Journal of Electronics and Information Science (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/jeis.2024.090217
ISSN 2371-9524 Vol. 9 Num. 2

A Study of Tennis Match Momentum Based on Random Forest Model and AHP Approach

Renyang Xiong[1,*]

[1]*College of Automation Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, China*
[*]*Corresponding author: renyangxiong@163.com*

Keywords: Random Forest Modelling, AHP, Spearman's Correlation Coefficient

Abstract: This study explores the role of momentum in tennis by developing a mathematical model. Firstly, entropy weighting and hierarchical analysis were used to determine the weights of the athletes' competitive performance index, and the final results were obtained through the game theory combination weighting method. Then, the link between the scores of both sides of the match and momentum was analysed by Spearman's correlation coefficient, and the correlation coefficient between the two was found to be 0.610, which proved the significant influence of momentum on the results of the match. Further, a random forest model was used to predict the turning point of the match, and indicators such as running distance and winning points were found to have a significant effect on the outcome of the match. These findings provide important insights for a deeper understanding of the role of momentum in tennis matches, which can help optimise athletes' competitive performance and tactical strategies.

1. Introduction

Momentum plays an important role in tennis matches, yet its exact impact is difficult to quantify. The aim of this study is to develop a mathematical model to systematically analyse the impact of momentum on match outcomes. By defining a competitive performance index and applying different analytical methods, we selected a match from the 2023 Wimbledon tournament as a case study to validate the importance of momentum in the game. Our findings will provide practical data support for athletes and coaches to optimise match performance and tactical decisions. Through this study, we aim to reveal the key role of momentum in tennis and provide useful insights for the development and practice in the field of sports science [1].

2. Tennis match evaluation model

2.1 Fraction size

The score difference between players can help us better understand the current state of the game and the possible outcomes of the game. At the same time, the size of the score difference tends to affect the players' game strategy and mental state. Therefore, we use the average score difference in a game to measure the readiness of the players.

Table 1: The weights of the API

Object	Indicators	Description
AA	FWP	First serve win percentage
	NWP	Net winning percentage
	WP	Winning point
DA	FRP	First serve return percentage
	UE	Unforced error
	TNSR	The number of strokes in a round
	RD	Running distance
4	SS	Serving speed
	BE	Break efficiency
	DS	Different score
	DF	Double fault
	CG	Critical game

Based on the Table 1 indicators, we further determine the weights of these indicators to achieve a combination of indicators. We need to consider the impact of each indicator on the athlete's performance, so we use the entropy weight method to comprehensively evaluate multiple indicators. The entropy weight method is based on the concept of information entropy to determine the importance of indicators and assign the corresponding weights [2].

Before using the entropy weight method, we need to standardize these indicators so that the best and worst values after each variable change are 1 and 0, respectively. The evaluation variables are $X_1, X_2 \cdots \cdots X_k$, where $X_i = (x_{1i}, x_{2i} \cdots \cdots x_{ni}), k$ and n are the defined evaluation indicators and the number of matches respectively.

We need to standardize data for different indicators so that comparisons can be made on the same scale. For different types of data, we use different normalization methods. For the "cost attribute type", that is, the smaller the data type, the better it is, we use the following methods for standardization:

$$x_{in}^{*} = \frac{\max\{X_i\} - x_{in}}{\max\{X_i\} - \min\{X_i\}} \tag{1}$$

For "benefit attribute types", the larger the data type, the better it is, we normalize using the following formula:

$$x_{in}^{*} = \frac{x_{in} - \min\{X_i\}}{\max\{X_i\} - \min\{X_i\}} \tag{2}$$

Where x_{in}^{*} is the standardized value of each evaluation index of the match, $\max\{X_i\}$ and $\min\{X_i\}$ are the maximum and minimum values of evaluation index X_i respectively. The $\max\{X_i\} = \max\{x_{1i}, x_{2i} \cdots \cdots x_{ni}\}, \min\{X_i\} = \min\{x_{1i}, x_{2i} \cdots \cdots x_{ni}\}$.

After standardization, we successfully substitute x_{in}^{*} to represent a certain evaluation index X_i of a country. Then, we introduce:

$$p_{in} = \frac{X_{in}^{*}}{\sum_{i=1}^{k} X_{in}^{*}} \tag{3}$$

According to the concepts of information and entropy in information theory, we can calculate the information entropy of each evaluation index so that we can get:

$$E_i = -\ln(k)^{-1} \sum_{n=1}^{N} p_{in} \ln(p_{\text{in}}) \tag{4}$$

On the basis of information entropy, we further obtain the weight of each evaluation index we defined before:

$$w_i = \frac{1-E_i}{k-\sum E_i} \tag{5}$$

We can then derive the athletic performance Index, and on the basis of calculating these weights, we have:

$$API_n = \sum_{i=1}^{k} w_i X_{in}^* \tag{6}$$

The analytic hierarchy process (AHP) is to decompose the relevant elements of the goal problem into the levels of goals, criteria and plans, and objectively quantify people's subjective judgment with a scale, and substitute a certain degree of objectivity into the subjective. Since this is a traditional method, we ignore its calculation process. Finally, we obtain the test coefficient CR to pass the consistency test [3].

Combined weighting method is to design an algorithm different from subjective and objective weighting, which first uses different methods to assign weights, then uses the algorithm to adjust the weight of different methods, and finally synthesizes the weight value to achieve objective and reasonable weight matching. The game combination weighting model is to find a game equilibrium solution in the weights obtained by different single methods, so as to minimize the difference between the metric weights of each method and the metric weights of game combination weighting. The specific steps are as follows:

First, create a set of weight vectors $Q = \{q_1, q_2\}$ according to the weights of the above two methods, where q_1 and q_2 are the weight vectors of subjective and objective weights respectively, let $b = \{b_1, b_2\}$ be the linear combination coefficients, and the linear combination of these vectors can be obtained:

$$Q = b_1 q_1^t + b_2 q_2^t \tag{7}$$

Second, under the guidance of game theory, the linear combination coefficient b is continuously optimized. The purpose of optimization is to minimize the deviation between Q and q_i, so as to obtain the most ideal index weight vector Q. Using this method, the objective function is determined as:

$$\min \|Q - q_i\|_2, i = 1,2 \tag{8}$$

Thirdly, according to the differential property of the matrix, the first derivation of the optimization corresponding to the above model is as follows:

$$\begin{pmatrix} q_1 q_1^T & q_1 q_2^T \\ q_2 q_1^T & q_2 q_2^T \end{pmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} q_1 q_1^T \\ q_2 q_2^T \end{bmatrix} \tag{9}$$

Fourth, according to the above formula, the weight vector set is obtained and normalized, and the weight of the game theory combination can be obtained:

$$Q = b_1^* q_1^T + b_2^* q_2^T, b_1^* = \frac{b_1}{b_1+b_2}, b_2^* = \frac{b_2}{b_1+b_2} \tag{10}$$

Finally, we get the weights of 12 indicators, as shown in Figure 1:

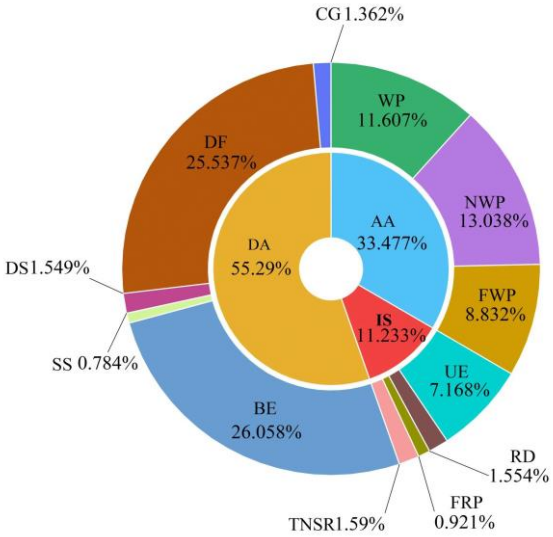

Figure 1: The weights of 12 indicators

2.2 Visualization of the game

In tennis, momentum can be interpreted as a player's morale, confidence, and performance during a match, but these metrics are difficult to quantify, so we use the Athletic Performance Index to replace the reflection of momentum in a tennis match, with scores per match reflecting a player's performance.

Then we apply the evaluation model to the 2nd set and 4th set of 2023-wimbledon-1701 to obtain the scores and performance indicators of players Carlos Alcaraz and Novak Djokovic in each game, as shown in Figure 2:

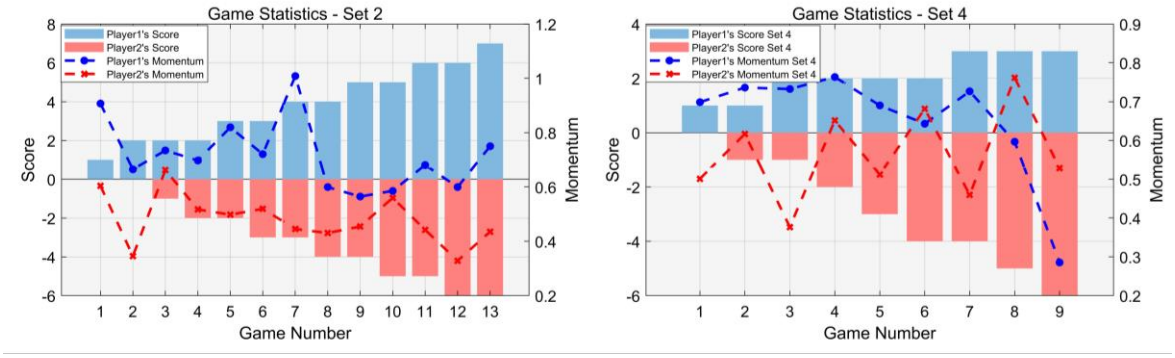

Figure 2: Players' scores and performance reviews

In the second set, we can notice that Carlos Alcaraz's performance index is much higher than Novak Djokovic's, and we can assume that Carlos Alcaraz's momentum is higher than Novak Djokovic's. So in terms of points, despite the tug, Carlos Alcaraz took the final victory in the second set.

In Set 4, Carlos Alcaraz's momentum, the Performance Index API, is declining despite his dominance in the first four games, while Novak Djokovic is fluctuating but still rising. As a result, Novak Djokovic outperformed Carlos Alcaraz in the next few games, while also coming from behind in the score, and eventually became the winner.

3. Momentum index evaluation model

In order to clarify whether momentum is a key factor affecting players' performance, we explored in depth the momentum difference, score difference and other related indicators of players in the game, and constructed a model using Spearman's correlation coefficient to further analyse the relationship between these factors and the outcome of the game, so as to reveal the importance of momentum in the game and the stochastic nature of players' performance [4].

3.1 Spearman correlation coefficient

Spearman correlation coefficient is a non-parametric index that measures the dependence between variables. The monotone equation is used to measure the degree of correlation between hierarchical ordering variables to evaluate the correlation of statistical variables. Spearman's correlation coefficient is usually defined as the Pearson correlation coefficient between rank variables. For a sample size of n, n raw data are converted to registration data, and the correlation coefficient ρ is:

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \tag{11}$$

Spearman's correlation coefficient indicates the correlation direction between independent variable X and dependent variable Y. When X increases, Y tends to increase, and spearman's correlation coefficient is positive. When X increases, Y tends to decrease, and spearman's correlation coefficient is negative. A spearman correlation coefficient of zero indicates that Y is not trending as X increases.

3.2 The Establishment of Model

In order to reflect that the performance of two athletes in the game is related to momentum, we choose to use the momentum difference of the athletes in the game, that is, the competitive performance index difference, to replace the momentum difference, and the score difference of each game to reflect the athlete's performance.

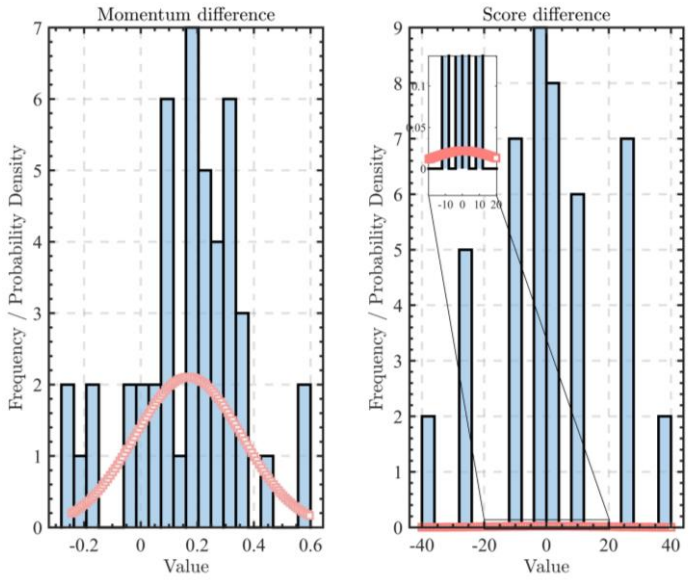

Figure 3: Statistical chart of momentum and score difference

Spearman correlation coefficient analysis is a non-parametric statistical method, which is widely used to analyze different types of data including non-normal distribution. Through data analysis, as

shown in the Figure 3, we can see that the difference of athletic performance index is not normally correlated with the difference of score. At the same time, we can get that there is a monotone relationship between athletic performance index difference and score difference, so we apply spearman correlation coefficient method.

At the same time, we conduct spearman correlation analysis on 12 indicators in the athletic performance index, and the results are shown in Figure 4.

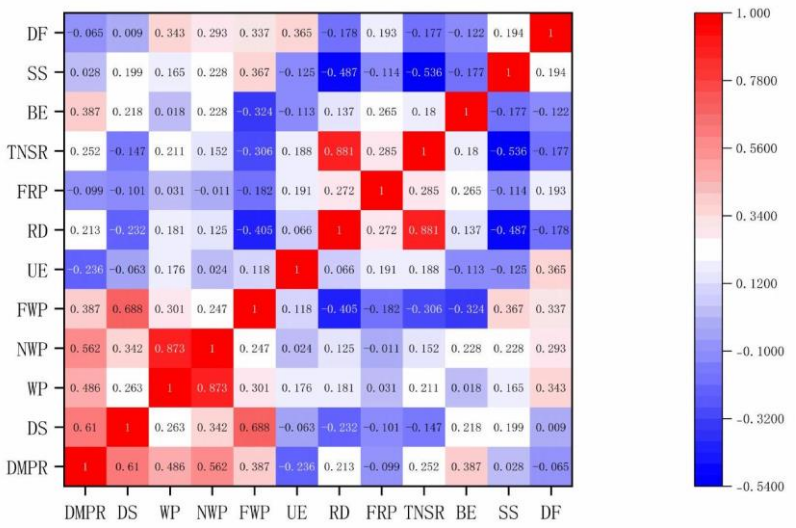

Figure 4: Spearman correlation coefficient analysis

3.3 The Solution of Model

Finally, we get that the correlation coefficient between the competitive performance index difference and the score difference is 0.610, which proves that a player's swing and success in the match are not random, and there is a significant relationship between momentum and match performance. At the same time, we can also draw from the above Figure 4:

1) There is a strong positive correlation between winning percentage in front of the net and winning points, that is, the stronger the player's ability to score in front of the net, the greater the advantage for winning points, and then affect the player's momentum and game performance.

2) There is a strong positive correlation between the batting times of a round and the running distance, which can be explained as the more batting times of a round, the greater the running distance of players, which conforms to the actual situation, and proves the rationality of this model.

4. Transition point judgment model

The data provided by the competition was used to analyse and process the indicators in the athletes' competitive performance metrics. Based on domain knowledge and relevant expert opinions, the turning points of the competition were quantitatively defined and predicted using machine learning, and a turning point judgement model based on the random forest model was constructed [5].

4.1 Data Processing

Before establishing the correlation regression model, we first preprocess the data. Based on the statistical principle, we adopt the boxplot detection method to eliminate outliers, as shown in Figure 5. Boxplot is an outlier detection method based on statistical principles, which uses the quartile and quartile distance of the data to determine whether there is an outlier.

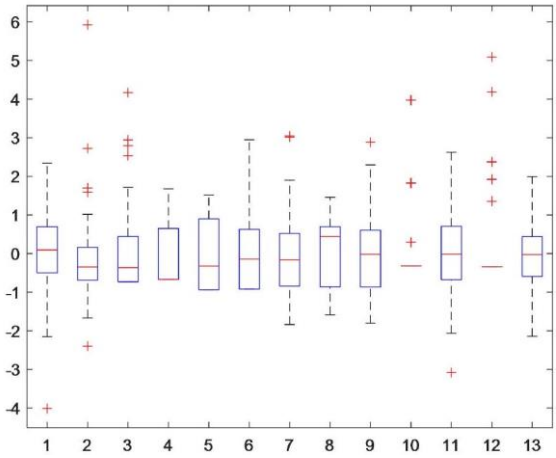

Figure 5: Outliers are eliminated by box plot detection

4.2 Random forest model

Random forest model (RFM) is an ensemble learning model, which trains samples by constructing multiple decision trees and combines their prediction results to meet the prediction accuracy and stability [6].

When dealing with a large number of features and data such as tennis match data, random forest model has great advantages. On the one hand, the decision tree of random forest can train and predict in parallel, which means that the computational efficiency can be significantly improved when dealing with a large number of data. On the other hand, random forest improves the overall prediction accuracy by inheriting the prediction results of multiple decision trees and reduces the error of using a single model.

At the same time, we compare the Gaussian Process Regression model (GPR), Decision tree model (DTM) and SVW model, and their fitting effects are shown in Figure 6 below.

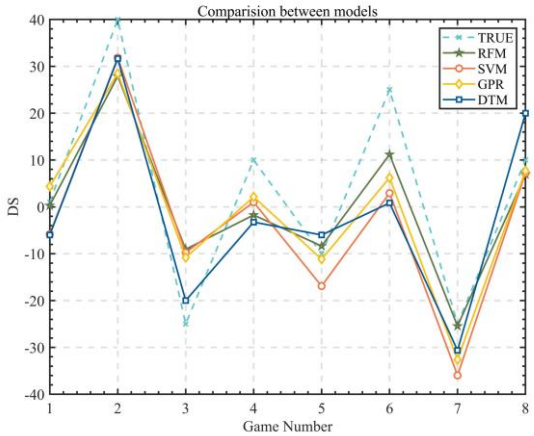

Figure 6: Comparison of four models

Overall, the Random Forest model has the best fit, so we choose the Random Forest model.

In addition, we select data for fitting analysis of turning points, the end result is shown in Figure 7:

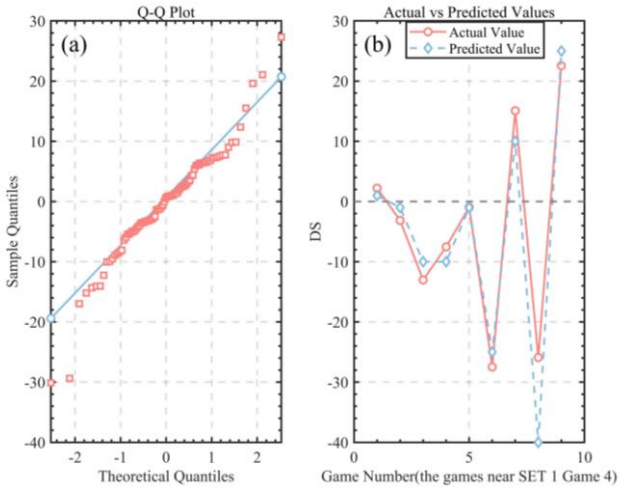

Figure 7: (a) Model fitting analysis (b) Fitting graph near Set1 Game4

As can be seen from the above graph and the turning point fitting analysis, the model has a good fit, so we completed the prediction of the differences in scores across rounds using the random forest model.

Based on the prediction results in Table 2, we think that a point like Set 1 Game 4, for example, is considered a turning point if the margin of victory is negative for two or more consecutive predictions that follow.

Table 2: Game turning point prediction

Set	The game of turning point
Set 1	Game 4
Set 2	Game 11
Set 4	Game 7
Set 5	Game 6

5. Conclusions

In this paper, the competitive performance index model was successfully established, and the momentum of athletes in the match was quantitatively assessed by entropy weighting, hierarchical analysis and game theory combination weighting methods, revealing the actual impact of momentum in tennis matches. Secondly, the results of Spearman's correlation coefficient analysis showed a significant correlation between the score and momentum, confirming the importance of momentum on the outcome of the match. Finally, the Random Forest model was used to successfully predict the turning point of the match, and the significant influence of indicators such as running distance and winning points on the outcome of the match was found.

References

[1] Xia Lei. Research on specialised physical characteristics and training strategies of competitive tennis [J]. Sports Science and Technology Literature Bulletin, 2024, 32(04): 91-94+201. DOI: 10.19379/j.cnki.issn.1005-0256. 2024. 04. 022.

[2] Luo Qian, Li Yongmei, Wang Tenghua, et al. Constructing an evaluation system of rational medication use indicators in clinical departments based on the improved entropy weight method combined with TOPSIS method [J]. Clinical rational use of medication, 2024, 17(15): 170-172+ 177. DOI: 10.15887/j.cnki.13-1389/r.2024.15.049.

[3] He Yinshui, Xiao He, Luo Canghai, et al. Autonomous decision-making of welding position for arc welding of thick plate T-joints based on hierarchical analysis process [J]. Journal of Jilin University (Engineering Edition), 2024, 54(03):

657-662. DOI: 10.13229/j.cnki.jdxbgxb.20220518.
[4] S.Z. Zhao, Y.H. Shin, Y.S. Zhou. A study on the credibility of post qualification of power grid enterprises between zones based on Spearman's correlation coefficient [C]//China's power enterprise management innovation practice (2022). Yunnan Power Grid Limited Liability Company; 2024: 4. DOI:10.26914/c.cnkihy.2024.001241.
[5] Shi Cheng, Yao XF. Research progress of machine learning for Parkinson's disease diagnosis [J]. Chinese Journal of Medical Physics, 2024, 41(05): 640-645.
[6] He Qingxin, Chen Chuanfa, Wang Yuhui, et al. A fusion method for multi-source remote sensing daily precipitation data: a random forest model with consideration of spatial autocorrelation [J]. Journal of Geo-Information Science, 2024, 26(06): 1517-1530.