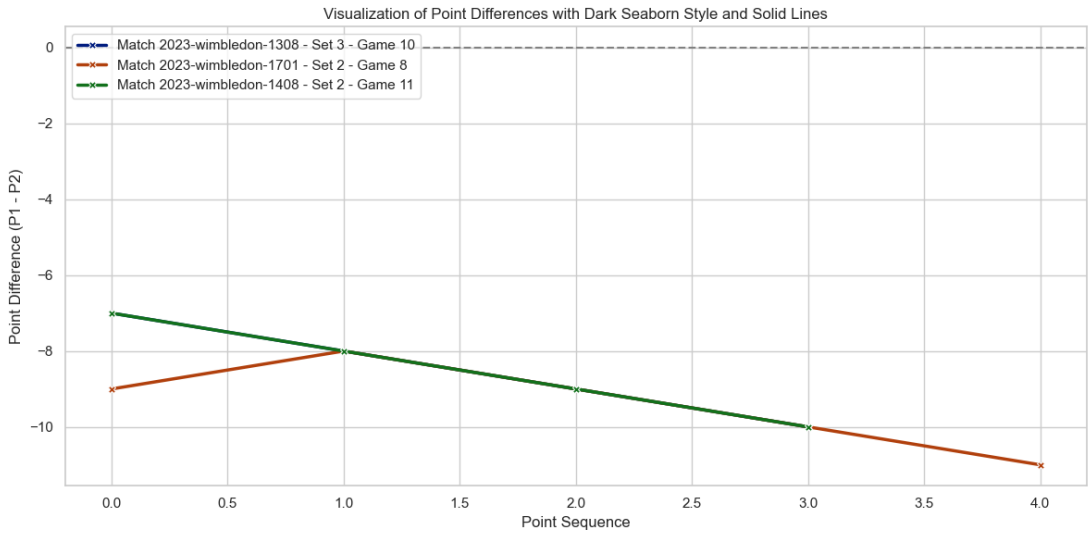

Figure 1 Table analyzing the difference in scoring between the two players

As well, this paper counts the number of ACEs on both sides of the serve. an ACE indicates that the player hit a winning serve that was out of reach of the opponent. If a player has completed an ace at that time, the player's current state is updated to a relatively better state in terms of ball feel and physical condition, which is a characteristic indicator that needs to be counted in the player's performance[8]. And serve ACE is relatively easier to hit, so the weight of serve ACE needs to be considered comprehensively. In this paper, the data of three matches are randomly selected, and the corresponding ACE serve situations in the matches are visualized in Figure 2, in which the folded line represents the difference between the serve ACEs of both players. This paper can visualize the difference between the two sides' serve ACE balls from the figure.

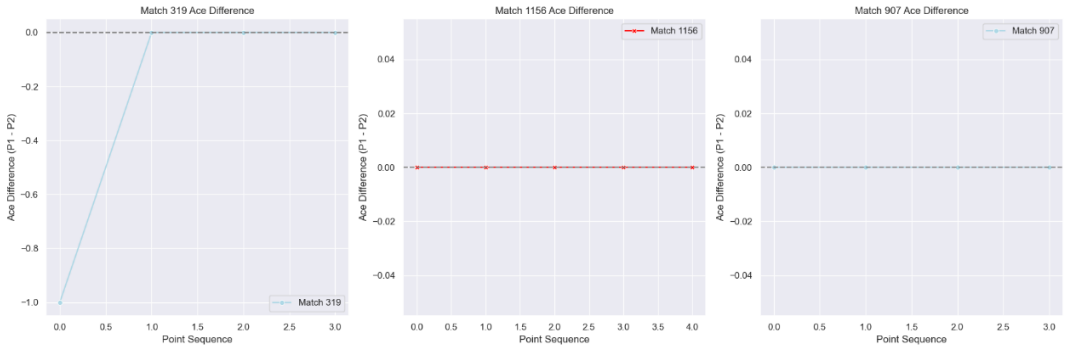

Figure 2 Scoring of Ace Served by Both Players

Below are the statistics of the difference between the two sides in terms of winner point, which means that the player has hit a winning ball that is out of reach of the opponent, similar to an ACE ball, which indicates that the player is in a good condition or is about to turn the situation around. However, compared with the ACE ball, the winner point is easier to hit, so in the subsequent statistics, compared with the ACE ball, this paper will reduce the weight assigned to the winner point. The difference between the two players in the winner points is shown in Figure 3.

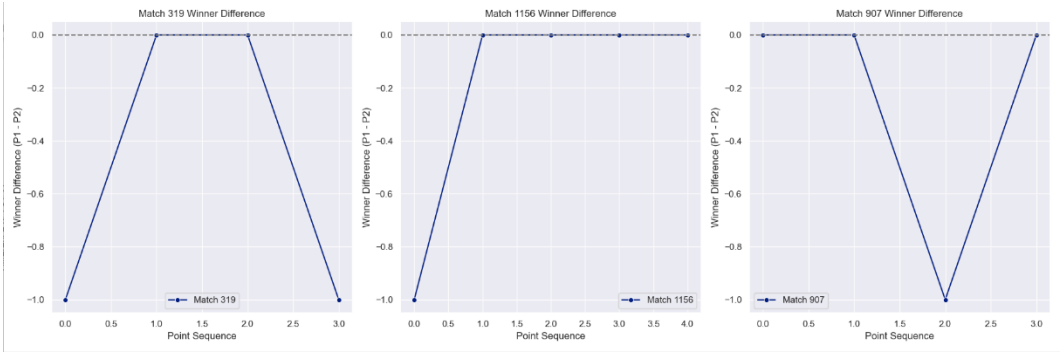

Figure 3 Winner ball difference of both players

This paper also counts the Serve Holds and Break Points of both players in the match, as shown in Figure 4. Break Points means that the player hits a break point in the opponent's serve, which can indicate that the player's hand is hotter and in better condition[9]. Serve Holds means that the player on the serve side wins the current set, and the player's mind will be more stable, which is also beneficial to the player's play. Serve Holds says that if the player on the serve side wins the current set, the player's mindset will be more stable, which is also favorable to the player's performance. Of course, it is certain that Break Points will have a more significant impact on a player's current momentum.

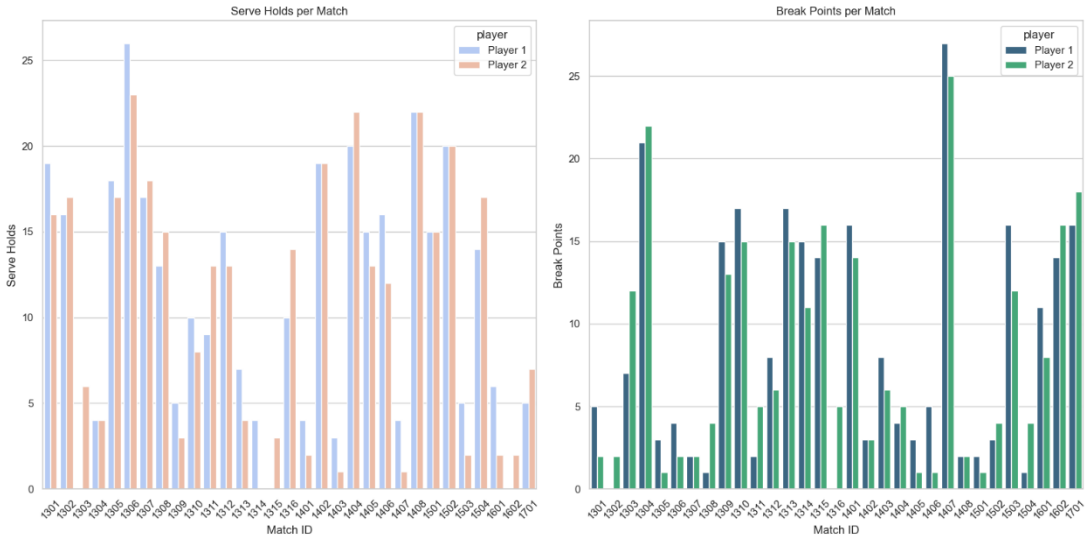

Figure 4 Serve Holds and Break Points for both players

In the process of data processing, this paper analyzes the length of the game experienced by each score, the longer the time experienced by a score represents a more indistinguishable duel between the players, which can be regarded as the performance of both players is good. Therefore, this paper counts the length of the game experienced by each score, and calculates certain weights in the subsequent analysis. The length of the match experienced by each score is visualized as shown in Figure 5, with the vertical axis indicating the time the match was played.

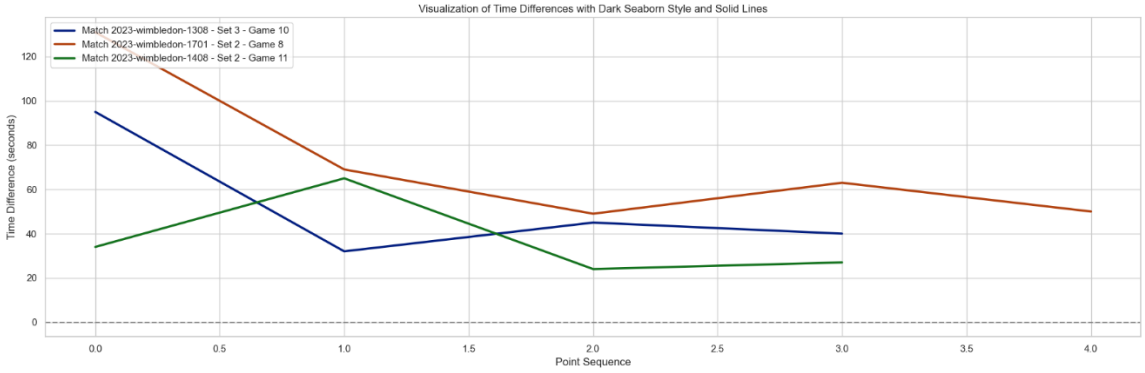

Figure 5 Length of game per scoring experience

Similarly, this paper analyzes the errors made by the players during the course of the match. Errors leading to loss of points represent the loss of momentum of the players' form that occurred during the course of the game, which will be given some weight in the analysis. The statistical results of this paper are shown in Figure 6, where the difference between the players' errors in the three matches is randomly selected, with positive values of the vertical coordinate representing that player 1 had fewer errors and negative values representing that player 2 had fewer errors.

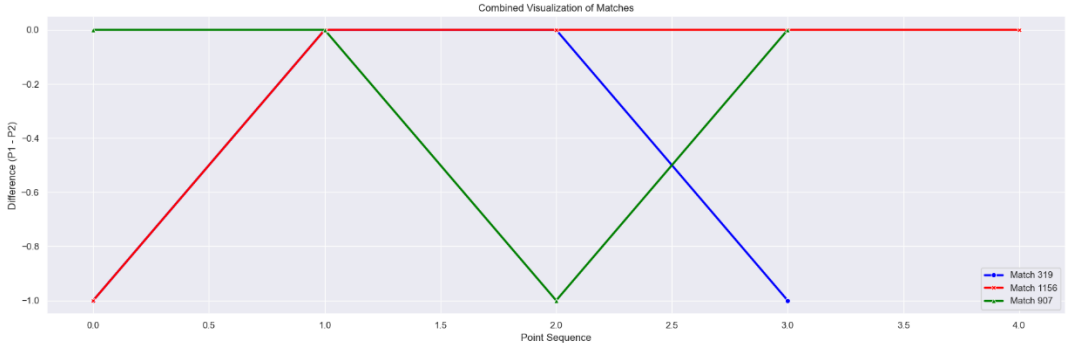

Figure 6 Difference in errors by both players

The paper concludes with statistics on the wins of both players in a set to set up a control group that is used to control the accuracy of the features collected above, the data visualization is shown in Figure 7. Looking at the graph it can be seen that player 2 performed quite well in all three matches and was able to score quick winners to take the match. This is mirrored by the ACEs and winners as well as the player's errors as described above.

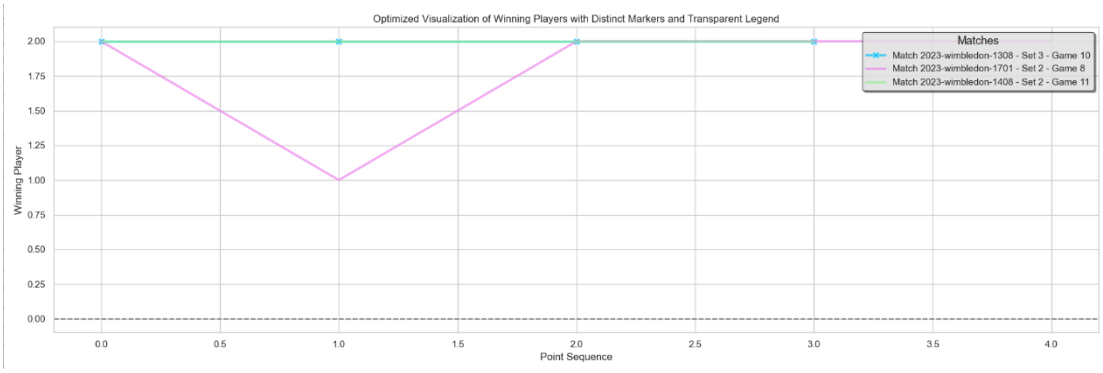

Figure 7 Players' wins

In summary, this paper has completed the preliminary processing of data and the extraction of indicator features, the following paper will be based on the extracted features for the evaluation of player performance.

3.3. Analysis of race-time performance based on improved entropy weighting method

Next, the performance of both players was evaluated using the improved entropy weight method. Combined with the extracted features, the evaluation model of the players' performance was established, and the modeling flow is shown in Figure 8.

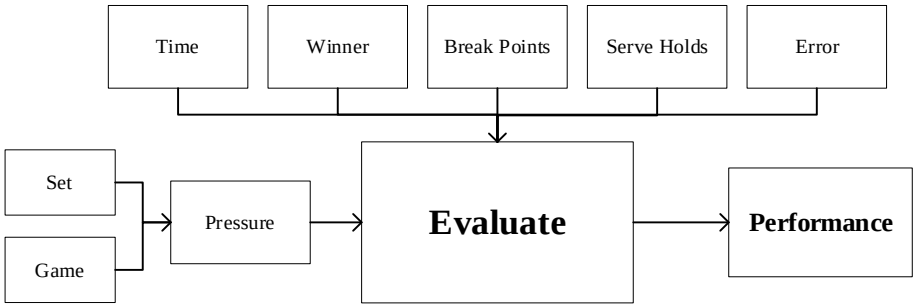

Figure 8 Modeling flow

When a player wins a game or a set, he or she gets closer to winning the game, which increases the player's confidence and the psychological pressure on the opponent, thus making the player perform better. This phenomenon is called "cumulative advantage" in this paper, and cumulative advantage can lead to optimization of the player's performance, so this paper incorporates it into the model of the player's performance. Therefore, this paper includes this phenomenon in the player performance model, and the player's performance score will be improved by winning the set and game.

The following model is constructed to calculate the player's performance, assigning corresponding weights to each feature with the following formula.

$$\begin{aligned} Performance = \frac{Time}{100} + 1.5*\text{Break Points won} + 0.5*\text{Serve Holds} + \\ 0.5*\text{winner} - 0.5*\text{error} + \text{pressure} + \text{win}*\alpha \end{aligned} \tag{12}$$

Where break point won indicates that the player breaks serve and wins, serve holds indicates that the player holds serve and wins, winner indicates whether the player wins or not, 1 if he wins, 0

otherwise; α represents the winning streak coefficient, α=1.2 when the player's last ball is also a score, otherwise α+1.

By this point, this paper completes the modeling of the player performance model, and this paper evaluates the player performance for one game, and the results are shown in Figure 9. The vertical axis represents the player's performance score in the corresponding ball, and the horizontal axis represents the number of each ball. It can be seen that the performance score of player 1 is slightly higher than that of player 2 as a whole, so player 1 performs better than player 2 in the match.

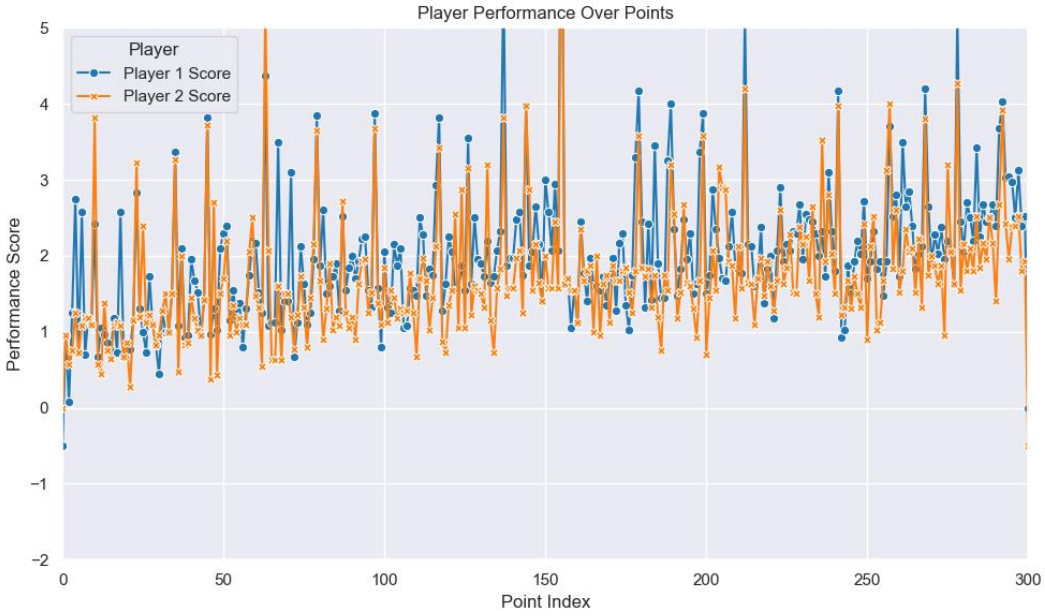

Figure 9 Evaluation of the performance of the parties

In order to further see the difference in player performance, this paper randomly intercepted a section from the middle of the game to clearly show the difference in performance between the two, and the results are shown in Figure 10 below. The results are shown in Figure 10 below. It can be seen that in this section of the game, the performance of player 1 is better than that of player 2.

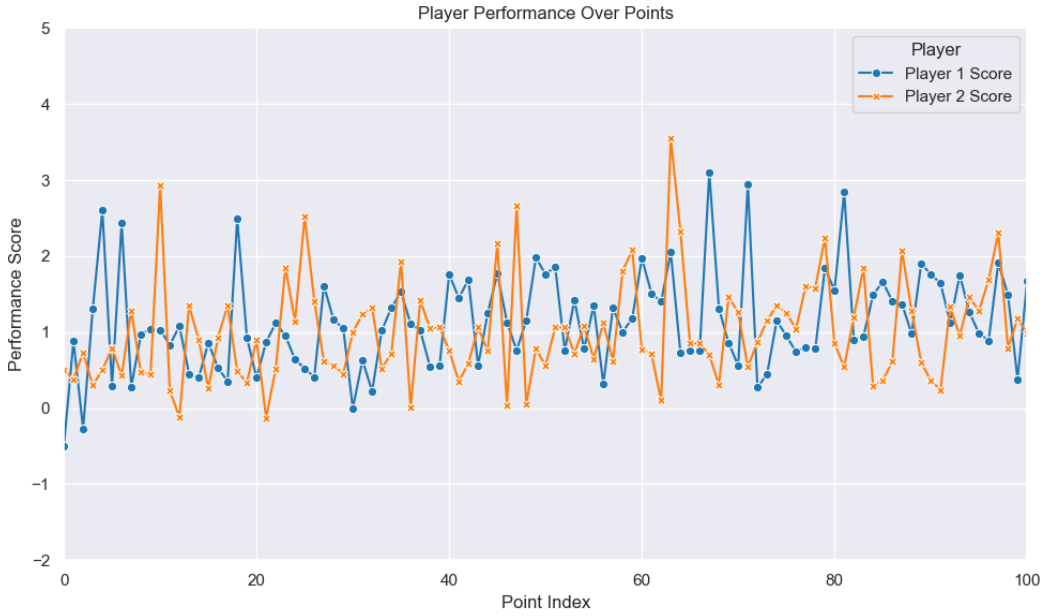

Figure 10 Snippets of data from the evaluation of the performance of the parties

3.4. Personal correlation analysis of player performance and game fluctuations

In order to validate the questioning of the correlation between match fluctuations and match results, in particular, the relevance of quantifying the performance profile of the players through the use of statistical analysis. In doing so, this paper employed the Pearson correlation coefficient as a statistical tool to ensure the comprehensiveness and reliability of the analysis. The results of the analysis reveal a Pearson correlation coefficient of 0.9995, a finding that suggests a high level of correlation between player performances.

The results of the high Pearson correlation coefficients specifically point to the existence of significant linear correlations between the features used to calculate the player performance scores[10] Together, these results reinforce this paper's understanding of the correlates of player performance, showing that the accuracy and explanatory power of predictions of player performance can be effectively enhanced by integrating the correlations between the different features.

Further data analysis revealed a low degree of correlation between most of the data, suggesting that this paper needs to be alert to possible low correlation data during the application of the model. These low correlated data may indicate that the model needs to be adapted or a more detailed analysis of a specific data set in a given situation to ensure the accuracy of the predictions and the generalization ability of the model.

Next, this paper conducts further correlation analysis on the obtained data to find out the possible correlations, as shown in Figure 11, most of the data have a low degree of correlation with each other.

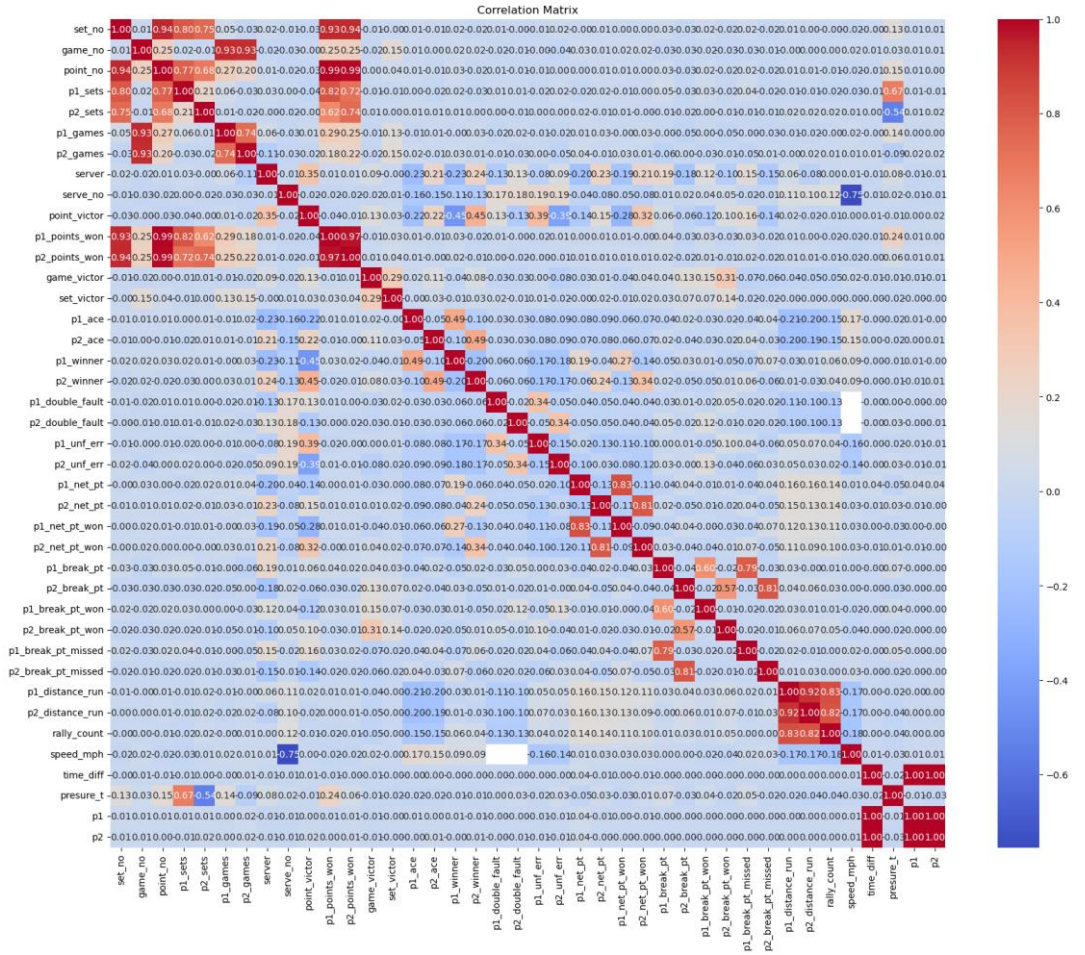

Figure 11 Degree of correlation between data

3.5. Cumulative process analysis of player performance

In the previous section, this paper calculated the performance of both players in the game, and the following further analyzes the cumulative process of the players' performance.

$$\Delta p=p1-p2 \tag{13}$$

where ΔP represents the difference in performance between the two players, with a positive value indicating a better performance by player 1 and a negative value indicating a better performance by player 2.

After constructing the performance differences between the two players, this paper needs to calculate the cumulative performance differences between the two sides during the game. Points are calculated for the performance differences of the players during the whole game to calculate the cumulative performance, and some of the results are displayed in Table 1.

Table.1. Accumulated performance of players throughout the entire game

1411	1412	1413	1414	1415	1416	1417	1418	1419	1420
80.2	-122.6	-95.4	-2.3	27.3	182.2	147.8	-91.2	13.1	76.4

3.6. Correlation between tournament fluctuations and player performance

Next, analyze the correlation between the fluctuation of the game and the performance of the players, this paper locates the data of the last ball of each game, the win or loss of this ball has a direct correlation with the win or loss of the whole game, this paper carries out the Boolean operation on the cumulative performance, through the dichotomous value of whether or not it indicates which side is more advantageous, combined with the win or loss of the game, this paper plots the situation as shown in Figure 12, which is positive for 1 indicates that the player of the game 1 is more advantageous. The positive value of 1 indicates that player 1 is more advantageous in the game. The negative value of -1 indicates that player 2 is more advantageous. From the figure, it can be seen that there is a high degree of consistency between the player's cumulative advantage in the game and the result of the game.

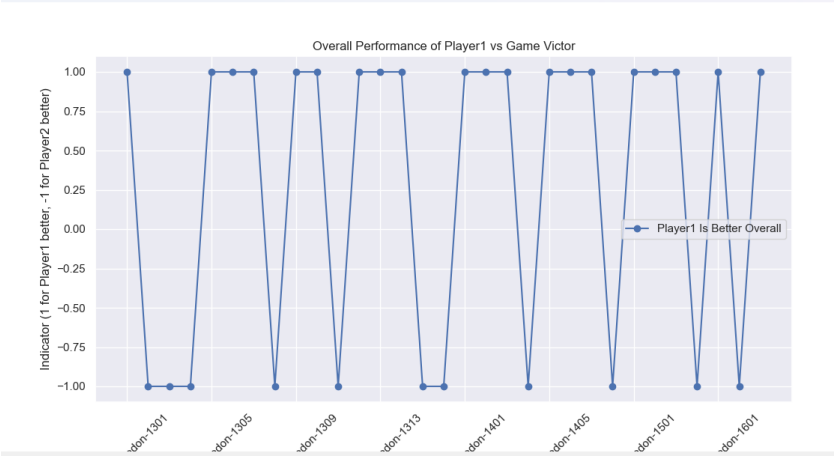

Figure 12 Analysis of the performance advantages of both parties

The following paper then carries out correlation analysis, carries out the calculation of the Pearson correlation coefficient obtained correlation coefficient of -0.7832. It can be seen that there is also a strong correlation between the two, comprehensive above analysis, this paper can dismiss the coach's questioning, the fluctuation of the advantage during the game and the player's success or not is not random, and there is a strong correlation between the two.

4. Conclusions

In this paper, we conducted an in-depth analysis of tennis players' match-time performance, cumulative advantage among players and its effect on match momentum, correlation between players' performance and match results, correlation between players' performance and match fluctuations through the improved entropy weight method and personal correlation coefficient, aiming to explore the issues related to tennis players' momentum and match analysis. After a rigorous research process,

we obtained a series of meaningful findings. First, this study confirms the strong positive correlation between good and bad player performance and match results, a result that has important theoretical implications for understanding the role of player momentum in tennis matches. Second, we found the cumulative advantage among tennis players and its effect on match momentum, a finding that provides practical guidance for the development and modification of tactics in tennis matches. In addition, this paper explores quantitative evaluation methods for tennis players' performance, enriching the research in this area.

Despite the results achieved in this study, there are still some limitations. For example, with the increase of the extracted indicator features, the fitting effect of the player performance model will be better, but the complexity of the model also increases. Therefore, the extraction of indicator features can be optimized according to the specifics of the prediction demand, and how to introduce more features to make the player performance evaluation system more complete needs further research. The correlation analysis draws more credible conclusions when enough match data are selected for correlation analysis. The sample match data used in this study is small, so attention can be paid to the reasonable selection of match samples, such as selecting more match data in different environments or women's match data.

References

[1] Fang Xiangzhong. Uncertainty makes sports competition full of charm[J]. China Statistics,2022(2):21-23.

[2] Du Zirui. Current status of technical and tactical research on competitive tennis in China based on visualization analysis[J]. Sports Goods and Technology,2023,14(14):153-155.

[3] Wang Jiarui,Zou Zhumeiyi.Analysis of the factors of Djokovic's victory in 2021 Australian Open[C]// Hubei Sports Science Society. Proceedings of the First Hubei Sports Science Conference(Volume I). Wuhan Institute of Physical Education;,2023:2.

[4] Rong Doudou,Xu Xiangfeng. Research on the characteristics of competitive ability of amateur male tennis players in China -an analysis based on the men's singles champion of the 14th National Games mass competition tennis final[J]. Sports Excellence,2022,41(6):89-92,95.

[5] WANG Wei,ZHOU Shu. Research and analysis of game characteristics of professional tennis competition[C]// Hubei Sports Science Society. Proceedings of the First Hubei Sports Science Conference (Volume I). Hubei Normal University;,2023:3.

[6] LI Qingwen,PEI Huafu,SONG Huaibo,et al. Prediction of PSO-SVR-NGM slope displacement based on the optimized combination of entropy weight method[J]. Journal of Engineering Geology,2023,31(3):949-958.

[7] WANG Jinwei,SAN Yuhan,SAN Baohai. Analysis of short track speed skating competition data based on Pearson's correlation coefficient[J]. Ice and Snow Sports,2023,45(04):9-12+17.

[8] Chen Chen,Zhao Qianchang.Characterization of technical and tactical analysis of Wu Yibing's serve phase in 2022 US Open[J]. Sports Science and Technology Literature Bulletin,2023,31(07):98-100+124.

[9] MA Jiacai,CHEN Qingguo,LIANG Cairong,et al. Model construction and research analysis of match-winning factors of professional tennis players[J]. Sports Science and Technology Literature Bulletin,2023,31(4):66-68,118.

[10] DONG Lipeng,NIE Qinghao,SUN Xiaokun et al. Analysis of the influence of shield tunneling parameters on surface settlement based on Pearson's correlation coefficient method[J]. Construction Technology,2024,53(01):116-123.