Momentum Quantification and Prediction of Tennis Match Based on Time Series and Logistic Regression

Yichen Han*

Jinan University-University of Birmingham Joint Institute, Jinan University, Guangzhou, China, 511443

*Corresponding author: yxc247@student.bham.ac.uk

Abstract. In the highly competitive sport of tennis, it is crucial to understand and capitalize on every small match advantage, and despite the attention paid to the concept of momentum, its quantification, analysis, and prediction remain a challenge. The purpose of this study was to examine the role of momentum quantification, momentum influencing factors, and accurate prediction of momentum shifts in enhancing athletes' performance in tennis. By analyzing the competitive data of tennis players, a tennis-based "Big Data 'Momentum' Shift Prediction Model" was constructed, which not only specifically quantifies momentum, but also captures the impact of potential information on the changing situation of the game, and predicts the momentum of a match with an accuracy of around 80% accuracy in predicting how the situation will change during the match. Specifically, the construction of high-quality models helps athletes and coaches make timely adjustments to their strategies during the game. These findings are important guidance for coaches, athletes, and researchers in related fields to advance sports science and improve athletes' performance.

Keywords: Quantify, Time Series Model, Correlation Analysis, Prediction Model, Logistic Regression.

1. Introduction

In highly competitive sports, especially in a sport like tennis which is highly adversarial and highly dependent on individual performance, it becomes crucial to understand and capitalize on every little advantage in the game. In addition to the strength of the players themselves, momentum is one of the key factors influencing the variability of the outcome[1-4]. Although momentum is a widely discussed concept, its impact and methods of quantification as well as prediction of momentum in sports, especially in tennis, have been difficult to study. Momentum is often considered an indicator that can significantly influence the performance of the athlete and the outcome of the match[3-8]. However, although the existence of momentum is generally recognized, its abstract nature makes scientific quantification and practical application a challenge.

In recent years sports competitions have developed rapidly and athletes have become increasingly competitive before. The exploration of the laws of the game has received more and more attention. Among them, the quantification of momentum, the influencing factors of momentum, and the prediction of momentum shifts have always been of interest[1, 2, 4-8]. Previous studies on momentum usually use time series models, correlation analysis, and logistic regression algorithms, but they are usually limited to one area. For example, Zhang et al. quantified and predicted the average speed of marathon runners in 10,000-meter segments based on a time series autoregressive (AR) model[2]. Shen et al. used correlation analysis to calculate PEARSON correlations between athletes' physical exertion metrics and total match scores in women's singles matches at tennis grand slams to determine which metrics had a more significant relationship with total match scores[7]. Jiang et al. applied a logistic regression model to predict the athletes' match results based on the match data of the four Grand Slam tournaments, and the men's singles professional tennis tournaments from 2014-2018[5]. Few previous studies have explored the whole in depth. Since actual matches can be affected by a variety of factors, it is difficult to effectively interpret the results with only a single analysis. Therefore, this paper needs to synthesize multiple aspects to analyze the momentum.

The data in this paper comes from https://www.contest.comap.com. In this paper, data preprocessing was first performed on the raw data to obtain data for data analysis; after that, a quantitative model of momentum is constructed using the time series model, which lays a good foundation for the construction of subsequent models; next, correlation analysis is used to analyze the factors that have a significant impact on momentum; finally, using logistic regression algorithms, an attempt is made to construct a momentum shift prediction model. At the end of the article, the results of the model were synthesized with the aim of providing theoretical support for athletes and coaches to make decisions during the game and to make different recommendations for different game situations.

2. Preparation and Development of the Momentum Assessment and Forecasting Model

2.1. Data Preprocessing

First, the existing data is analyzed and organized, and the data is converted into a format acceptable to the model. For example, character-based data is converted to a numeric format. After completing the data conversion, data cleaning is required to deal with outliers, duplicate values, and missing values using appropriate methods. Preprocessing the data, including label coding, removing duplicate values and outliers, filling in missing values, and processing the data columns, ultimately gives us data for quantifying momentum as well as predicting momentum shifts.

2.2. Constructing a Quantitative Momentum Scoring Model

By considering the impact of different factors on momentum, one momentum scoring system was constructed to quantify "momentum" and a time series model was developed. The momentum scoring system consists of positive and negative impacts[1, 2]. For example, serving rights, winning streaks, and high-level scoring shots will cause a positive increase in a player's momentum, while losses, forced errors, and two service errors will cause a decrease in a player's momentum.

The primary influences on a player's momentum are “winner of the point” (point_victor), “server of the point” (server), and “sequential victories” (based on the data), and the secondary influences are “player 1 hit an untouchable winning serve” (p1_ace), “player 2 hit an untouchable winning shot” (p2_winner), “player 1 double serves faults” (p1_double_fault), and “player 1 unforced errors” (p1_unf_err).

Reintroducing a formula for momentum; Player 1 The change in momentum is a continuous process that needs to take into account the momentum of the previous moment. The momentum at time point t is denoted as w_t, plus the product of the momentum change (winner) p_t ($point\_victor = 1$ then $p_t = 1$ else $p_t = -1.1$), the player's serve at time t, S_t ($server = 1$ then $S_t = 1.2$, else $S_t = 1$), and the player's winning streak at time t, C_t (the first win 1, each winning streak 0.2); and then add the winning serves A_t(0.02), the game-winning goal W_t (0.01), the double serves faults D_t (-0.2), and the unforced errors E_t (-0.1) respectively.
The calculation formula is:

$$w_t = w_{t-1} + (p_t * S_t * C_t) + A_t + W_t + D_t + E_t \quad (1)$$

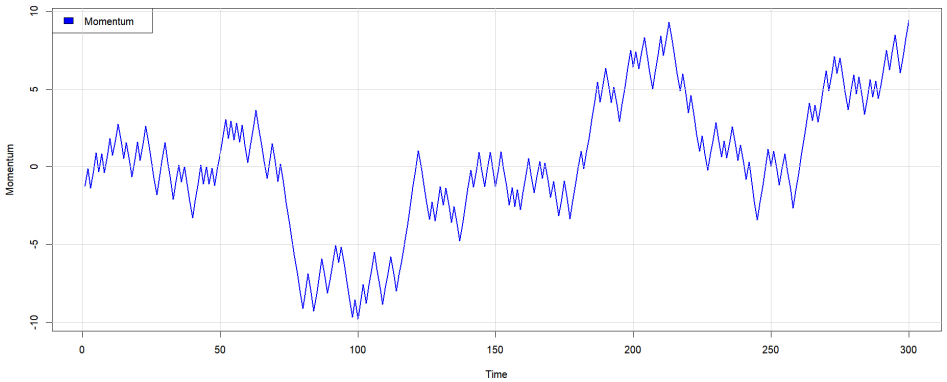

Figure 1: Momentum over Time

Quantification of momentum is illustrated in the Figure 1. Thereafter, a model has been developed, "Momentum Shift Prediction Model", which is used to predict when the situation in the match will change and on which side the advantage lies, i.e., calculate is needed and then based on the momentum change of both players, if it has been increasing positively, it means that the advantage of the match lies with player 1 during this period, and vice versa, it means that it lies with player 2.

With this data as the target data, a prediction model is constructed to predict which side has the advantage at a certain point in time, thus knowing how the situation changes during the match.

3. Exploring Potential Indicators for Tennis Match

To explore which variables have a significant effect on momentum, this paper performed a correlation analysis on the data[6-8]. The variables 'server', 'serve_no', 'p1_points_won', 'p1_winner', 'p1_unf_err', 'p1_net_pt', 'p1_distance_run', 'rally_ count', 'speed_mph', 'serve_width', 'serve_depth', 'return_depth' were chosen for correlation analysis.

3.1. Judgment of Normality

First determine whether the data follow a normal distribution by utilizing the Anderson-Darling test[9]. The Anderson-Darling test evaluates the fit of sample data to a theoretical distribution, often for testing normality, by measuring discrepancies between the observed and theoretical cumulative distribution functions through the Anderson-Darling statistic (AD statistic). It starts with a hypothesis that assumes a specific distribution, such as normal, and calculates the AD statistic to assess the fit. The test's outcome, based on the AD statistic's significance, determines if the sample aligns with the hypothesized distribution. However, it's sensitive to large sample sizes, requiring careful interpretation.

The A-D test results are shown in Table 1. Using R to perform an A-D test on the data and the final results yield the selected variables 'server', 'serve_no', 'p1_points_won', 'p1_winner', 'p1_unf_err', 'p1_net_pt', 'p1_ distance_run', 'rally_count', 'speed_mph', 'serve_width', 'serve_depth', and 'return_depth' all have p-values much less than 0.05; therefore, at 5% significance, the null hypothesis that all of these variables are normally distributed is rejected.

Table 1: AD Statistic and P-value of Selected Variables

Symbols	AD Statistic	P-value
server	1308.9	< 2.2e-16
serve_no	1444.6	< 2.2e-16
p1_points_won	69.536	< 2.2e-16
p1_winner	2062.3	< 2.2e-16
p1_unf_err	2273.5	< 2.2e-16
p1_net_pt'	2378.3	< 2.2e-16
p1_distance_run	233.36	< 2.2e-16
rally_count	341.55	< 2.2e-16
speed_mph	79.95	< 2.2e-16
serve_width	290.6	< 2.2e-16
serve_depth	1434.2	< 2.2e-16
return_depth	633.68	< 2.2e-16
y	1308.9	< 2.2e-16

3.2. Correlation Analysis

The use of the Pearson correlation coefficient may not be appropriate when the data does not conform to a normal distribution or shows a skewed distribution. It can be shown that the data is not normally distributed, in which case the Spearman correlation coefficient is a better choice because it is more lenient on the data distribution[10].

This text analyze player 1. First compute the difference of the 'momentum' column, then compute the value of the 'momentum_shift' column based on the value of the 'momentum_diff' column, merge X and y, compute the Spearman correlation coefficient of the features and the 'momentum_diff' column, and finally generate a heat map of the correlation coefficients and show it as follows:

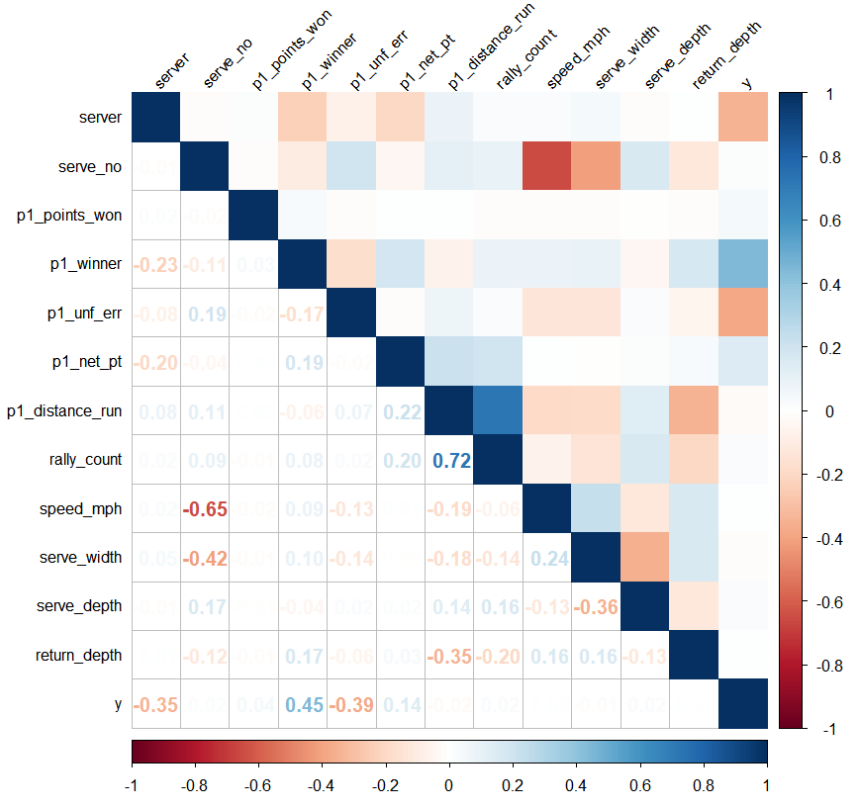

Figure 2: Correlation Coefficient Heat Map

As can be seen from Figure 2, 'server', 'p1_winner', 'p1_unf_err', 'p1_ net_pt' have a strong influence on the model; that is, factors such as the server, players hitting unreturnable scoring balls,

and player errors, and players hitting the ball toward the net have a strong influence on the momentum as well as the prediction model.

The initial analysis indicates that scoring, consecutive scoring, and unreturnable shots play a significant role in enhancing momentum. indicated by their high positive correlation coefficients. Balls hit to the net also positively affect momentum, though the impact is minor (correlation coefficient of 0.14). Serving has a notable negative influence on momentum, with a correlation coefficient of -0.35, implying that the opponent's serve increases their winning probability and negatively impacts the serving side's momentum. Player errors are particularly detrimental to momentum, evidenced by a substantial negative correlation coefficient of -0.39. Further parametric analyses and p-value assessments are recommended to identify the independent variables with the most significant effects on the model.

3.3. Potential Problems

With the results from figure 2, it is found that "p1_distance_run" and "rally_count", "serve_no" and "speed_mph", "serve_no" and "serve_width", "p1_distance_run" and "p1_count". distance_run" and "return_depth" have the darker colors, which means that the absolute value of the correlation coefficients between the variables is high; therefore, there may be multicollinearity between the independent variables in the regression, leading to bias in the subsequent model predictions. Therefore, there may be multicollinearity between the independent variables of the regression, which leads to the bias of the subsequent model[11]. To eliminate multicollinearity, highly correlated variables can be removed, correlated variables can be combined, and covariance-based principal component analyses such as Ridge Regression and Lasso Regression can be used[11, 12].

4. Reconstructing the Momentum Shift Prediction Model

The "Momentum Shift Forecasting Model" not only integrates the results of the previously constructed time series model but also innovatively introduces the logistic regression algorithm. Logistic regression, as an efficient binary classification tool, is particularly suitable for accurately predicting the evolution of the situation[3-5, 13, 14]. In intense tennis matches, it is crucial for coaches and players to accurately capture the subtle moments of situation changes. Thanks to the in-depth analysis of logistic regression, it is possible to gain insight and predict these key changes in advance based on multiple factors in the game, such as players' status and match progress, providing a strong basis for the adjustment of the game strategy[15].

4.1. Momentum Shift Prediction Model

4.1.1 Building Logistic Regression

In this paper, logistic regression analysis is performed using the overall data, and the original dataset is first divided into a training set and a test set, where the training set accounts for 80% and the test set accounts for 20%. This paper constructed a logistic regression prediction model and obtained the estimated regression parameters and the p-value of each parameter.

Table 2: Summary of the Logistic Regression

Coefficients	Estimate	P-value
Intercept	7.409e-01	0.22039
server	-1.754e+00	< 2e-16 ***
server_no	8.662e-01	3.62e-14 ***
p1_points_won	2.207e-03	0.01674 *
p1_winner	1.945e+01	0.95543
p1_unf_err	-2.028e+01	0.95994
p1_net_pt	1.152e-01	0.44642
p1_distance_run	1.890e-02	0.00158 **
rally_count	-1.299e-01	1.42e-07 ***
speed_mph	9.814e-03	0.01712 *
serve_width	-8.254e-02	0.01337 *
serve_depth	-5.753e-03	0.94583
return_depth	-1.111e-01	0.06438 •

Note: Significance '***' Indicates $P < 0.001$, '**' Indicates $P < 0.01$, '*' Indicates $P < 0.05$, '•' Indicates $P < 0.1$.

By analyzing the results in Table 2, it is easily to observe that the variables 'server', 'serve_no', 'p1_points_won', 'p1_distance_run', 'rally_count', 'speed_mph', and 'serve_width', are significantly not equal to 0 at 5% significance, indicating that these variables significantly impact the model's performance.

4.1.2 Performance of the Logistic Regression Prediction Model

Next, the Accuracy, Confusion matrix, Precision, Recall, and F1 score of the logistic regression model were calculated.

In calculating accuracy, the number of correctly predicted samples is usually divided by the total number of samples to obtain a percentage value indicating how accurately the model predicts. The formula for calculating Accuracy is as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{2}$$

Accuracy of logistic regression is a commonly used metric to assess the performance of a classification model, which measures the proportion of all samples that the model correctly classifies. In logistic regression, accuracy can be calculated by comparing the categories predicted by the model with the true categories.

It calculated Accuracy ≈ 0.80.

The Confusion Matrix for logistic regression is a 2×2 matrix used to evaluate the performance of a classification model. In a binary classification problem, the confusion matrix is as follows:

$$\begin{bmatrix} TN & FP \\ FN & TP \end{bmatrix} \tag{3}$$

In the confusion matrix, the rows represent the actual categories and the columns represent the categories predicted by the model.

The confusion matrix for logistic regression can be calculated by comparing the predicted and actual values of the model. Based on the predicted and real values, the sample can be categorized into positive and negative categories and the number of *TP*, *TN*, *FP*, and *FN* for each category can be calculated, which in turn populates the confusion matrix.

The confusion matrix is an important tool for evaluating the performance of a classification model, in addition, and it can also be used to calculate metrics such as accuracy, precision, recall, etc., which help to understand the model's classification ability and error.

It calculated the Confusion Matrix as $\begin{bmatrix} 980 & 213 \\ 269 & 966 \end{bmatrix}$.

The Precision of logistic regression is a measure of how many of the samples predicted as positive categories by the classification model are positive categories. Precision can be calculated by the following formula:

$$Precision = \frac{TP}{TP+FP} \tag{4}$$

In other words, the precision rate is the percentage of true positive categories in the samples that the model predicts as positive categories. In logistic regression, the precision rate is one of the most important metrics for evaluating the model's categorization performance. It tells us how accurate the model is in identifying positive categories and is an important metric for whether the model misclassifies.

It calculated the Precision ≈ 0.82.

Recall for logistic regression, also known as Sensitivity or True Positive Rate, is a measure of the proportion of all actual positive examples that are correctly recognized by the classification model. Recall can be calculated by the following formula:

$$Recall = \frac{TP}{TP+FN} \tag{5}$$

Recall tells us the proportion of actual positive examples that the model recognizes. In some scenarios, recall is more important than precision in logistic regression, recall is one of the most important metrics for evaluating the classification performance of a model, and along with precision, it provides an assessment of the overall performance of the model.

It calculated the Recall ≈ 0.78

The F1 score of logistic regression is the reconciled average of Precision and Recall, which is used to comprehensively assess the performance of the classification model. The F1 score can be calculated by the following formula:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{6}$$

The calculation of the F1 score combines precision and recall to provide a comprehensive assessment of the performance of the classification model. It is a value between 0 and 1. The closer to 1 indicates better model performance and the closer to 0 indicates worse model performance.

In practice, the F1 score is a commonly used evaluation metric, especially in unbalanced categories and datasets. It takes into account both the accuracy and omission rate of the model and is therefore useful in evaluating classifier performance.

It calculated the F1 score ≈ 0.80.

4.1.3 Analysis of Model Prediction Performance

Table 3: Value of Evaluation Indicators

Accuracy	Confusion Matrix	Precision	Recall	F1 score
0.80	$\begin{bmatrix} 980 & 213 \\ 269 & 966 \end{bmatrix}$	0.82	0.78	0.80

With the results from the Table 3, it can conclude that on the whole the logistic regression prediction model fits well with an overall prediction accuracy of about 80%. That is to say, the model performs well on the overall sample. The following are the reasons for the good fitting results of the logistic regression prediction model:

1. High Accuracy and F1 Score: The logistic regression model has high accuracy and F1 score on the test dataset. This indicates that the model correctly predicts the categories of most samples and has a good balance between positive and negative categories.

2. Good Precision and Recall: The logistic regression model has good precision and recall, which means that the model can accurately identify the true positive categories in samples predicted to be positive and is less likely to misclassify negative categories as positive.
Overall, the model performs well in several aspects and can effectively explain and predict the results in the dataset.

5. Model Discussion

When momentum starts to decline, coaches must swiftly adjust tactics while players need to maintain focus, especially during opponent serves, to prevent momentum loss. It's crucial to adapt strategies and mindset in response to increasing point differences, unreturnable shots by the opponent, and errors. Minimizing mistakes and stabilizing mentality on the court are essential to counteract negative momentum impacts. The ability to remain calm and adjust tactics is key to reversing unfavorable momentum trends and securing victory. Both coaches and players must work to turn the tide against losing points and momentum.

6. Conclusion

This study developed a quantitative model of momentum in sports by utilizing time series modeling, correlation analysis, and logistic regression. Time series modeling laid the groundwork for data analysis and model construction. The findings from correlation analysis indicate that scoring patterns, particularly consecutive scoring and unreturnable shots, positively influence momentum, suggesting that players in advantageous positions should aim for more points. Conversely, momentum is negatively affected when the opponent serves, highlighting the importance of concentration and error avoidance during these times. Additionally, player errors were found to detrimentally impact momentum, emphasizing the need for mental adjustment to mitigate negative momentum trends. Using logistic regression, this study constructed a model predicting momentum shifts with approximately 80% accuracy, validating the approach to quantifying momentum and predicting game dynamics effectively.

This study introduces an innovative approach that integrates sports science with advanced algorithms, demonstrating its feasibility and potential for broad applications in sports science. By employing data mining, time series modeling, correlation analysis, and logistic regression, this text accurately quantified athletes' momentum, analyzed influencing factors, and predicted score trends. This method offers new perspectives in sports science, enabling the quantification of momentum, understanding its trends, and aiding in the development of scientific training and competition strategies. Overall, it serves as a valuable example of interdisciplinary research, offering insights and inspiration for future work in sports science.

References

[1] Xin Chi, Zhao Xueqing. Analysis and prediction of sports performance based on time series analysis[J]. Science and Technology Wind, 2012(16): 185-186.

[2] Zhang Xiaolong. Application of time series autoregressive (AR) model in sports forecasting[J]. Journal of Beijing Sport University, 2010,33(02): 86-88.

[3] Zhenhui Li, Yushan Zhang. Predicting NBA games using long and short-term memory networks[J]. Computer Applications, 2021,41(z2): 98-102.

[4] Rong Zhang. Predicting results and analyzing players in tennis matches[D]. Yunnan University, 2022.

[5] Jiang T, Li Q. The match results of men's singles matches in professional tennis [D]. Research on the influencing factors of winners and losers of men's singles matches in professional tennis tournaments - based on the four Grand Slam tournaments from 2014 to 2018[J]. China Sports Science and Technology, 2021,57(07): 62-68.

[6] Zhong Tingting. Research on the competitive performance characteristics of excellent male singles players in badminton today[J]. Anhui Sports Science and Technology, 2022,43(3): 27-32, 45.

[7] SHEN Jingyao, ZHU Yuanwren, MEN Fangyu. Characteristics of athletes' physical energy consumption in women's singles matches of tennis grand slam and its correlation with the victory or defeat of the matches[J]. Hubei Sports Science and Technology, 2023,42(2): 184-188.

[8] WANG Junsheng, GUO Zizhao, LIU Bing, et al. Analysis of performance and improvement strategies of elite women beach volleyball players in China[J]. Sports Science and Technology Literature Bulletin, 2023,31(10): 65-69.

[9] Liu Jindong, Wang Ticheng, Chen Yonghong, et al. Improved method and application based on Anderson-Darling normality test[J]. Practice and Understanding of Mathematics, 2023,53(04): 184-193.

[10] Zhang Weifeng, Xu Weichao. Robustness analysis of Spearman's simple correlation coefficient and Gini gamma correlation coefficient against impulse noise[J]. Electronic World, 2020(10): 81-82.

[11] Liu F, Dong Fenyi. Research on diagnosis and treatment methods of multicollinearity in econometrics[J]. Journal of Zhongyuan Institute of Technology, 2020,31(1): 44-48, 55.

[12] Lin Leyi. Application of ridge regression in eliminating multicollinearity[J]. Journal of Liaodong College (Natural Science Edition), 2020,27(4): 274-278.

[13] Fu Wei. Sports team performance prediction based on data-driven and data envelopment analysis[J]. Journal of Shandong Sports Institute, 2021,37(4): 102-111.

[14] CHEN Zhenxiang, LIU Lingjun, HAN Dong. Exploration on the evaluation standard of serving technique utilization of Chinese outstanding male volleyball players--an empirical study based on the type of serve, ball speed and effect[J]. Journal of Beijing Sport University, 2023,46(12): 80-91.

[15] HU Qian, HU Yao, LIU Wei. Principal component estimation and empirical analysis based on logistic regression[J]. Economic Mathematics, 2020,37(04): 123-129.