Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 11, No. 3, 2024

Research on the Prediction of Tennis Match Momentum Based on BP Neural Network and Linear Regression Analysis

Jiaqi Guo [1], Jianju Zhang [1], Zhong Wang [1], Xingrong Zheng [1,*], Kaiqiang Xie [2]

[1] Faculty of New Energy, University of Longdong, Qingyang, China
[2] School of Materials Engineering, Shanghai University of Engineering Science, Shanghai, China

*** Corresponding author:** Xingrong Zheng

Abstract: This study conducted an in-depth analysis of the momentum shift phenomenon during the men's singles matches at the 2023 Wimbledon Open, with the aim of quantifying momentum changes and their impact on match outcomes through a comprehensive data analysis framework that incorporates multivariate statistical methods such as BP neural networks, principal component analysis (PCA), entropy weighting analysis, and multivariate linear regression. The model, after sensitivity analysis and quantification through player rankings, is capable of predicting momentum shifts during a match and providing athletes with strategic adjustments. Additionally, the model's generalization ability was validated, demonstrating its potential applicability across various competitions, courts, and sports. The research outcomes not only offer coaches and athletes strategies for momentum shifts but also highlight the model's broad potential for application in sports match analysis.

Keywords: Power; Linear Regression Analysis; ANOVA Test; BP Neural Network.

1. Introduction

Momentum transfer in tennis is a complex and fascinating phenomenon that involves multiple dimensions, including mental, physical and technical. In high-level tennis duels, the skill gap between players is often tiny, so momentum management and conversion become one of the key strategies to win the match [1]. However, quantification and prediction of momentum remains a challenge in the field of sport science. Traditional analytical methods often rely on qualitative judgment and subjective experience, lacking in-depth understanding and accurate prediction of the changing patterns of momentum.

In this study, a prediction model based on BP neural network and multivariate linear regression is constructed to analyze and predict the momentum changes in a tennis match. And the model is used to predict the turning points and the dynamics of the winning percentage over time [2]. Finally, the model is applied to real matches, including pre-match prediction and post-match prediction [3].

2. Momentum Analysis

2.1. BP Neural Network-based Data Missing Value Processing

Table 1. Elapsed time after Python processing

	Match_id	Elapsed_time	Time_diff
0	2023-wimbledon-1301	00:00:00	0.0
1	2023-wimbledon-1301	00:00:38	38.0
2	2023-wimbledon-1301	00:01:01	23.0
3	2023-wimbledon-1301	00:01:31	30.0
4	2023-wimbledon-1301	00:02:21	50.0

First, we analyzed the data and calculated the time in seconds as shown in Table 1. For convenience, the score for AD lead status was replaced with 50 points. The dataset of the tournament contains 7284 game score data.

We use BP neural network to interpolate the missing values. For the nodes with missing values and outliers in the missing data, this is exactly the data we need to study the subsequent problems.

Missing value interpolation based on BP neural network is a promising method. First, we can construct a BP neural network model by taking the non-missing values in the dataset as the training set and the features corresponding to the missing values as the target values. Through continuous iterative training, the neural network can learn the complex relationships between features and thus accurately predict the missing values [4-5].

2.2. Establishment of Indicators and Assessments

The key factors related to the player's victory, in addition to being a server, are also related to the player's fatigue level, personal technical ability, and real-time mentality during the game. Based on the athletes' performance on the field, we list some important indicators as the key indicators to determine whether the athletes can win the game, as shown in Fig. 1.

2.3. Principal Component Analysis Models

We have three levels of stratified analysis: Goal level: predicted wins (for scoring); Criteria level: each metric used to predict real-time wins; Program level: two participants. We analyzed the players' mentality in these three types of metrics, with data derived from variables that affect the players' mentality [6].

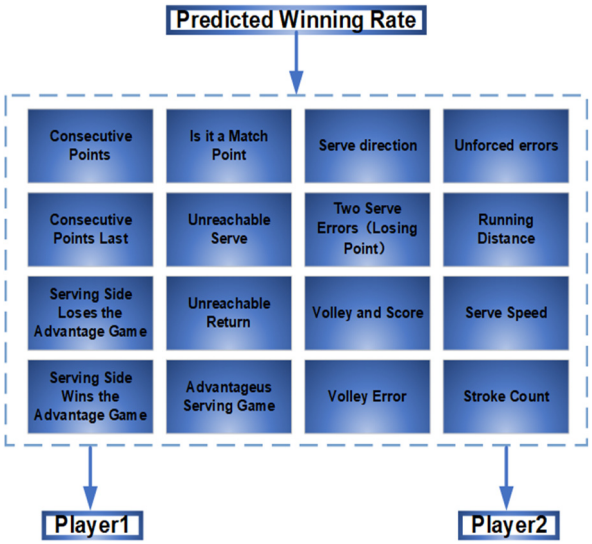

Fig 1. Indicator System Diagram

In order to achieve a more hierarchical classification, the dimensionality of the above three types of factors was reduced. The external relationship reconstruction model was utilized to construct the indicators as shown in the following equation.

$$\begin{aligned} x_1 &= a_{11}F_1 + a_{12}F_2 + \ldots + a_{1m}F_m + e_1 \\ x_2 &= a_{21}F_1 + a_{22}F_2 + \ldots + a_{2m}F_m + e_2 \\ x_3 &= a_{31}F_1 + a_{32}F_2 + \ldots + a_{3m}F_m + e_3 \end{aligned} \quad (1)$$

where x_1, x_2 , x_3 is an observable random variable denoting the continuity given in the question. f is a common factor, an unmeasured random vector such as the indicators affected by depletion under the conditions described above. and e are independent of each other, to simplify the model we can rewrite the factor score model as.

$$\begin{aligned} F_1 &= \omega_{11}x_1 + \omega_{12}x_2 + \ldots + \omega_{1p}x_p \\ F_2 &= \omega_{21}x_1 + \omega_{22}x_2 + \ldots + \omega_{2p}x_p \\ F_3 &= \omega_{31}x_1 + \omega_{32}x_2 + \ldots + \omega_{3p}x_p \end{aligned} \quad (2)$$

KMO and Bartlett's test were then performed to determine if a principal component analysis could be performed. The results obtained are shown in table 2, table 3 and table 4.

Table 2. KMO test and Bartlett's test of mentality indicators

value of KMO		0.855
Bartlett Sphelicity test	Approximate Chi-squared value	13868.993
	df	66
	P	0.000***

Table 3. KMO test and Bartlett's test of technical indicators

value of KMO		0.829
Bartlett Sphelicity test	Approximate Chi-squared value	87106.585
	df	136
	P	0.000***

Table 4. KMO test and Bartlett's test of exhaustion indicators

value of KMO		0.895
Bartlett Sphelicity test	Approximate Chi-squared value	18340.277
	df	15
	P	0.000***

Note: ***, **and * represent the significance levels of 1%, 5% and 10%, respectively

After testing, we got the following conclusions for KMO value:0.8 is very suitable for component analysis. And p is less than 0.05, the original hypothesis is rejected, which means that principal component analysis can be used.

Then, by analyzing the negative coefficients of the principal components and the heat map correlation coefficient matrix, the importance degree of the hidden variables in each principal component is analyzed. The distribution pattern is shown in Fig. 2. We obtained coefficients of 0.99287 and 0.60321 respectively, which shows that the two players present a high level of performance between each other.

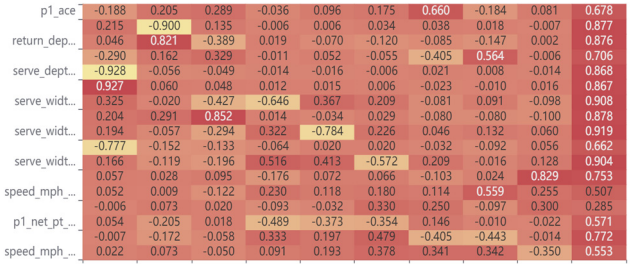

Fig 2. Correlation icon

3. Predictive Modeling

After establishing the indicator system and evaluation mechanism, we combine machine learning methods with traditional models for subsequent prediction. In classical statistics, the probability function that the aggregate depends on a certain parameter θ is denoted $p(x;\theta)$. He noted that different θ's in the parameter space θ correspond to different distributions. In Bayesian statistics, this is denoted as $p(x|\theta)$.

It represents the conditional probability function of the aggregate when the random variable θ takes a given value. The prior distribution $\pi(\theta)$ can be determined from the prior information of the parameter θ. From a Bayesian point of view, the generation of sample $X = (x_1 \cdots x_n)$ consists of two steps: Imagine generating individual θ_0 from the prior distribution $\pi(\theta)$; and generating another set of samples from $p(x|\theta_0)$, which has a joint conditional probability function of.

$$p(X|\theta_0) = p(x_1, \cdots x_n|\theta_0) = \prod_{i=1}^{n} p(x_i|\theta_0) \quad (3)$$

This distribution integrates aggregate and sample information.

θ_0 is the unknown, generated by the prior distribution $\pi(\theta)$.

In order to integrate the a priori information, we have to consider not only θ_0, but also the possibility of other values of θ. Therefore, we should integrate $\pi(\theta)$. The joint distribution of sample X and parameter θ is

$$h(X,\theta)=p(X|\theta)\pi(\theta) \tag{4}$$

This joint distribution combines aggregate, sample, and prior information, as well as all three available information.

In the absence of sample information, inferences can only be made from the prior distribution.

With the sample observation $X=(x_1\cdots x_n)$, θ should be inferred from $h(X,\theta)$. If $h(X,\theta)$ decomposes as follows:

$$h(X,\theta)=\pi(\theta|X)m(X) \tag{5}$$

where $m(X)=\int_\Theta h(X,\theta)d\theta=\int_\Theta p(X|\theta)\pi(\theta)d\theta$ is a marginal probability function of X, is independent of θ, and does not contain any information about θ.

Therefore, only the conditional distribution $\pi(\theta|X)$ can be used to infer B, as follows:

$$\pi(\theta|X)=\frac{h(X,\theta)}{m(X)}=\frac{p(X|\theta)\pi(\theta)}{\int_\Theta p(X|\theta)\pi(\theta)d\theta} \tag{6}$$

This conditional distribution is called the posterior distribution of θ, and it concentrates all the information about θ from the population, the sample, and the prior. The formula for the posterior distribution $\pi(\theta|X)$ is the Bayesian formula expressed as a density function and the result of the adjustment of the prior distribution $\pi(\theta|X)$ with the population and sample, which is closer to the reality of θ than the prior distribution $\pi(\theta)$. The posterior distribution $\pi(\theta|X)$ is the Bayesian formula with the density function expressed as a density function and the result of the adjustment of the prior distribution $\pi(\theta|X)$ with the population and sample. All inferences in Bayesian statistics are made on the basis of the posterior distribution. We most often apply the mean of the posterior distribution as a point estimate of θ, called the posterior expectation estimate, denoted $\hat{\theta}_B$. The mean of the posterior distribution is the mean of the posterior distribution.

Based on the above conditions, we visualize the performance of the two athletes as shown in Fig. 3.

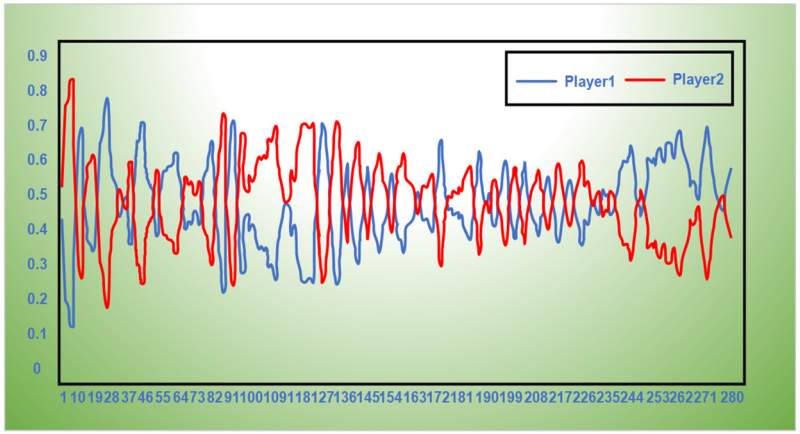

Fig 3. 2023-winbledon-1301 Performance of the competition athletes

We combine the above prediction model with multiple linear regression, a commonly used statistical analysis model for studying the effects of multiple independent variables on the dependent variable. The general form of the multiple linear regression model is:

$$y_0=a+a_1x_1+a_2x_2+a_3x_3+\ldots+a_nx_n \tag{7}$$

where n is the number of explanatory variables and a_n *(i = 1,2,..., n)* is the regression coefficient of the corresponding variable.

The multiple linear regression model holds true in most real-world problems where there is not one but multiple factors affecting the dependent variable, which fits the context of the problem. Therefore, this paper uses multiple linear regression analysis to further process the data.

In this paper, we will evaluate the athlete's “motivation” from the three aspects of personal technical ability (y_s), fatigue (y_e) and real-time mentality (y_m), which are summarized into three parts based on the data of the above three aspects, and each part establishes a regular multiple linear regression model based on the data affecting this part, and three multiple linear regression models are obtained.

$$y_s=a_0+a_1x_{sn}+a_2x_{sm}+a_3x_{swb}+a_4x_{swbc}+a_5x_{swbw}+a_6x_{swc}+a_7x_{sww}+a_8x_{sdc}+a_9x_{sdn}+a_{10}x_{rdd}+a_{11}x_{rdn}+a_{12}x_{ace}+a_{13}x_w+a_{14}x_{npw}+a_{15} \tag{8}$$

where y_s is the player's individual ability, a_i *(i = 0,1,2,..., 15)* is the regression coefficient of the corresponding variable, x_{sn} is the score of the first serve, x_{sm} is the speed of the serve, x_{swb} x_{swbc} , x_{swbw} , x_{swc} , x_{sww} is the Body, Body/Center, Body/Side, Center, Wide in

the direction of the serve, x_{sdc} is the depth of the serve near the line, x_{sdn} is the depth of the serve not near the line, x_{rdd} is the depth of the return depth, x_{rdn} is the depth of the return depth shallowness, x_{ace} is the ACE ball, x_w is the player hitting an untouchable scoring shot, and x_{npw} is a player scoring at the net.

$$y_e = b_0 + b_1 x_t + b_2 x_{sn} + b_3 x_{gn} + b_4 x_{dr} + b_5 x_{rc} + b_6 \quad (9)$$

where y_e is the athlete's fatigue level, b_i $(i = 0,1,2,..., 6)$ is the regression coefficient of the corresponding variable, x_t is the time of the game, x_{sn} is the number of sets, x_{gn} is the number of sets in the set, x_{dr} is the distance run by the athlete during the game, and x_{rc} is the number of shots taken in the section.

$$y_m = c_0 + c_1 x_t + c_2 x_s + c_3 x_g + c_4 x_{sc} + c_5 x_{df} + c_6 x_{ue} + c_7 x_{np} + c_8 \quad (10)$$

where y_m is the player's mentality, c_i $(i = 0,1,2,..., 8)$ is the regression coefficient of the corresponding variable, x_t is the time of the match, x_s is the hand, x_g is the player winning the match in the current set, x_{sc} is the player winning the match in the current set, x_{df} is the player serving twice and scoring a point, x_{ue} is the player's unforced error, and x_{np} is the player at the net.

After evaluating the athletes' technical ability, fatigue level and real-time mentality, these three variables were taken as independent variables affecting the "momentum", and a multiple linear regression model was established to evaluate the "momentum" of the athletes (y_i).

$$y_i = d_0 + d_1 y_s + d_2 y_e + d_3 y_m + d_4 \quad (11)$$

where d_i $(i = 0,1,2,3,4)$ is the regression coefficient of the corresponding variable.

4. RESULTS

4.1. Visualization of Predictive Model Results and Turning Point Prediction

We verified the accuracy and robustness of the model using cross-validation and Holdout techniques. Cross-validation (CV) is a widely used technique for model evaluation [6].

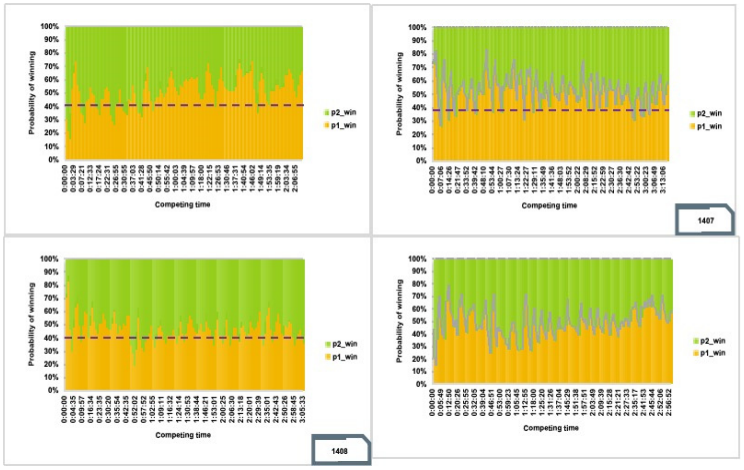

Fig 4. Visual analysis of winning score based on prediction model

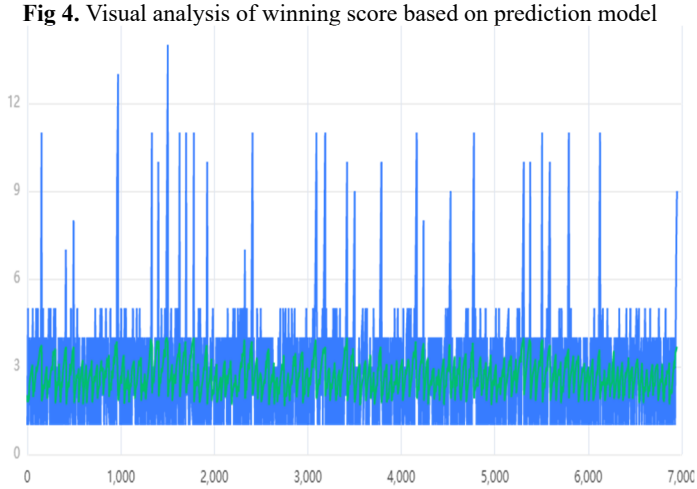

Fig 5. Validation of model results by influence indicators

The above prediction model allows us to know at which point in time or stage of the game the trend of the winning percentage changes significantly. Based on the calculation results of the above model, we randomly selected the players of four matches, and visualized their scores and winning percentages, as shown in Fig. 4.

Then, the model analyzes the historical data of past players and retrains the model to simulate the battle process. As shown in Fig. 5. With the visualization above, it is not difficult to arrive at a specific reason for the momentum error conclusion: In past matches, the player has built up an advantage, such as winning one more game or even more games, which continues to build up an advantage for the player. We get this judgment contrary to the fact. So, we need to redesign the cumulative momentum between players and predict the momentum based on that.

Here's how to design momentum. We have the following equation.

$$Momentum = \frac{Time}{100} + ace + 0.5 \times winner - 0.5 \times error + win \times \beta \tag{12}$$

Among other things, we improved the win factor by adding the effect of multiple past balls on the current momentum, not just the effect of the last ball on the current momentum.

$$\beta = 1 + 0.2 \times win \times \left(1 - 0.5^{n}\right) \tag{13}$$

where n is the number of wins in the last n matches. The exponential correlation of power affecting wins is obtained, as shown in Table 5.

Table 5. Correlation degree of predicted winning rate through improved momentum

Evaluate item	Correlation degree	ranking
p1_double_fault	0.977	1
p2_unf_err	0.966	2
p1_net_pt	0.963	3
p1_unf_err	0.962	4
p2_net_pt	0.962	5
p1_sets	0.958	6
p2_sets	0.958	7
p1_games	0.957	8
p2_games	0.957	9

We have weakened the effect of existing advantages on momentum and further strengthened the effect of the winning mechanism on momentum in the short term.

4.2. Extension and Feasibility Analysis of Momentum Conversion Models

We used our model to test it on other matches. To improve readability, we inverted the results of the same matches and our predicted momentum into the same table. The results are shown in Fig. 6. The blue dashed line represents our prediction of the momentum of the game, where 1 denotes momentum against player_1, -1 denotes momentum against player_2, and 0 denotes that there is no significant inversion of momentum.

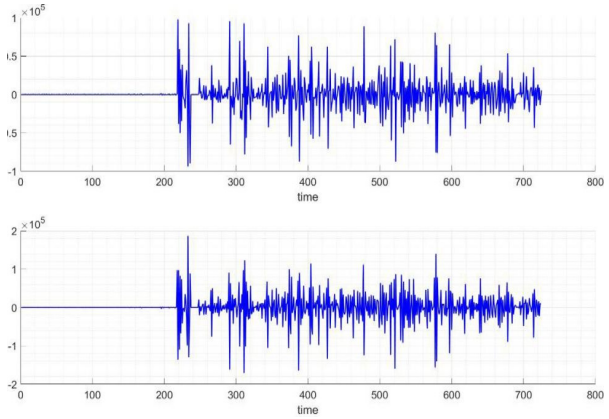

Fig 6. Momentum prediction results for different iterations

4.3. Analysis of the Results

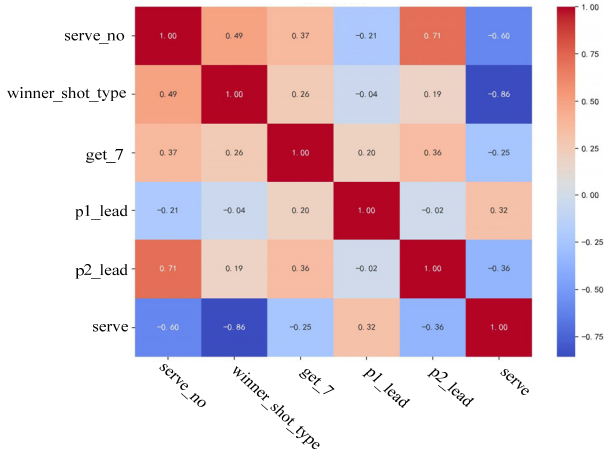

Fig 7. Correlation Matrix between Main Indicators

The situation related to unexpected performance in the game was analyzed, and the results of the game were compared with the actual results based on the momentum changes in the model. A series of accurate metrics for interpreting turning points, breakpoints and consecutive scores were obtained to obtain the relationship between player 1 and player 2 being able to win under the influence of the metrics, as shown in Figure 7.

5. Conclusion

In this research work, a dynamic regression model was constructed to predict the change of momentum during the match, highly accurate prediction results were obtained, the ability of the model to predict the change of momentum in tennis matches was comprehensively tested and its validity and accuracy was verified by the case study of an open tournament, and the effective analysis helped us to identify the other factors that need to be taken into account.

Despite the remarkable results of this study, there are some limitations. For example, the hierarchical analysis method may become cumbersome when dealing with extremely complex decision problems and requires high data quality. Future research could further explore these issues and attempt to develop more efficient and robust models.

References

[1] Scientific Platform Serving for Statistics Professional 2021. SPSSPRO. (Version 1.0.11) [Online Application Software]. Retrieved from https://www.spsspro.com.

[2] Draper, N.R. and Smith, H. Applied Regression Analysis. Wiley Series in Probability and Statistics. 1998.

[3] Benedikt Langenberg, Markus Janczyk, Valentin Koob, Reinhold Kliegl, Axel Mayer. A tutorial on using the paired t

test for power calculations in repeated measures ANOVA with interactions. 2022.

[4] Richard G. Brereton. Introduction to analysis of variance.2018.

[5] M S Mayo,M D Conerly. Evaluating overall significance levels in multifactor ANOVA. 2018.

[6] Assaf Rabinowicz, Saharon Rosset.Cross-Validation for Correlated Data. 2020.