# Research on the Prediction of Tennis Momentum Transition Based on Multiple Linear Weighting and Logistic Regression

**Lingxin Kong[1, *, #], Chuan Ju[1, #], Diming Wu[1, #]**

[1]School of Computer Science, China University of Geosciences (Wuhan), Wuhan, China

*Corresponding author: 3180348092@qq.com

[#] These authors contributed equally to this work

**Keywords:** Tennis Match Analysis; Momentum Transition Prediction; Multiple Linear Weighting; Logistic Regression; Principal Component Analysis; Analytic Hierarchy Process

**Abstract:** This paper investigates the prediction of tennis match momentum transitions based on multiple linear weighting and logistic regression. To analyze the match situation more comprehensively, we considered the scenarios of consecutive scoring and losing points and introduced 11 new variables to measure match momentum. Through Principal Component Analysis (PCA), a dimensionality reduction technique, we consolidated 14 variables into 9 key components to simplify the model and enhance its generalizability and conciseness. Furthermore, we employed the Analytic Hierarchy Process (AHP) to assign weights to independent variables, which were then used for stepwise regression analysis to select influencing factors. By applying multiple linear weighting, we constructed a performance scoring model and predicted match turning points using a logistic regression model. We defined "swings in the match" as changes in the state of a particular player and built a model to quantify the impact of various indicators on match swings and predict the specific moments when the win rate transitions from one state to another. Our experimental results demonstrate that the proposed model has high accuracy in real-time win rate prediction and momentum transition prediction. We conclude that the model can provide effective predictions for tennis match momentum transitions, offering valuable insights for coaches and athletes.

## 1. Introduction

Tennis matches are often won or lost by momentum shifts at crucial moments, which can be caused by consecutive points or lost points. Momentum transition not only affects the immediate outcome of the game, but also has a profound impact on the players' mental state and game strategy [1, 2]. However, most of the existing research focuses on the static analysis of player performance, and relatively few research on the prediction of dynamic process such as momentum transition. In order to fill this gap, this paper proposes a momentum transformation prediction model for tennis matches based on multiple linear weighting and logistic regression [3, 4].

This study first collected and preprocessed the Wimbledon official match data, including filling in missing values and converting qualitative variables to binary variables. In order to analyze the match situation more comprehensively, we introduced 11 new variables to measure the match momentum and reduced the dimensionality of the data using principal component analysis (PCA) techniques, thus simplifying the model and improving its universality and simplicity [5].

Further, this paper uses the analytic Hierarchy process (AHP) to assign weights to independent variables, and uses these weights for stepwise regression analysis to select influencing factors [6]. Using multiple linear weighting, we construct a performance scoring model and predict the turning point of the match through logistic regression model. We define "volatility in a game" as a change in a particular player's state, and construct a model to quantify the impact of various metrics on match volatility and predict the specific moment at which a win percentage switches from one state to another. The purpose of this study is to provide a more accurate prediction of momentum conversion in tennis matches by constructing a prediction model that takes into account many factors.

Copyright © (2024) Francis Academic Press, UK  DOI: 10.25236/iwmecs.2024.010

## 2. Data Analysis

### 2.1 Data Preprocessing

This paper has collected official match data provided by Wimbledon as the dataset. Initially, we need to conduct data preprocessing.

#### 2.1.1 Fill Missing Values

Upon scrutinizing the dataset, we detected missing values in columns such as spee_mph. Consequently, it becomes imperative to address these gaps by imputation. We have chosen mean imputation, as the addition of data reflecting the average performance per game is not anticipated to exert a substantial influence on predictions of player performance.

#### 2.1.2 Data Addition and Processing

The variables "serve depth" and "return depth" were initially encoded as "CTL: Close To Line, NCTL: Not Close To Line" and "D: Deep, ND: Not Deep," respectively. For the sake of convenience in subsequent analysis, we will transform them into binary variables (0-1). We have also added 11 variables as shown in Table 1. To conduct a more comprehensive analysis, we will consider situations involving consecutive wins and losses.

Table 1. Competition variables and their interpretation

| Variables | Explanation |
|---|---|
| get_7 | This set, is it a tiebreak? |
| m_point_1 | Is Player 1 in possession of a set point in this set? |
| m_point_2 | Is Player 2 in possession of a set point in this set? |
| p1_le50 | Is Player 1 currently leading? |
| p2_le50 | Is Player 2 currently leading? |
| p1_mb | Is Player 1 scoring consecutively? |
| p2_mb | Is Player 2 scoring consecutively? |
| p1_ss | Is Player 1 losing scores consecutively? |
| p2_ss | Is Player 2 losing scores consecutively? |
| p1_rm | Does Player 1 move more? |
| p2_rm | Does Player 2 move more? |

When calculating consecutive points, we utilized the following formula:

$$x[i] < x[i+1] < x[i+2]. \tag{1}$$

In the context of p1_ss and p2_ss, a consecutive loss of 3 points is treated as a continuous decline in momentum. Subsequently, each detection of a 1 result in a reduction of momentum by one unit. Moreover, a greater running distance may potentially affect the upcoming game, leading to a predicted decrease in the winning rate.

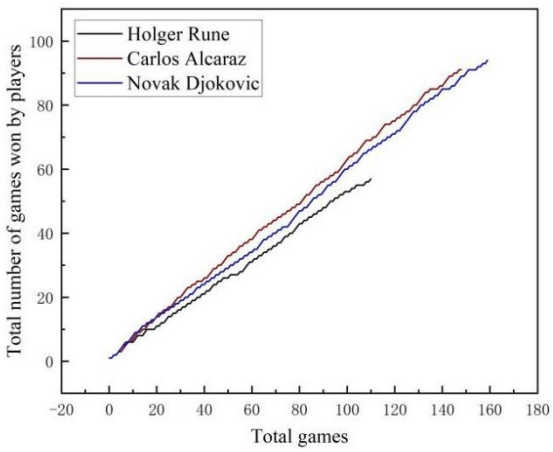


Fig. 1 Line chart of total match victories for three players

We have selected three players, each from the final and match with the identifier 1501 (with one being repeated). In subsequent analyses, we will utilize data from above matches. Therefore, we use these three players as examples for illustration. The total match win rates for the three players are depicted in Fig. 1.

### 2.2 Data Selection

To consider it more perfectly, we strive to encompass a broad range of elements such as consecutive wins. While this approach enhances the comprehensiveness of our model, it also runs the risk of introducing excessive indicators in the primary content.

We employ PCA algorithm to reduce dimensionality and construct comprehensive indicators by grouping similar metrics into distinct categories. This not only alleviates the complexity of establishing multiple regression equations in the future but also enhances the generalizability and conciseness of our overall model. Feature extraction is shown in Table 2. The 14 variables are defined by us as "momentum".

Table 2. Feature extraction

| **Specific Indicators** | |
|---|---|
| Consecutive Points | Winning Match Point |
| Consecutive Points Lost | Players Lead |
| Lost Opportunities to Win what he other is Serving | Made it to the Net |
| Won the Game The Other is Serving | Won While at the Net |
| Missed both Serves and Lost the Point | Made Unforced Error |
| Untouchable Winning Serve | Running Distance |
| Untouchable Winning Shot | Serving Player |

### 2.3 Principal Component Analysis (PCA)

Large datasets are increasingly widespread in many disciplines. In order to interpret such datasets, methods are required to drastically reduce their dimensionality in an interpretable way, such that most of the information in the data is preserved. Principal component analysis (PCA) is one of the most widely used. Its idea is simple—reduce the dimensionality of a dataset, while preserving as much 'variability' (i.e. statistical information) as possible.

Firstly, standardize the raw data to mitigate the impact of scales on the extraction of principal components. Then, calculate the covariance matrix for the standardized dataset.

$$\sum = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}) (X_i - \bar{X})^T. \tag{2}$$

Where, $n$ denotes the sample size, $X_i$ represents the ith sample, and $\bar{X}$ is the sample mean. Perform eigenvalue decomposition on the covariance matrix to obtain the eigenvalues and their corresponding eigenvectors.

$$\sum v = \lambda v. \tag{3}$$

Where, $v$ denotes the feature vector, $\lambda$ represents the eigenvalues. Order the eigenvalues in descending order and select the corresponding eigenvectors for the top k eigenvalues as principal components, where k represents the desired reduced dimension. Finally, the original data is projected to create dimensionally reduced data.

$$Y = XV_k. \tag{4}$$

$Y$ denates the data after dimensionality reduction, $X$ represents the raw data, and $V_k$ contains the first k feature vectors.

We employ Principal Component Analysis (PCA) for dimensionality reduction and create a composite indicator by grouping together similar metrics.

The fundamental concept of Principal Component Analysis (PCA) involves utilizing dimensionality reduction techniques to transform multiple variables into a limited number of principal

components. These principal components retain the significant majority of information from the original variables.

$$y_1 = a_{11}x_1 + a_{21}x_2 + \cdots + a_{p1}x_p. \quad (5)$$

$$y_2 = a_{12}x_1 + a_{22}x_2 + \cdots + a_{p2}x_p. \quad (6)$$

$y_n$ denotes the simplified latitude, $x_n$ represents the original latitude data, $a_n$ denotes the coefficient. In the initial phase, perform the Kaiser-Meyer-Olkin (KMO) and Bartlett's tests to assess the feasibility of conducting Principal Component Analysis (PCA). The results are shown in Table 3

Table 3. KMO Test and Bartlett's Test

| Test | value | |
|---|---|---|
| KMO Value | 0.6 | |
| Bartlett's Test of Sphericity | Approximate chi-square | 22048.708 |
| | df | 91 |
| | P | 0.000*** |

By means of Principal Component Analysis, we have consolidated 14 data points into 9 key components to facilitate subsequent analysis, which is contained in Fig. 2.

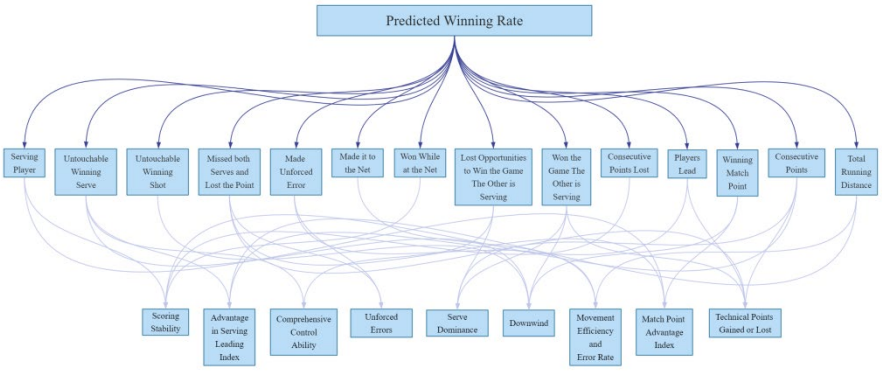


Fig. 2 PCA Result Plot

## 3. Establishment of Model

### 3.1 Analytic Hierarchy Process (AHP)

Multicriteria decision-making (MCDM) has also been used to exploit the search space after exploring the search space with nature-inspired optimization techniques. The analytic hierarchy process (AHP), one of the well-regarded MCDM tools, is attributed to Thomas Saaty. It has been widely used in many different fields for the last forty years. In AHP the factors, which can influence the decisions, are identified and then these factors are arranged into a hierarchal structure of different levels to reduce the complexity of the decision problem.

Before engaging in multiple linear regression analysis, the Analytic Hierarchy Process (AHP) can be employed to establish the weights for each independent variable. Subsequently, these weights facilitate stepwise regression analysis for factor selection. Ultimately, the weights serve as coefficients or constraints in the multiple linear regression model during the regression analysis.

The weights assigned are as follows: "Scoring Stability" contributes 14.193%, "Advantage in Serving Leading Index" holds a weight of 31.11%, "Comprehensive Control Ability" is weighted at 4.125%, "Unforced error" contributes 11.424%, "Serve Dominance" holds a weight of 4.092%, "Downwind" is weighted at 7.43%, "Movement Efficiency and Error Rate" contributes 6.698%, "Match Point Advantage Index" holds a weight of 12.589%, and "Technical Points Gained or Lost" has a weight of 8.338%.

The results of the Analytic Hierarchy Process (AHP) calculation show that the maximum eigenvalue is 10.127. Referring to the Random Index (RI) table, the corresponding RI value is found

to be 1.451. Therefore, the Consistency Ratio (CR), calculated as CI / RI, is 0.097, which is less than 0.1, indicating a satisfactory outcome in the one-time consistency check.

### 3.2 Multiple Linear Weighting

Above, we have determined the respective proportions for each item.

We designate the nine principal components as $x_1$ through $x_9$. Let the performance score be denoted as $Y$. Therefore, we can express it with the following formula:

$$Y = 0.142\,x_1 + 0.311\,x_2 + 0.041\,x_3 + 0.114\,x_4 + 0.041\,x_5 + 0.074\,x_6 + 0.067\,x_7 + 0.126\,x_8 + 0.083\,x_9. \quad (7)$$

Following this, we consider the 1501 match as an example, inputting the data for the two players to calculate their individual scores. Normalization is necessary in this context to convert the scores into decimals, thereby enhancing the prediction of win rates.

### 3.3 Logistic Regression

To predict turning points in a race we use a logistic regression model. We define "swings in the match" as a change in the state of a particular player. For instance, if the opponent has been leading from the start until a certain point, but suddenly the player starts exerting more force, improving their state—although the score may not have surpassed the opponent's— we still consider the match to have undergone a transition. To achieve this, we need to construct a model that not only quantifies the impact of various indicators on the "swings in the match" but also predicts the specific moments when the win rate transitions from one state to another.

In cases where the dependent variable has only two categories (0/1), it is referred to as binomial logistic regression. When dealing with multiple categories, it is termed multinomial logistic regression.

The general formula for linear regression is:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon. \quad (8)$$

The regression equation $E(y_i) = \beta_0 + \beta_1 x_i$ forecasts the average value of the dependent variable when the explanatory variable is $x_i$. In the context of a binary dependent variable (0/1), the logistic regression equation predicts the probability of the dependent variable being $p = 1$ given the explanatory variable is $x$. Thus, the general form of the logistic regression equation is:

$$y_{p=1} = \beta_0 + \beta_1 x. \quad (9)$$

## 4. Results

### 4.1 Real-time Win Rate Prediction

The eventual output is illustrated in Table 4 (displaying only a partial dataset).

Table 4. The prediction of win rates

| Player1 | Player2 | Elapsed_time | Player1 | Player2 |
|---|---|---|---|---|
| Carlos Alcaraz | Holger Rune | 0:00:00 | 0.582919 | 0.417081 |
| Carlos Alcaraz | Holger Rune | 0:00:54 | 0.602693 | 0.397307 |
| Carlos Alcaraz | Holger Rune | 0:01:25 | 0.697955 | 0.302045 |
| Carlos Alcaraz | Holger Rune | 0:02:11 | 0.69615 | 0.30385 |
| Carlos Alcaraz | Holger Rune | 0:02:55 | 0.74288 | 0.25712 |

Using the match identified as 1501 as an example, a visual representation, depicted in Fig. 3, can be generated. This visualization delineates the scoring dynamics of both players throughout the match, offering a graphical portrayal of the game's unfolding progression.

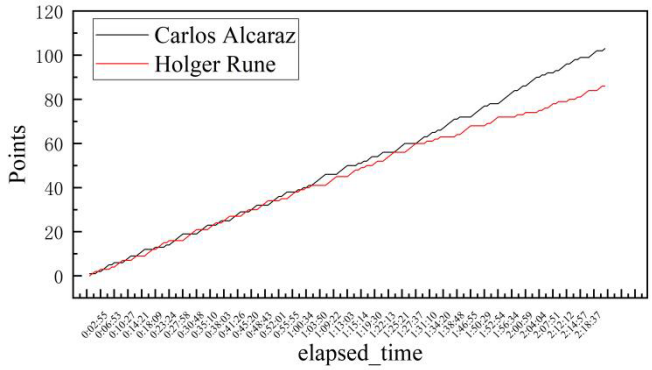


Fig. 3 Line chart depicting the progress of match 1501

Examining Fig. 4, it is evident that the deep blue player consistently demonstrates a superior overall performance, with predicted values hovering around 0.6. Conversely, the light blue player exhibits relatively weaker performance, consistently subdued both at the beginning and towards the conclusion of the match.

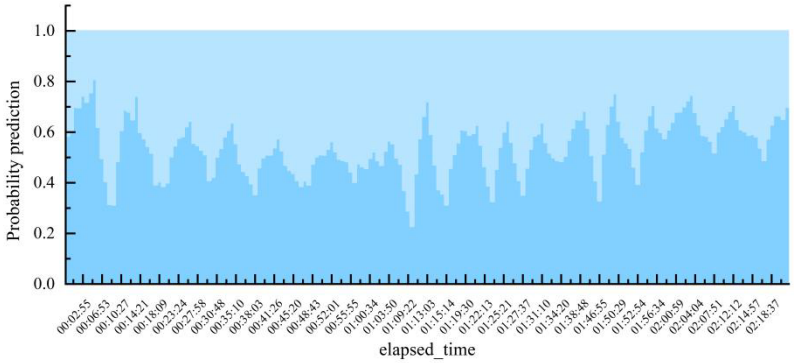


Fig. 4 Prediction of player performance during the game

In the initial analysis, Carlos Alcaraz emerged as the winner in this match, consistent with the advantageous trend indicated by the deep blue color in our predicted chart. Notably, as the match reached the 1/3 mark, we observed a minor decline in Carlos Alcaraz's performance, mirrored by a prompt decrease in the corresponding deep blue win rate on the chart. These findings underscore the strong predictive capabilities of our model.

### 4.2 Momentum Transition Prediction

Through binary logistic regression analysis, the momentum transition prediction results obtained are shown in Table 5 and Table 6.

Table 5. Result between "swings in play" and "momentum"

| Experimental Group | Regression Coefficients | Standard Error | Wald | P | OR | upper limit | lower limit |
|---|---|---|---|---|---|---|---|
| Constant | -7.237 | 1.192 | 36.857 | 0.000*** | 0.001 | 0 | 0.007 |
| player1.1 | 12.113 | 2.076 | 34.041 | 0.000*** | 182282.842 | 3115.144 | 10666290.133 |

Table 6. Result between "runs of success" and "momentum"

| Experimental Group | Regression Coefficients | Standard Error | Wald | P | OR | upper limit | lower limit |
|---|---|---|---|---|---|---|---|
| Constant | -6.186 | 1.124 | 30.293 | 0.000*** | 0.002 | 0 | 0.019 |
| player1.1 | 9.664 | 1.93 | 25.062 | 0.000*** | 15741.014 | 357.971 | 692177.484 |

Subsequently, we utilized five evaluation metrics to assess the accuracy of the model's predictions: Accuracy, Recall, Precision, F1, and AUC, with the results presented in Table 7. Confusion matrix heat map is shown in Fig. 5. The most effective methods for evaluating classification models are

typically the confusion matrix and the ROC curve (AUC). When the AUC falls between 0.7 and 0.85, the performance is considered good.

Table 7. The evaluation results of the prediction model

| | Accuracy | Recall | Precision | F1 | AUC |
|---|---|---|---|---|---|
| swings in play | 0.725 | 0.725 | 0.72 | 0.72 | 0.791 |
| runs of success | 0.693 | 0.693 | 0.667 | 0.668 | 0.75 |

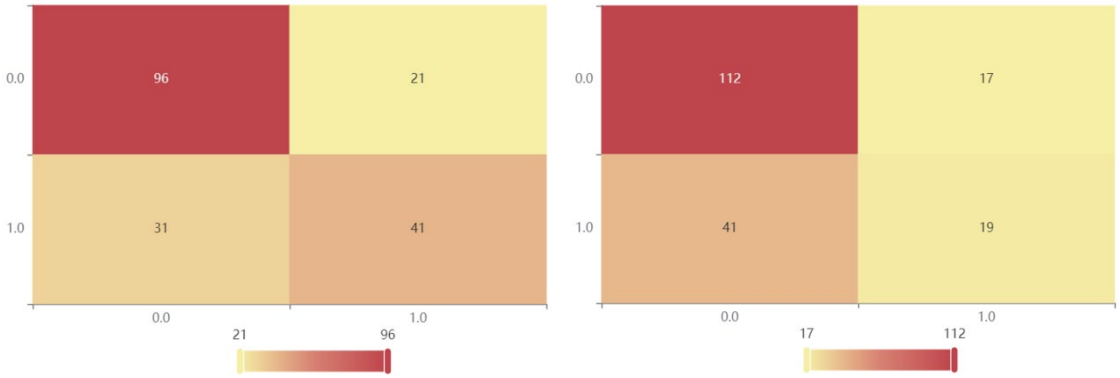


Fig. 5 Confusion matrix heat map of p1_le50 (Left) and p1_mb (Right)

## 5. Conclusion

This study successfully developed a tennis match momentum transition prediction model based on multiple linear weighting and logistic regression. Through an in-depth analysis of official Wimbledon match data, we introduced 11 new variables to measure match momentum and effectively reduced data dimensions using Principal Component Analysis (PCA), while retaining the core information of the data. Furthermore, the Analytic Hierarchy Process (AHP) was utilized to assign weights to influencing factors, thereby enhancing the predictive accuracy of the model.

The experimental results demonstrate that the proposed model performs exceptionally well in real-time win rate prediction and momentum transition prediction, with high predictive performance indicated by evaluation metrics such as accuracy, recall, precision, F1 score, and AUC. Notably, the AUC values fall between 0.7 and 0.85, confirming the model's robust predictive capabilities. Future work will further explore the applicability and robustness of the model in different match scenarios, as well as the integration of the model into actual match analysis and decision support systems.

## References

[1] Dietl, Helmut, and Cornel Nesseler. "Momentum in tennis: Controlling the match." UZH Business Working Paper Series 365 (2017).

[2] O'Donoghue, Peter. "Momentum in tennis matches in Grand Slam tournaments." Performance Analysis of Sport IX. Routledge, 2013. 174-179.

[3] Curtis, David. "Multiple linear regression allows weighted burden analysis of rare coding variants in an ethnically heterogeneous population." Human Heredity 85.1 (2021): 1-10.

[4] LaValley, Michael P. "Logistic regression." Circulation 117.18 (2008): 2395-2399.

[5] Abdi, Hervé, and Lynne J. Williams. "Principal component analysis." Wiley interdisciplinary reviews: computational statistics 2.4 (2010): 433-459.

[6] Podvezko, Valentinas. "Application of AHP technique." Journal of Business Economics and management 2 (2009): 181-189.