|  |  |  |
| --- | --- | --- |
| **Problem Chosen** C | **2024 MCM/ICM Summary Sheet** | **Team Control Number** 2401509 |

**Gather Your Momentum: Model Capture of Momentum in Tennis**

**Summary**

Momentum is the invisible indicator of the competition between players.

**Keywords:** keyword1; keyword2; keyword3; keyword4

Contents

[1 Introduction 3](#_Toc158067027)

[1.1 Problem Background 3](#_Toc158067028)

[1.2 Restatement of the Problem 4](#_Toc158067029)

[1.3 Our Work 4](#_Toc158067030)

[2 Assumptions and Justifications 5](#_Toc158067031)

[3 Notations and Glossary 5](#_Toc158067032)

[3.1 Notations 5](#_Toc158067033)

[3.2 Glossary 6](#_Toc158067034)

[4 The Data 6](#_Toc158067035)

[4.1 Data Description and Preprocessing 6](#_Toc158067036)

[5 Dynamic Match Flow Depiction Model 6](#_Toc158067037)

[5.1 Description of Player Performance 6](#_Toc158067038)

[5.1.1 Determination of component *V* 6](#_Toc158067039)

[5.1.2 Determination of component *Q* 8](#_Toc158067040)

[5.2 Model Solution and Match Flow Visualization 8](#_Toc158067041)

[6 Correlation Analysis of Momentum and Match Result 10](#_Toc158067042)

[6.1 Swings in Play Analysis 10](#_Toc158067043)

[6.2 Runs of Success Analysis 11](#_Toc158067044)

[6.3 ARIMA and White Noise Test 11](#_Toc158067045)

[7 Match Swing Prediction and Affecting Factors 12](#_Toc158067046)

[7.1 Kalman Filtering Analysis 12](#_Toc158067047)

[7.2 Bayesian Changepoint Detection 12](#_Toc158067048)

[8 Model Evaluation and Extension 12](#_Toc158067049)

[8.1 Confusion Matrix Analysis 12](#_Toc158067050)

[9 Conclusions 13](#_Toc158067051)

[10 Memorandum 14](#_Toc158067052)

[References 15](#_Toc158067053)

[Appendices 16](#_Toc158067054)

# Introduction

## Problem Background

Like many other competitive sports, tennis is highly intensive, fast-paced and stochastic. The final outcome of a battle between tennis masters not only depends on strength, but is also a matter of "luck", the "feeling", or, the "momentum".

Some have been arguing about the existence of "pace" in many intense events where there is a non-stop flow, like tests and competitions. It seems like there are some critical moments that could get a person to catch up with the pace of a game, making him feel at ease or even leading him all the way to victory. In other words, the "momentum" is what makes swings occur when it gathers up, and you begin to reign superiority and feel that the Goddess of Victory is smiling at you. It sounds just like the VS bar in duel games. You get to win when you prevail in momentum.

There are indeed something indicating the validity of the notion of "momentum". For example, many of us may have such experience that we just "get the feel" after successfully catching that particular shot. The phenomenon could be attributed to complex possible factors including mental motivation and physical adaption, etc. However, the existence of such "momentum" is still rather ambiguous. The perception of momentum in games is much a subjective feeling. Meanwhile, it is always hard to identify when "momentum" is about to gather, and in what form it is coming in your favor. Is it a serve? Or the next shot? If players want to gain their momentum, it is vital to figure out what makes momentum, when it tends to rise, and pay special attention to the potential "critical moments".

|  |
| --- |
|  |

Figure 1. Momentum report from Wimbledon official website[1].

## Restatement of the Problem

Our main goal is to verify the existence of momentum in tennis games through quantified analysis of Wimbledon 2023 men's matches. Then we want to design a feasible strategy to predict and capture momentum in games and put forward a paradigm for better momentum utilization in favor of tennis players.

Considering both our primitive goals and real-life demand, we summarized the problems as follows.

* Develop a model to capture the flow of a game as every point occurs. Apply the model to one or more of the Wimbledon 2023 men's matches. The model should be able to identify which player is playing better as well as how much better they are performing at a given time. Then present a visualization of the match flow based on the model.
* Use the model to assess the claim of a tennis coach that swings and success are random instead of a result influenced by momentum.
* Find out possible indicators that help determine swings. Specifically, complete the following two tasks:
* Develop a model using data from at least one match that predicts the swings and find out if there are most related factors.
* Provide advice for a player going to play a new match against a different rival.
* Test the model with data of one or more of the other matches, and evaluate the quality of swing prediction. If the model would sometimes perform poorly, find out some omitted factors. Evaluate how generalizable the model is to other matches.
* Provide a report of no more than 25 pages with a one- to two-page memo of the research result and advice for coaches as for the role of momentum and how to help players better utilize it.

## Our Work

|  |
| --- |
|  |

Figure 2. Work Overview.

# Assumptions and Justifications

Our model is established on the basis of the following assumptions:

* **Data authenticity.**

The only data used in this paper is the provided data from Wimbledon 2023 men's matches. The validity of our research results is based on the assumption that all data is authentic with all matches played under Wimbledon's rules and regulations.

* **Possible abnormalities like injuries in games are not considered.**

We assume that there is no exceptional situation in all recorded matches that could have significant impact on the player's performance. Such situation includes a twist in the ankle, stimulant usage, and at the same time any mental trauma the player is going through. Considering that such events are especially rare in official contests, they were taken out of consideration so that we could focus on the effects of momentum.

* **The effects of players' mental state and technical skill on their performance satisfy the Markov property.**

We assume that the players' mental state and technical skill at each point have Markov property. In other words, the current performance of any player is only determined by his current mental state and the type of the current point, but not any past mental change and past points. We made this assumption because it is sensible that a player is influenced by his current state, and how he got this state matters little.

# Notations and Glossary

## Notations

The key mathematical notations used in this paper are listed in Table 1.

Table 1: Notations used in this paper

|  |  |
| --- | --- |
| **Symbol** | **Description** |
|  | State vector of the player's performance. |
|  | Mental state factor in affecting player performance. |
|  | Technical performance factor in affecting player performance. |
|  | Mental effect from memorized points of the game. |
|  | Mental effect from the score of prior game. |
|  | Mental effect from current estimated win rate. |
|  | Number of memorized points in the sliding window. |
|  | Whether the player won the i-1th point before the current point. |
|  | Weight of the  memorized point before the current point . |

## Glossary

**Technical Performance.** The technical performance of the current point means the technical features of the shot, including whether the player is the server, whether the player hits an ace, the player's distance ran during the point, etc. The items involved are determined using PCA (Principal component analysis). For the complete list of data items involved in the technical performance variable, please refer to Appendix 1.

**Mental state.** The mental state of the player at current point is decided by the player's scoring history. It is determined by the current score of the game.

# The Data

## Data Description and Preprocessing

The only data used in this paper is the provided data of Wimbledon 2023 men's matches. The data covers every point after each match's first two rounds.

We found missing values in the items *serve\_width*, *serve\_depth* and *return\_depth*. The missing values were simply dropped out because of lack of effective approach of filling. All values in the items *rally\_count* and *serve\_width* of the match indexed 1310 were invalid. The match was taken out of consideration when the model involved these two items.

# Dynamic Match Flow Depiction Model

## Description of Player Performance

Before looking into the effects of momentum on players' performance and the flow of play on the whole, we need to quantify our description of players' performance first.

We investigated the features of the data set for possible components that could most indicate a player's current performance at a time point, and we found that the data includes two main types of information: one is the player's scoring history, and the other is the technical performance of the current point. Considering that the player's current mental state and technique are most likely to affect the match results, and the player's mental state depends mainly on his current scoring, we chose the scoring history and technique factor as the two indicators of player's current performance.

A player's overall performance is denoted by the state vector *Z*, which is defined as follows:

|  |  |
| --- | --- |
|  | () |

Here *V* and *Q* are seen as functions of the current point numbered *N* in this match. Separately, *V* is the player's mental state fluctuation estimated by his recent scoring and *Q* is the technical performance reflected by the current shot. The components have range [0,1]. When this player is superior to his rival in this factor, the value of this component is closer to 1.

### Determination of component *V*

We applied the sliding window method and the Markov chain model to determine the component *V*. We used the weighted average of the player's recent scoring history to describe the overall situation of the player's scoring. *V* is represented by the equation below:

|  |  |
| --- | --- |
|  | () |

where  is the variable describing the mental effect from the performance of the player in recent memorized points of shots and is calculated as the weighted average of these memorized points.  is the estimated mental effect from the score of the prior game.  is the mental effect from the estimated possibility of winning based on the player's performance in former sets. Note that  takes into consideration the time point when the point occurs, while  and  are only affected by the current score.

The determination of  applied the sliding window technique. For each point we consider the effect of the past memorized points (denoted as ) on the current point. The value of  depends on the order of the point. Specifically, if the current point is the  point of the game,  takes value:

|  |  |
| --- | --- |
|  | () |

Considering that points with more time proximity have more effect on the player's current mental state,  is represented as the weighted average:

|  |  |
| --- | --- |
|  | () |

 is the virtual variable denoting whether the player wins the  point before the current point in this match numbered . The weight of the  memorized point  before the current point  follows the definition presented below.

|  |  |
| --- | --- |
|  | () |

Similarly, the determination of  is also by weighted average. The weight is decided by the score of the prior game. Note that when the total score of the two players in the prior game is n, and the score of player 1 is m, The values of  of the two players should add up to 1. Therefore, the weight is determined using combinatorial number.

|  |  |
| --- | --- |
|  | () |

so that  satisfies the constraint:

|  |  |
| --- | --- |
|  | () |

The determination of  used the Markov chain model. We assumed that the effects of players' mental state on their performance have Markov property. In other words, the estimated possibility, or the player's performance, should only depend on the player's current status. The match is modeled as Markov chain to predict the possibility of winning. The match status is described by the score. The value of  equals the win rate under current status:

|  |  |
| --- | --- |
|  | () |

Therefore, based on our modeling of the match status, the state transition process of Markov chain with each time step could be depicted by the graph below.

|  |
| --- |
|  |

Figure 3. State transition diagram of Markov chain.

When the state space of Markov chain is finite, the Markov chain can be depicted with the state transition matrix. Arranging the single-step state transition probability of all possible states in the matrix, we got the state transition matrix:

|  |  |
| --- | --- |
|  | () |

### Determination of component *Q*

The component *Q* is determined using PCA (Principal component analysis) for data dimension reduction. Based on the 29 indicators reflecting technical performance obtained from processing of the data set, we generated 10 synthesized principal components which are the linear combination of the items (Appendix 1).

Assuming that there are *p* original factors and *m* principal components, the principal components are linear combination of original technical performance indicators:

|  |  |
| --- | --- |
|  | () |

When all principal components are mutually independent and each is the linear combination that satisfies the condition and has largest variance, the dimension of indicator can be reduced while retaining information to the utmost.

## Model Solution and Match Flow Visualization

**Solution of component V**

Based on MATLAB solution, set ,, as 0.7, 0.2, 0.1.

For the solution of , under each possible score, the estimated win rate is presented in the table:

Table 2: Estimated win rate under different score

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| 0 | 0.50 | 0.57 | 0.57 | 0.92 | 0.99 | 0.99 | - |
| 1 | 0.43 | 0.50 | 0.66 | 0.79 | 0.99 | 0.99 | - |
| 2 | 0.22 | 0.34 | 0.50 | 0.69 | 0.90 | 0.95 | - |
| 3 | 0.08 | 0.21 | 0.31 | 0.50 | 0.72 | 0.98 | - |
| 4 | 0.01 | 0.15 | 0.10 | 0.28 | 0.50 | 0.76 | - |
| 5 | 0.01 | 0.01 | 0.05 | 0.02 | 0.24 | 0.50 | 0.65 |
| 6 | - | - | - | - | - | 0.35 | 0.50 |

**Solution of component Q**

In order to ensure model accuracy, we chose to generate 10 principal components. Use SPSS solution to determine the principal components. PCA result and total sum of contribution to variance is presented below. The variance explained by all principal components accounts for 80% of total variance.

|  |
| --- |
|  |

Figure 4. PCA result and principal components contribution to total variance.

We applied the model to the match 2023-Wimbledon-1701 between Carlos Alcaraz and Novak Djokovic. Components are calculated after smoothing process and normalization. The visualized match flow is presented below.

|  |
| --- |
|  |

Figure 5. Visualization of the match flow of match 2023-Wimbledon-1701.

According to the visualization, in the first set, the yellow blocks significantly prevail over blue blocks, showing that Djokovic was dominating in this set. In the second set, there are mostly green blocks and the values are close to axis, indicating that the two were evenly matched in power in this set. Similarly, in the third set, there are significantly more blue blocks, it means that Alcaraz was performing remarkably better. In the fourth set, there are more green blocks and the flow turned in favor of Djokovic. In the fifth set, the blue blocks slightly surpass the green blocks, and the set ended up in Alcaraz's victory.

The result depicts the real situation of the match well. It demonstrates the feasibility of our model.

# Correlation Analysis of Momentum and Match Result

## Swings in Play Analysis

**Swing Quantification**

To verify the correlation between momentum and the swings and results of the match, we conducted separate research. For investigation into the effects of momentum on swings in matches, we first completed quantification of swings in play. When the difference of player scores reaches 5, we consider that a swing happens.

**Monte Carlo Simulation**

To establish the random match flow model of the coach without consideration of effects of momentum, we conducted Monte Carlo simulation method to simulate random swing occurrence probability in match. Number of simulations was set to 200.

We considered the strength of player, whether the player serves the ball and random factors in the simulation model. The player's strength is approximately evaluated by the score of the final game, which we considered as most representative of player's relative strength. Empirically, the server tends to have higher probability of winning. When the player is the server, we take this factor into consideration and multiply the player's result with an advantageous factor.

|  |
| --- |
|  |

Figure 6. Monte Carlo Simulation result.

**Logistic Evaluation Model**

The result of Monte Carlo simulation (corresponding to the coach's random match flow model) is compared to our match flow depiction model. The mean and variance of match swing prediction of the two models are:

Table 3: Model prediction results of player 1.

|  |  |  |
| --- | --- | --- |
|  | Mean | Variance |
| Random Match Flow Model | 0.568 | 0.130 |
| Momentum Match Flow Model | 0.654 | 0.020 |

Table 4: Model prediction results of player 2.

|  |  |  |
| --- | --- | --- |
|  | Mean | Variance |
| Random Match Flow Model | 0.508 | 0.100 |
| Momentum Match Flow Model | 0.317 | 0.012 |

The variance show that the prediction of our model based on momentum is more stable and reliable.

For comprehensive evaluation of the prediction of the two models, we applied the prediction evaluation model based on the Logistic function. A simple Logistic function can be expressed by the following equation. It is continuous and monotonically increasing. When , . The function tends to 1 when the argument tends to  and tends to 0 when the argument tends to .

|  |  |
| --- | --- |
|  | () |

Consider the inverse of Logistic function:

|  |  |
| --- | --- |
|  | () |

When prediction tends to 1, the function's value tends to . Meanwhile, when prediction tends to -1, the function tends to . To avoid overly significant impact of the prediction of a single point on the overall score of the model, a single scoring in the model is restricted in [-2,2]. When the scoring goes beyond the range, it is counted as the boundary value.

We evaluated all predictions of each model and calculated the average score as the model's overall score. The Logistic score of our model predicting result of player 1 and 2 were 7 and 8.9, while the random match flow model scored 1.4 and 0.6. The momentum match flow model performed significantly better in swing prediction.

## Runs of Success Analysis

Runs of success is the consecutive scoring of player.

Similar to analysis of swings in play, we first conducted 200 times of Monte Carlo simulation to establish the coach's random match flow model. Then we compared the predictions of the two models. The variance shows that the result of our model is more stable.

Table 5: Model prediction results of player 1.

|  |  |  |
| --- | --- | --- |
|  | Mean | Variance |
| Random Match Flow Model | 0.612 | 0.130 |
| Momentum Match Flow Model | 0.624 | 0.020 |

Table 6: Model prediction results of player 2.

|  |  |  |
| --- | --- | --- |
|  | Mean | Variance |
| Random Match Flow Model | 0.468 | 0.080 |
| Momentum Match Flow Model | 0.387 | 0.020 |

The Logistic score of our model predicting result of player 1 and 2 were 5.6 and 5.7, while the random match flow model scored 3.1 and 1.9. It demonstrates that our model based on momentum is also significantly better in predicting occurrence of runs of success. These results provide valid proof for the important role of momentum in tennis.

## ARIMA and White Noise Test

We assumed that momentum itself is not random as well. To verify the hypothesis, we modeled momentum with ARIMA time series model. Make null hypothesis that the time series of momentum is significantly different from Gaussian white noise.

For the technical performance component *Q*, we selected the former 279 records of Wimbledon 2023 final match as training group, the remaining records of the match as testing group. MATLAB solution showed that the time series of the technical performance component *Q* fits ARIMA(1,1,5).

|  |
| --- |
|  |

Figure 7. ARIMA regression result of component *Q*.

For the mental state component *V*, we selected the former 320 records of the same match as training group, the rest left as testing group. The time series fits ARIMA(1,1,6).

|  |
| --- |
|  |

Figure 8. ARIMA regression result of component *V*.

The results show ARIMA's ideal regression effect of momentum. This proves that momentum is not random, but is correlated to the time factor.

# Match Swing Prediction and Affecting Factors

## Kalman Filtering Analysis

We applied Kalman Filter to our dynamic model to predict occurrence of swings in matches.

## Bayesian Changepoint Detection

To find out possible factors most related to determining swings in matches, we used the Bayesian Online Changepoint Detection to process the technical performance of each applicable match. Note that we chose the technical performance factor as the indicator to avoid the impact of winning or losing scores on the mental state of the player. This is because in higher levels of matches, the mental factor plays minor role in affecting the performance of players. This could help us focus more on the technical and actionable causes of winning or losing points.

Bayesian Changepoint Detection helps detect the position of parameter changepoint caused by systematic factors in the time series. The method predicts the posterior distribution of the run length of current time point since last changepoint occurred, and calculates the posterior distribution of the next point with recursion method. Using the Bayesian formular, the

posterior distribution of run length  is:

|  |  |
| --- | --- |
|  | () |

Then, the distribution of the next point  could be expressed as:

|  |  |
| --- | --- |
|  | () |

Since the beginning of the match is definitely a changepoint, set the initial value . Apply Bayesian Changepoint Detection to all applicable matches. The following are examples of Bayesian Changepoint Detection results of match 1501, 1502, 1503, 1504, 1601 and 1602.

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |

Figure 6. Bayesian Changepoint Detection of match 1501-1602.

129 significant breakpoints were detected in analysis result. We investigated the primitive causes of each point, while the factor of who is the server was sorted out. Then, we picked out 5 factors with highest frequency of occurrence:

Table 7. 5 Most related factors affecting match swings.

|  |  |  |
| --- | --- | --- |
| Factor | Explanation | Frequency |
|  |  | 31/129 |
|  |  | 27/129 |
|  |  | 23/129 |
|  |  | 19/129 |
|  |  | 18/129 |

## Advice for players

# Model Evaluation and Extension

## Confusion Matrix Analysis

# Conclusions

|  |
| --- |
|  |

# Memorandum

**To:** Coaches

**From:** MCM Team #2401509

**Date:** February 5, 2024

**Subject:** The important role of momentum and how to prepare your players for it

Dear Sir or Madam:

The result of competition between tennis players, especially elite players, depends not only on their professional skills and technique, but also to a large extent depends on their mental state, physical adaptation and many other sophisticated factors. Through our research, we have verified the particular role of "momentum" in tennis matches using mathematic model, and we would like to offer you some related advice as we introduce our model.

Yours sincerely,

MCM Team #2401509

# References

[1]: AELTC. “Home-The Championships, Wimbledon.” Wimbledon, 2 January, 2024.

https://www.wimbledon.com/index.html.

[2]: Adams, Ryan P. and David John Cameron MacKay. “Bayesian Online Changepoint Detection.” arXiv: Machine Learning (2007): n. pag.

# Appendices

|  |
| --- |
| Appendix 1 |
| Introduce: Data items involved in technical performance variable. |
| **Appendix Table 1: 29 original technical performance indicators**   |  |  | | --- | --- | | **Variable** | **Description** | | who\_to\_serve | whether the player is the server | | consumption\_of\_strength | physical comsumption of recent 20 points | | consumption | total physical consumption | | p1\_ace | player 1 hit an untouchable winning serve | | p2\_ace | player 2 hit an untouchable winning serve | | p1\_winner | player 1 hit an untouchable winning shot | | p2\_winner | player 2 hit an untouchable winning shot | | p1\_double\_fault | player 1 missed both serves and lost the point | | p2\_double\_fault | player 2 missed both serves and lost the point | | p1\_unf\_err | player 1 made an unforced error | | p2\_unf\_err | player 2 made an unforced error | | p1\_net\_pt | player 1 made it to the net | | p2\_net\_pt | player 2 made it to the net | | p1\_net\_pt\_won | player 1 won the point while at the net | | p2\_net\_pt\_won | player 2 won the point while at the net | | p1\_break\_pt | player 1 has an opportunity to win a game player 2 is serving | | p2\_break\_pt | player 2 has an opportunity to win a game player 1 is serving | | p1\_break\_pt\_won | player 1 won the game player 2 is serving | | p2\_break\_pt\_won | player 2 won the game player 1 is serving | | p1\_break\_pt\_missed | player 1 missed an opportunity to win a game player 2 is serving | | p2\_break\_pt\_missed | player 2 missed an opportunity to win a game player 1 is serving | | p1\_serve\_speed\_ave | player 1's average speed when serving | | p2\_serve\_speed\_ave | player 2's average speed when serving | | width\_1 | player 1's direction of serve | | width\_2 | player 2's direction of serve | | depth\_1 | player 1's depth of serve | | depth\_2 | player 2's depth of serve | | depth\_return\_1 | player 1's depth of return | | depth\_return\_2 | player 2's depth of return | |

|  |
| --- |
| Appendix 2 |
| Introduce: 这里放上附录2的介绍 |
|  |