ISSN 2618-1576 Vol. 6, Issue 3: 29-36 DOI: 10.25236/FSR.2024.060304

Prediction of the Swings in Plays Based on Momentum of Tennis Matches

Xinlin Zhu[1,#], Shuojie Wang[2,#], Zhaonan Wu[1,#]

[1]School of Automotive and Traffic Engineering, Jiangsu University, Zhenjiang, 212013, China
[2]College of Computer and Information Engineering (College of Artificial Intelligene), Nanjing Tech University, Nanjing, 211816, China
[#]These authors contributed equally.

Abstract: *Tennis is an increasingly popular sport worldwide. This paper focuses on the study of "momentum" in tennis matches, which is utilized to predict match fluctuations. In this paper, we quantify the momentum about physical momentum and construct an evaluation model. Using principal component analysis to analyze the long-term influencing factors, four principal components are obtained; considering that the influence of the actual game situation has periodical timeliness, a mathematical model of the short-term factors is constructed. Then, based on the established evaluation model, this paper quantifies the fluctuation, and based on the data of the finals, conducts ARIMA time series prediction of the fluctuation, and uses the neural network model to determine the most relevant factors affecting the fluctuation.*

Keywords: *Principal Component Analysis, ARIMA Time Series Prediction, Neural Network*

1. Introduction

As a globally recognized competitive sport, professional tennis has always been in the spotlight[1]. With the development of time and the rise of data science, a series of new technologies have been applied to the sport, generating a large amount of athletes' game and training data. Subsequently, researchers in the field of data science began to incorporate knowledge to solve momentum problems. Previous researchers have incorporated Momentum into race models by integrating a form of motion. This form of motion proposes to represent these moments in terms of conditional probabilities and empirical Bayesian estimates, ultimately merging Monte Carlo simulations through a unified hybrid approach based[2-3]. This approach is subsequently applied to volleyball data to significantly enhance the prediction of match outcomes. Researchers have also delved into the complexity of basketball coaches' emotions and perceptions related to technical fouls (TFs). Through detailed thematic content analysis, the coach's ability to identify and manipulate the psychological "momentum" on the court was enhanced, and the coach was able to use a thorough gatekeeper. However, it has been found that analyses conducted in the field of data science have been less frequently performed in tennis, mainly due to the complexity of the rules of the game, the large number of rounds in each match, and the large sample size, which makes data processing challenging. This paper attempts to address this gap by collecting and studying data from the 2023 Wimbledon Championships, quantifying and analyzing the "momentum" of the athletes' matches, predicting the flow of the matches when scoring occurs using ARIMA, BP neural network and decision tree models, and identifying which players will score at particular points in the match. The ARIMA model, BP neural network model and decision tree model were used to predict the flow of the game at the time of scoring, and to identify which players were performing better at specific times in the game. (Data from the website: www.comap.com)

2. Establishment and Application of the Evaluation Index Model of Athletes' "Momentum"

2.1 Quantification of The Concept of Momentum

For tennis players to score in a tennis match depends on the comprehensive ability of their own technical experience. However, athletes will encounter many uncertain influencing factors in the actual match, such as: serving errors, the influence of mindset brought about by the opponent's scoring, and so on. As the match time passes, many uncertain influences will be synthesized and processed into one

ISSN 2618-1576 Vol. 6, Issue 3: 29-36 DOI: 10.25236/FSR.2024.060304

influencing factor that affects the athlete's level throughout the match time.

This essay is analogous to the definition of momentum in classical mechanics $P = M \cdot V$ [4], and M is analogous to C (a composite metric) to explain a player's ability to maintain strength,which is used as aggregate metrics of the Momentum Indicator Evaluation Model in this article. In classical mechanics, the greater the mass, the more difficult it is to change the state of motion of an object. Similarly, the higher the value of the "C" index, the more stable their on-field performance is, and the less likely they are to be disturbed by their opponents. Secondly, for stability, mass is constant in classical mechanics, which is consistent with the characteristics of the composite index established based on the analysis of past performance. v is analogous to F (impact factor) to reflect the changes in the performance of the player. In classical mechanics, a change in velocity indicates that the state of motion of an object is changing. In tennis, changes in the flow of the game[5-6] (e.g., scoring consecutive points, hitting unstoppable winners, etc.) can greatly affect the direction of the game, and the paper can respond to this through changes in the value of the impact factor. By analogy, this paper gives a formula for measuring the "momentum" of a player's game:

$$M = C \cdot F \tag{1}$$

2.2 Development of Composite Indicators

2.2.1 Determine Model Parameter Variables

The establishment of a comprehensive ability of the athlete's judging indicators is based on the accumulation of long-term competition experience, embodied in the data in the successive rounds of the competition, the paper needs to analyze the data-related variables that can measure the athlete's ability. Considering the large number of its related variables, based on the data preprocessing in the previous section, the related indicator variables are listed:

Serve side scoring rate under x_1, average number of serves x_2, round scoring rate x_3, serve direct scoring rate x_4, unanswered ball scoring rate x_5, error rate x_6, net scoring rate x_7, average distance moved/per round x_8, speed_mph x_9, number of sets ahead of opponent when winning x_{10}.

Due to the excessive number of indicator variables, a dimensionality reduction process was considered and an appropriate principal component analysis was chosen[7].

2.2.2 Principal Component Analysis (PCA) Process

The paper plotted a mixture of line and bar charts for the Cumulative Variance Explanation Rate as in Fig 1.

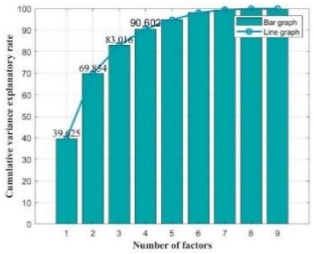


Figure 1: Principal Component Expression Variable (PCEV) Plot

It was found that the variable can be well represented when the principal component variable is five (94.776%) and after five, the variation curve of the analyzed graph tends to flatten out with less variation, so the parameter of the number of principal components is adopted as five.

Build the modeling equation:(F_n is the impact factor of the momentum indicator evaluation model)

$$C = \frac{0.396}{0.906} \times F_1 + \frac{0.302}{0.906} \times F_2 + \frac{0.132}{0.906} \times F_3 + \frac{0.076}{0.906} \times F_4 \tag{2}$$

ISSN 2618-1576 Vol. 6, Issue 3: 29-36 DOI: 10.25236/FSR.2024.060304

Where: F_1, F_2, F_3, F_4, and F_5 are the analyzed principal components, and have a linear relationship with the previous 10 relevant indicator variables x1-x10, as follows:

$$F_1 = 0.264x_1 - 0.177x_2 + 0.061x_3 + 0.182x_4 + 0.014x_5 - 0.242x_6 + 0.177x_7 + 0.075x_9 + 0.216x_{10} \quad (3)$$

$$F_2 = -0.037x_1 + 0.209x_2 + 0.343x_3 - 0.168x_4 + 0.237x_5 + 0.138x_6 + 0.26x_7 + 0.013x_9 + 0.183x_{10} \quad (4)$$

$$F_3 = 0.178x_1 + 0.228x_2 - 0.02x_3 + 0.333x_4 + 0.319x_5 - 0.079x_6 - 0.03x_7 - 0.721x_9 - 0.143x_{10} \quad (5)$$

$$F_4 = 0.054x_1 - 0.002x_2 - 0.338x_3 + 0.268x_4 + 0.927x_5 + 0.133x_6 - 0.234x_7 + 0.574x_9 - 0.117x_{10} \quad (6)$$

2.3 Development of Impact Factors

Since the impact factor is a short-term variable generated in real-time, the paper can not define the parameters first, in the parameter value of the correlation analysis, our thinking is: according to the past game data, infer a strong correlation between a few types of parameters, the paper carry out correlation tests to verify the conjecture[8]. Then some of the parameters are normalized, in the definition of indicator variables for analysis, the explicit procedure corresponds to the following steps.

2.3.1 Determination of Impact Factor Parameters

The selected indicator variables are shown in Table 1.

Table 1: Indicator Variables

Notation	Definition	Value
y_{1n}	Whether the player lost the set point score on two service errors	0 or -1
y_{2n}	Whether the player gained a set point on a non-serve	0 or 1
y_{3n}	Whether the player scored a set point off serve	0 or 1
y_{4n}	Whether the player failed to score a set point on an off-serve.	0 or -1
y_{5n}	Whether the player scored a direct point on serve	0 or 1
y_{6n}	Did the player lose a point due to an error	0 or -1
y_{7n}	Did the player hit an unanswered shot to score a point	0 or 1
y_{8n}	Whether the player scored a set point	0 or 1
y_{9n}	Whether the player scored a set point	0 or 1
y_{10n}	Is the player scoring a set point	0 or 1
y_{11n}	Whether the player has scored an inning-point	0 or 1
y_{12n}	Has the player achieved consecutive runs in past T overs? (j=0,1,...<T)	0 or e^{j-1}
y_{13n}	Is the player's running distance below average for the match	-1 or 1
y_{14n}	Whether the player serves	0 or 1

Note: n indicates the number of rounds that have elapsed since the start of the match.

ISSN 2618-1576 Vol. 6, Issue 3: 29-36 DOI: 10.25236/FSR.2024.060304

2.3.2 Modeling Process

Taking the average round time of the match as the interval t, incremental scoring weights are assigned to the time segments formed by the time series, and in the nth time interval, the scoring weights are as follows.

$$N_n = (1+\mu)^n \tag{7}$$

μ is the newly defined time distribution coefficient, At is the average round time of the game, Tt is the total time of the game, and the relationship between the three indicated as:

$$\mu = \frac{At}{Tt} \tag{8}$$

Considering that the athlete has an advantage when he/she is on the serving side, defined α as the serving weight coefficient, Sp as the serving percentage of this athlete, and Spc as the opponent's serving percentage:

$$\alpha = \frac{Sp}{Spc} - 1 \tag{9}$$

When the player is at the nth match interval, compute the score a_n resulting from the player's performance in the round, a_n being the series over time and denoted by S_n as the sum of its first n terms:

$$a_n = (1+\alpha y_{14n}) N_n \sum_{i=1}^{13} y_{in} \tag{10}$$

Taking into account the player's ability to regulate, the performance of the player's previous rounds of scoring on their impact is time-sensitive, and the need to calculate a period T, T is the average number of rounds of each game.

The cumulative sum of the player's per-round performance scores over the period T is used as the numerical magnitude of the final impact factor F_n, which represents the athlete's impact factor at the nth time interval. The athlete's impact factor F_n will be segmented because the athlete's previous rounds performance is less than one time-lapse period number T when the athlete first takes to the field of play:

$$F_n = \begin{cases} S_n & , \quad n \leqslant T \\ S_n - S_{n-T} & , \quad n > T \end{cases} \tag{11}$$

To summarize, the expression for the athlete's "momentum" in the n-th round of a match is:

$$M_n = C \cdot F_n \tag{12}$$

3. Proof That The Swings in Play are Not Randomized

To determine who would have scored in the match, it is sufficient to compare who had more momentum in that round. The proof is as follows: In the 2023 Wimbledon-1601 semifinal match between Carlos Alcaraz (P1) and Daniil Medvedev (P2), for example, the two players played a total of 159 rounds. The difference in momentum between the two players is calculated by subtracting the Mn of player 1 from the Mn of player 2, denoted as Mnd, and the results are shown below in Fig 2. The horizontal axis

ISSN 2618-1576 Vol. 6, Issue 3: 29-36 DOI: 10.25236/FSR.2024.060304

shows the number of rounds and the vertical axis shows Mnd.

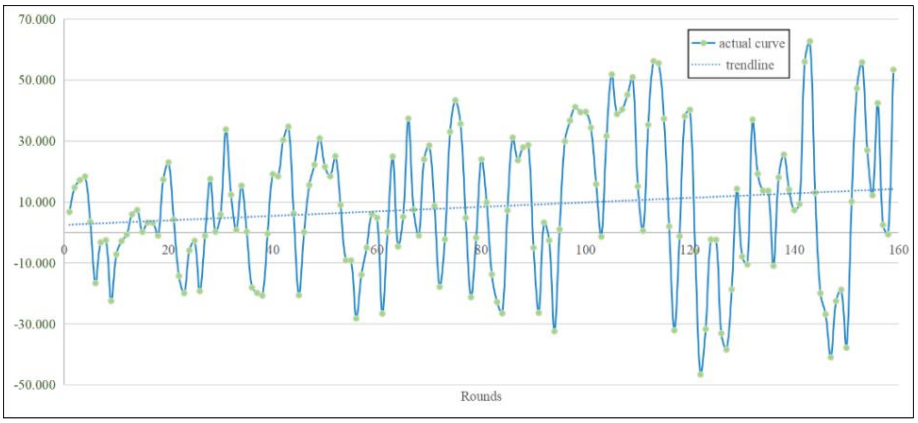


Figure 2: Momentum Difference Plot`

If Mnd>0 it means that player 1 scored in the round and player 2 did not score in the round; if Mnd<1 it means that player 1 did not score in the round and player 2 did.

Let's replace player 1's score with number 1 and player 2's score with number 2. This gives us a graph of predicted vs. actual values for the first 45 rounds in Fig 3. The horizontal axis represents the number of rounds and the vertical axis represents the values.

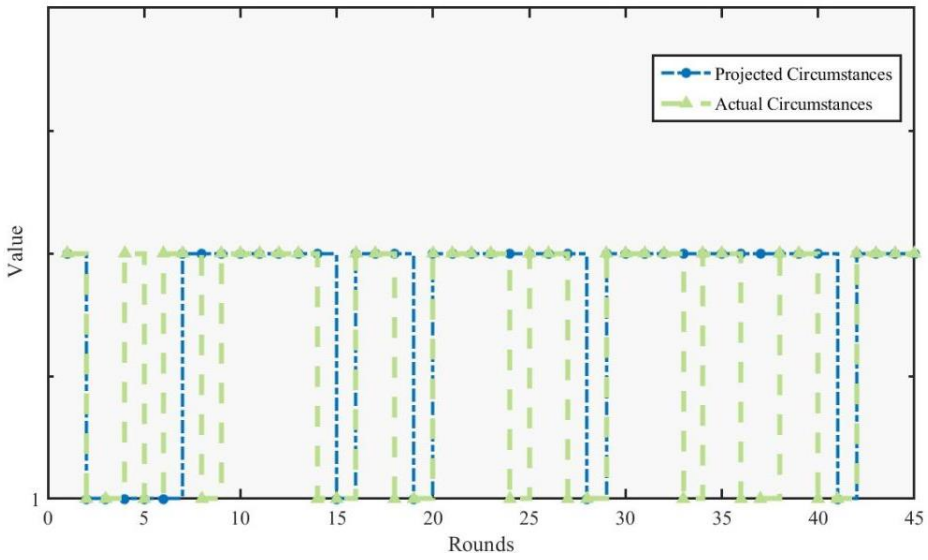


Figure 3: Plot of Predicted vs. Actual Values

As can be seen from the Fig, the predicted and actual values highly overlap. With this method, this paper calculates the agreement rate of the overall data and comes up with an agreement rate of 74.21%. It can be proved that the fluctuation of the match is not random**.**

4. In-depth Analysis of Swings

4.1 Quantification of Swings

In this paper, fluctuations are interpreted as changes in the flow of the game. The quantification of fluctuation is based on the "Momentum" evaluation model established above. First, the momentum Mn of the players is calculated, and then the fluctuation On is defined as Momentum fluctuations of players in the n-th round after the start of the match, computed as the quotient of the difference between the momentum of a player's current round and the difference of the momentum of the previous round, with the expression: (Note: a negative sign indicates a bias in favor of the (Note: a negative sign indicates a bias in favor of the opponent)

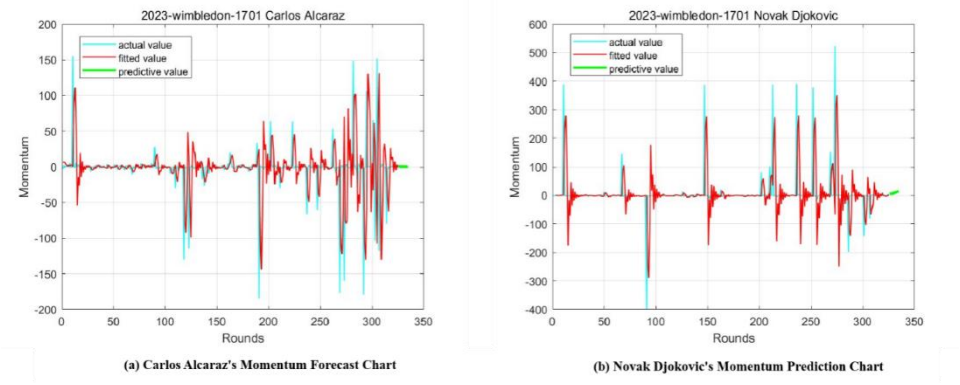
ISSN 2618-1576 Vol. 6, Issue 3: 29-36 DOI: 10.25236/FSR.2024.060304

$$O_n = -\frac{(M_{(n+1)} - M_n)}{(M_n - M_{(n-1)})} \tag{13}$$

4.2 Building an ARIMA time series forecasting model to predict swings

By analyzing the data, it can be found that the fluctuation data on the game's flow is distributed over time, so the ARIMA time series prediction model is built to predict these fluctuations in the game[9].

The ARIMA model based on data from match 2023-wimbledon-1701 was built to fit the change in players' momentum during the match and predicted the magnitude of Carlos Alcaraz and Novak Djokovic's momentum if the match had not yet ended. This is shown in the following Fig 4.



(a) Carlos Alcaraz's Momentum Forecast Chart (b) Novak Djokovic's Momentum Prediction Chart

Figure 4: ARIMA Prediction Results Chart

Carlos Alcaraz's time series is 0.921 and Novak Djokovic's time series is 0.935, both close to 1 in absolute value, which is a good fit.

4.3 Using the BP Neural Network Algorithm to Determine the Contribution

4.3.1 Standardization of Swing O_n Indicators & Determination of Input Layer

To determine if there are metrics that can help determine when the flow of the game is about to change from favoring one player to another, the swings O_{n1} of the finals of the player Carlos Alcaraz were transformed into the player's round scores, i.e., the swings O_{n1} were normalized to be $Y_n = 0$ when $O_{n1} < 0$, and $Y_n = 1$ when $O_{n1} > 0$.

Based on the data indicators measuring the composite indicator C and the impact factor F_n in Problem 1, the following factors were selected in Table 2:

Table 2: Parameter Table

Parameters	Definition	Parameters	Definition	Parameters	Definition
x_1'	score difference	x_6'	an untouchable shot	x_{11}'	block print
x_2'	current server	x_7'	serve an ace	x_{12}'	opponent error
x_3'	successive scores	x_8'	serve width	x_{13}'	opponent loses points
x_4'	breakpoint	x_9'	serve depth	x_{14}'	turn run distance
x_5'	game point	x_{10}'	serve an ace	x_{15}'	winnershot_type

4.3.2 Implementation of The Model

Momentum fluctuations during matches were explored based on data related to five matches played by Spanish rising star Carlos Alcaraz at the 2023 Wimbledon Open. The data was divided in a ratio of

ISSN 2618-1576 Vol. 6, Issue 3: 29-36 DOI: 10.25236/FSR.2024.060304

7:3 for the training and test sets and analyzed using the BP neural network algorithm:

A BP neural network classification model[10] was built to train and test the dataset, and the results were as follows.

The accuracy of the training set is 89.45% and the accuracy of the test set is 82.80%. Therefore, the BP neural network is well-trained.

The contribution of each factor variable is shown in Table 3 below.

Table 3: Table of Contribution Rates

Parameters	x_1'	x_2'	x_3'	x_4'	x_5'	x_6'	x_7'	x_8'
Contribution Rate	6.81%	6.56%	10.72%	4.90%	8.11%	7.05%	5.53%	8.57%
Parameters	x_9'	x_{10}'	x_{11}'	x_{12}'	x_{13}'	x_{14}'	x_{15}'	
Contribution Rate	5.33%	7.77%	8.46%	4.27%	4.69%	7.28%	3.96%	

Based on the contribution rates in the table above, the four variables with the highest values were used as the variables with the highest correlation with winning scores. The x_3' consecutive scores had the highest correlation with the standardized variables with a contribution rate of 10.7%, followed by x_8' serving styles, x_{11}' scores at the net, and x_5' in-set scores.

5. Conclusions

This paper studies the scoring process of tennis matches, and establishes the "momentum" evaluation model and the prediction model of scoring fluctuation through physical connection and based on mathematical knowledge. In this paper, the principal component analysis method is well utilized, and the 9 relevant factors are downgraded to 4 principal components, and the application effect of the "momentum" evaluation model is also better, with the consistency rate reaching 74.21%; during the quantification of fluctuation, this paper innovates and repeats the definition of fluctuation, and predicts it with the ARIMA time-series model, with the fitting accuracy of 0.921 and 0.921. In the process of quantifying fluctuation, this paper innovatively repeats the definition of fluctuation and predicts it with ARIMA time series model, and the goodness of fit reaches 0.921 and 0.935; finally, the BP neural network algorithm is used to determine that the relevant factor with the largest contribution rate is the consecutive scores of players, and the contribution rate reaches 10.7%.

References

[1] Chen, H. (2022). A data mining-based model for evaluating tennis players' training movements. Discrete Dynamics in Nature and Society, 2022.
[2] Skublewska-Paszkowska, M., & Powroznik, P. (2023). Temporal pattern attention for multivariate time series of tennis strokes classification. Sensors, 23(5), 2422.
[3] Wang, J. (2022). Mining and prediction of large sport tournament data based on bayesian network models for online data. Wireless Communications & Mobile Computing (Online), 2022.
[4] Xu, W., Liu, Q., Koenig, K., Fritchman, J., Han, J., Pan, S., & Bao, L. (2020). Assessment of knowledge integration in student learning of momentum. Physical Review. Physics Education Research, 16(1).
[5] Berhimpong, M. W., Mangolo, E. W., Makadada, F. A., Hadjarati, H., Perdana, G. S., & Ilham. (2023). Exploring the impact of drills training and grip strength on tennis serve performance: A factorial experimental design research. Journal of Physical Education and Sport, 23(11), 3108-3118.
[6] Singh, A., Kaur Arora, M., & Boruah, B. (2024). The role of the six factors model of athletic mental energy in mediating athletes' well-being in competitive sports. Scientific Reports (Nature Publisher Group), 14(1), 2974.
[7] Jha, C., & Barnett, I. (2022). Confidence intervals for the number of components in factor analysis and principal components analysis via subsampling. Ithaca: Retrieved from https://www. proquest.com/ working-papers/confidence-intervals-number-components-factor/docview/ 2662171673/se-2.
[8] Aslam, M. (2024). Analysis of imprecise measurement data utilizing z-test for correlation. Journal of

ISSN 2618-1576 Vol. 6, Issue 3: 29-36 DOI: 10.25236/FSR.2024.060304

Big Data, 11(1), 4.
[9] Wang, Z., Tang, J., Hou, S., Wang, Y., Zhang, A., Wang, J., Han, B. (2023). Landslide displacement prediction from on-site deformation data based on time series ARIMA model. Frontiers in Environmental Science.
[10] Liu, H., Xu, Y., Wang, C., Ding, F., & Xiao, H. (2022). Study on the linkages between microstructure and permeability of porous media using pore network and BP neural network. Materials Research Express, 9(2), 025504.