A study of tennis match trend prediction based on momentum assessment and heuristic HMMs

Xuanjun He[1,*], Changxu He[2], Fuxian Ye[2]

[1]College of physics and optoelectronic engineering, Shenzhen University, Shenzhen, China, 518000

[2]School of design, South China University of Technology, Guangzhou, China, 510000

* Corresponding Author Email: hexuanjun9@gmail.com

Abstract. In competitive tournaments, there exists a currently unexplored factor, momentum, whose role affects the trend of the game to a certain extent and even produces surprising fluctuations, which is particularly evident in tennis.This paper proposes a feasible research programme for the problem of momentum change and trend prediction in tennis matches. Firstly, hundreds of matches of Wimbledon Tennis Championships were used as the data source, and the assessment indexes related to momentum were filtered and mathematical models were established through analysis and processing, and then the coefficient settings were optimised by using the self-looping simulated annealing algorithm. The final model better describes changes in match conditions, with a model fit of over 85% for 87% of the matches.At the same time, in order to verify the existence of correlation between the momentum and the direction of the game, this paper randomly selects a number of games, the first half of the event as a dataset for the momentum model training, and uses the heuristic HMM to predict the subsequent changes and compare them with the scoring difference in the second half of the actual game, and the test result is that the two have basically the same trend, and therefore it can be proved that the momentum change curve can be used as one of the indicators of the prediction of the game trend.

Keywords: Sports Momentum, Heuristic HMM, Match Prediction.

1. Introduction

Momentum is an oft-discussed but relatively vague concept in tennis. Apart from aspects such as a player's own skill and fitness level, momentum seems to be an important factor influencing the outcome of a match. For example, looking back at some of the iconic matches in history, such as Federer's intense duel with Nadal at Wimbledon in 2008 or Djokovic's sudden loss against the rising Spanish star in 2023, we find that momentum played a key role in these matches. In these cases, a situation where one side is clearly in a dominant position, only to lose the lead due to a sudden error or a great performance by an opponent, is often referred to as a turning point in momentum.

Momentum turns are critical for athletes and coaches. If they are able to accurately capture a change in momentum and understand the factors behind such a change, they may be able to respond with appropriate strategies to gain an advantage in the game. However, despite the widely recognised impact of momentum on match outcomes, quantifying and predicting changes in momentum is a complex task.

Existing studies provide some clues in understanding and explaining changes in momentum. For example, Professor Xiangyang Lin and Master Zhiying Huang used BP neural network algorithm to try to establish a prediction model for match performance by analysing tennis match data[1], and Haifu Jiang et al. from Jiangsu University of Science and Technology studied the influencing factors of football matches based on fuzzy grey correlation analysis[2]. In addition, Qi Hang, M.S. of China University of Geosciences and Yang Longfei of North Central University coincidentally used statistical analysis in their studies targeting the winning factors of badminton matches[3,4]. However, although these studies provide some valuable insights, we are still at an exploratory stage as to how to systematically capture changes in momentum and make accurate predictions.

Therefore, this study aims to fill this knowledge gap by proposing a feasible research programme. Taking hundreds of matches of Wimbledon Tennis Championships (https://www.wimbledon.com/) as the data source, we try to establish a mathematical model of momentum that can be relatively accurate by screening the judging indicators to construct the momentum expression through the statistical principle and optimising the parameters by using the self-loop simulated annealing algorithm. Then we predicted the trend of the matches through the heuristic HMM algorithm to verify the practicality of the momentum model. Through this study, we hope not only to better understand the role of momentum in the game, but also to provide more effective game strategies for athletes and coaches.

2. The Momentum modelling based on statistical foundations and self-recycling simulated annealing

2.1. Screening of assessment indicators

Since momentum is not a real physical quantity, it needs to be given a basic definition. In this paper, the score difference is used as a direct reflection of the overall momentum increase or decrease. At the same time, before establishing the momentum model, it is necessary to choose appropriate judging indicators to measure the change of momentum in the game.

For a tennis match, its complete data includes a number of indicators such as the number of sparring rounds, scoring rounds, return position, key points won, running distance and so on. Prior to the actual analysis, the data were transformed and merged to create a link with the score, for example, the running distance of two players was differed to reflect the physical exertion problem. In order to further simplify the problem, the rounds of scoring were used as the extraction criterion to count the number of times each indicator appeared, combined with the reference to the eighteen technical statistical indicators of singles matches listed on the official website of the Professional Tennis Federation[5], thus eliminating certain non-essential considerations, including the return position, the reception forehand and backhand, etc. The remaining indicators were then analysed to determine the number of times each indicator occurred. The remaining indicators were then normalised in order to eliminate the differences in magnitude between different indicators and make them comparable. Finally, we applied the Spearson's coefficient method to filter the key indicators. Spearman's correlation coefficient is able to quantify the trend of change between two random variables, and can be used to reflect the degree of correlation of data (electricity consumption) between regions.[6]

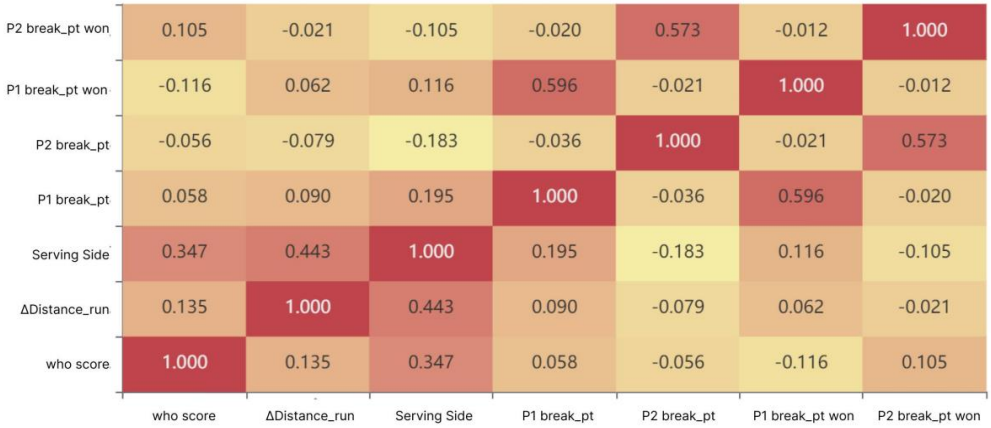


Figure 1. Heat map of the correlation coefficient matrix

The results obtained are shown in Figure 1, demonstrating the correlation between the different indicators. In the heat map, darker colours indicate stronger correlations, while lighter colours indicate weaker correlations. Based on the results of these analyses, we chose Points, Serve Points, Direct Serve Points, Break Points, Double French Errors, and Running Distance Differences as the key metrics for the subsequent building of the momentum model. These indicators cover the technical,

strategic and physical aspects of the match, which can comprehensively reflect the dynamic changes of the match, thus providing a solid foundation for the establishment of the momentum model.

2.2. Momentum modelling based on statistical foundations

In the previous section, it was mentioned that there is a mapping relationship between momentum and score, and considering the results of the analyses in Figure 1.1 and the fact that momentum change is a process that accumulates over time, the initial expression for the momentum model is as follows:

$$M_t = M_{t-1} + f(P_t, S_t, C_t, A_t, B_t, F_t, D_t) \tag{1}$$

P_t — gain/loss points(±1)、 S_t — serve weight 、 C_t — continuous score factor 、 A_t — *Ace* 、 B_t — break serve 、 F_t — double fault 、 D_t — poor standardised running distances

It can be observed that the screened indicators are all directly related to the score and have a cumulative effect, so it is considered that there is a linear relationship between them and momentum, and according to the statistical analysis in 1.1, the indicators are assigned initial coefficients, and so the formula can be specified as:

$$M_t = M_{t-1} + 1.2P_t \cdot C_t + 0.01A_t + 0.07B_t - 0.01F_t + D_t \tag{2}$$

In Figure 2, the effect of successive different scores on the next score is reflected. After 2 consecutive points, there is a limited increase in potential energy, after 3 consecutive points there is a significant increase in aura, and after 4 consecutive points there is almost no increase in potential energy. Numerical relationship equations were then established:

$$C_t = -0.2 \cdot x^3 + 1.2 \cdot x^2 - 2 \cdot x + 2 \quad (x = 1,2,3,4) \tag{3}$$

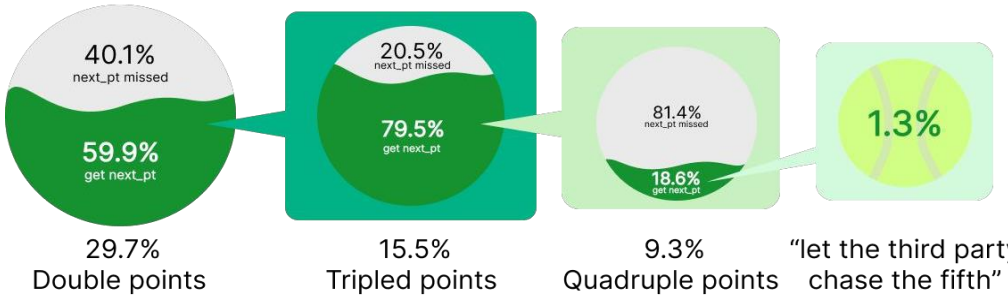


Figure 2. Statistical results of the impact caused by consecutive scores

After all the coefficients were reasonably determined, in order to test the model, a game was randomly selected to produce the results of the momentum change curve compared with the change in the percentage of points scored as shown in Fig. 3, which showed the same trend and verified the reasonableness of the model.

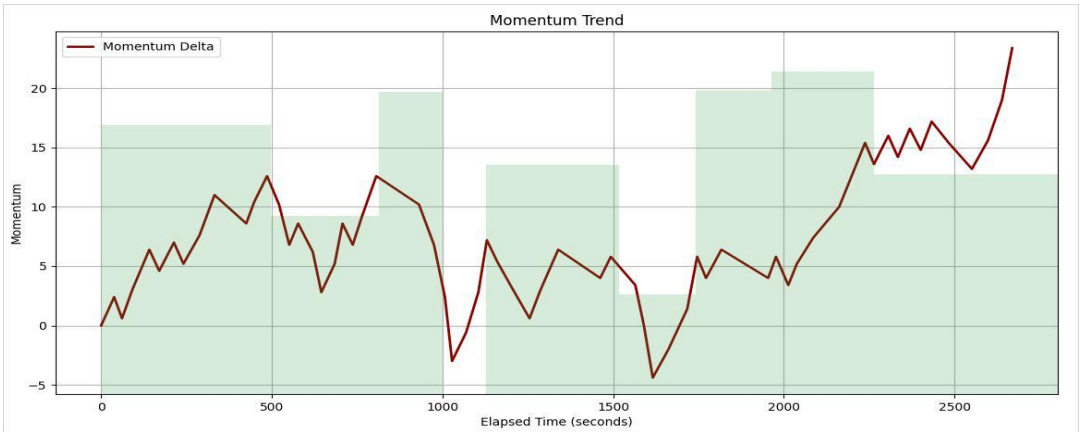


Figure 3. Comparison of Momentum Change and Score Share

2.3. Parameter optimisation based on Self-looping SAA

When establishing the momentum model, the fitting effect of the model may fluctuate to a certain extent when testing different competitions due to the subjectivity of choosing the coefficients of each indicator. In order to improve the stability and accuracy of the model, we introduced a self-looping simulated annealing algorithm for parameter optimisation.

The simulated annealing method uses Metropolis law for its inner loop and a cooling process for the outer layer. Because the simulated annealing algorithm has advantages in dealing with optimisation problems that other algorithms do not have, it has gained widespread attention and popularity among many scholars. Compared with other optimisation algorithms, the simulated annealing algorithm accepts a solution worse than the current one with a certain probability, and the computational process is simple, general and robust, and it is easier to find the global optimal solution through the heuristic probabilistic searching method to avoid falling into the local optimal solution[7]. However, for complex optimisation combination problems, the GA using random selection for searching has the problems of finding the local optimal solution in advance and poor search performance. In[8], we improved the algorithm to find the global optimal solution by setting multiple starting solutions through self-loop.

The key of this step is how to determine the judgement criteria. According to the momentum model, the momentum is correlated with the score difference, so the two values can be normalised and differed, and then the maximum of the absolute values is taken as the reflection of the fitting effect, i.e.:

$$\overline{M}_t = ([m_1, m_2, \cdots, m_{t-1}, m_t] - m_{\min}) / (m_{\max} - m_{\min}) \tag{4}$$

$$\Delta\overline{P}_t = ([\Delta p_1, \Delta p_2, \cdots, \Delta p_{t-1}, \Delta p_t] - \Delta p_{\min}) / (\Delta p_{\max} - \Delta p_{\min}) \tag{5}$$

$$\mathrm{F}itness = 1 - \left|\overline{M}_t - \overline{P}_t\right|_{\max} \tag{6}$$

Using 70% of the training set of 124 games and 30% as the test set, the optimised momentum expression is as follows:

$$M_t = M_{t-1} + 1.774P_t \cdot C_t + 0.148A_t + 0.054B_t - 0.818F_t + 1.359D_t \tag{7}$$

The fit based on it is shown in Table 1:

Table 1. Fitting effects

Fitness	>0.95	>0.90	>0.85
Match count	7(5.7%)	33(27.4%)	108(87.1%)

Not only that, the algorithm has another advantage in that it can be trained on different match data to optimise the output factor. Considering the existence of matches with different intensity levels such as gelling or crushing, the fitting results for the same type of matches are above 0.85 and reach 0.90 in 80.6% of the cases through targeted training.

3. The heuristic HMM-based match prediction

3.1. Model Selection

Tennis match trend forecasting is a time series forecasting problem. Traditional algorithms include time series ARIMA and Markov models. However, time series in practice usually have linear and non-linear characteristics, and a single traditional time series prediction method is a linear model, which shows some limitations in modelling time series[9], and is not applicable to real matches where the time taken for each score is different. In addition, match trends are also affected by some hidden factors, leading to limitations in Markov models based on visual data.

Hidden Markov Model (HMM) is a probabilistic model that can be used to predict sequential problems, HMM was first applied in speech recognition with great success. It has since been widely used in pattern recognition, image recognition, bioinformatics science, fault diagnosis and other fields with promising results.[10] has high recognition accuracy due to its short training time and no need to preconstruct the objective function. We believe that it is advantageous to use HMM to predict non-normally distributed and irregular time interval data.

3.2. Heuristic HMM modelling

Compared with the traditional HMM, heuristic HMM is able to capture the relationship between data more accurately by introducing domain knowledge and a priori information, thus improving the prediction accuracy. It is an effective means of solving nonlinear and nonconvex optimisation problems, with more advantages in terms of model representation, robustness, interpretability and generalisation ability. A swarm intelligence optimisation algorithm for global search was proposed in 1995 by observationally simulating the migration and flocking behaviour of birds during foraging.[11]

In this paper, Bayesian regularisation is used for model optimisation, where regularisation is conceived as the introduction of a priori information to compensate for the loss of information caused by the measurement process.[12] By normalising the coefficients, the degree of influence of each factor on the momentum is reflected as a prior probability distribution, which is introduced into the likelihood function of the model as a regularisation term. Through this process, the heuristic HMM model was completed.

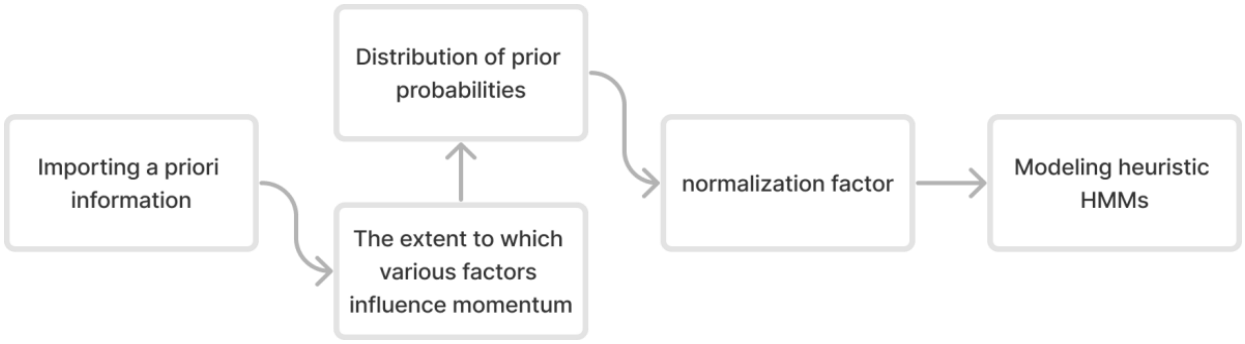


Figure 4. Flowchart of heuristic HMM algorithm

3.3. Test of match prediction results based on the momentum model

In this study, we examined the results of the momentum-based model for match prediction to assess the predictive performance and accuracy of the model. Specifically, the following steps were used to conduct the test:

Firstly, the data of a complete tennis match was divided into a training set and a test set, and the momentum function expression was constructed with the training set, introduced into the HMM model through Bayesian regularisation to predict the subsequent results, and compared with the test set, with the basis of the evaluation being the same as the method used in 1.3.

Table 2. Fitting effects based on different training sets

scale as a training set	50%	60%	70%
validity	0.599	0.626	0.754

As shown in Table 2, the fitting effect of using 70% as the training set is 0.754, which indicates that the momentum-based model-heuristic HMM algorithm is able to effectively predict the trend of the match. However, since momentum is not the only criterion for deciding matches, the model is suitable for providing a certain reference role.

In addition to this, this paper also conducts a sensitivity test on the prediction performance of the model, including four indicators: accuracy, precision, recall and F1 score.

Table 3. Sensitivity test

Tennis race（random selection）	Accuracy	Precision	Recall rate	F1 Score
1	0.743	0.442	0.454	0.491
2	0.652	0.326	0.362	0.497
3	0.721	0.500	0.411	0.386
4	0.674	0.419	0.500	0.430

From Table 3, the latter three indicators converge to 0.5, which coincides with the actual situation that momentum is based on the fact that the performance of the two players will fluctuate in both positive and negative directions, indicating that the prediction method is reasonable and objective.

4. Conclusions

In this study, we explored the importance of momentum in the game and the feasibility of the prediction model by building a prediction model for the trend of tennis matches based on momentum assessment and heuristic HMM. Firstly, we constructed a momentum model based on statistical foundation with self-looping simulated annealing by screening assessment indicators and building mathematical models. After parameter optimisation, the model demonstrated a high degree of fit on a training set of Wimbledon Tennis Championships data. Secondly, we explored a match trend prediction method based on heuristic HMM, which can capture the relationship between data more accurately and improve the prediction accuracy compared to the traditional HMM model. Finally, by comparing the actual match data, we verified that there is a high correlation between momentum changes and match trends, and demonstrated the feasibility of predicting match trends through the heuristic HMM.

In summary, this study provides a new method and idea for understanding and predicting tennis match trends, which is instructive for players and coaches to formulate match strategies. However, there are still some limitations in this study, such as the selection of dataset and the determination of model parameters, which are yet to be further optimised and improved. Future research can consider expanding the data scope, increasing the complexity of the model, and combining more domain knowledge to further enhance the predictive ability and practicality of the model.

References

[1] C.Y. Huang. Research on tennis match performance prediction model based on BP neural network [D]. Fujian Normal University, 2022.

[2] JIANG Hai-Fu, YU Hua-Long, WEI Lei. Research on the analysis of influencing factors of soccer game based on fuzzy gray correlation analysis [J]. Computer and Digital Engineering, 2023, 51(03): 555-560.

[3] Qi Hang. Discriminant analysis and outcome prediction of winning factors in ATP men's singles matches [D]. China University of Geosciences (Beijing), 2014.

[4] YANG Longfei, LIU Qingyao, ZHANG Jiangnan. Research and analysis on analyzing and predicting the results of men's singles badminton matches in China and the structure of winning and losing[C]//China Wisdom Engineering Research Association, China Bandi Association, Guangdong Physical Fitness Association. Proceedings of the 10th China Physical Training Science Conference (in Chinese). [Publisher unknown], 2023: 5.

[5] Rong Zhang. Result prediction and player analysis of tennis matches [D]. Yunnan University, 2023.

[6] WU Jun-Ying, LU Xin, LIU Hong et al. Ultra-short-term multiregional power load forecasting based on Spearman-GCN-GRU model [J/OL]. China Electric Power, 1-9 [2024-03-16].

[7] Kang Wenxuan. Urban delivery path planning for shared bicycle based on simulated annealing algorithm [J]. Technology and Innovation, 2022(13): 104-106+109.

[8] HAN J X, MA M Y, WANG K. Product modeling design based on genetic algorithm and BP neuralnetwork []. Neural Computing and Applications, 2021, 33(9): 4111-4117

[9] Ni Xuan, Chen Yichen. ARIMA model combined with LSTM neural network to establish a prediction model for drug purchasing decision [J]. Heilongjiang Science, 2024, 15(02): 76-78.

[10] Wang Xinwu. Research on online video personalized recommendation based on Hidden Markov Model [J]. Digital Communication World, 2023(09): 1-4.

[11] REN Yongfeng, XUE Yu, YUN Pingping et al. Markov prediction of multi-objective optimized energy storage system for smoothing power fluctuation in wind farms [J]. Power System Automation, 2020, 44(06): 67-74.

[12] Zhao Yue, Zhang Xueyong. Research on the application of Bayesian regularization in spherical equivalent source method [J]. Journal of Anhui University of Engineering, 2023, 38(02): 89-94.