Problem C	**2024 MCM Summary sheet**	**Team Number 2425485**

How tennis momentum shifts during matches

As tennis become a worldwide popular sport, there is always discussion about whether there is momentum exists in tennis matches or not. In this scenario, the word "**momentum**" is not the definition we usually talk about in physics, but is a special power that influences the wins or losses of sides in tennis matches. As this becoming a growing concern for tennis coaches and athletes, a model that can be used to depict and predict the flow pattern is necessary for future training of athletes.

Throughout our paper, we first describe the model we have for depicting the flow in matches, and then another model that is used to predict point wins and losses. In the end, we combine these two models to predict the flow of momentum in tennis matches.

Going into details, we first establish a model that is used to catch and depict the flow from matches with certain data. Since the first version does not work in the ways we want (Fig. 1), we develop another version and add more factors into considerations. After testing with the model, we generalize it to other matches (Fig. 2, 3).

Then, we build prediction models to **identify indicators** of change in flow between two tennis players in a given match. The first model is a **logistic regression model** that directly predicts the outcome of the next ball for one of the players, either winning or losing. We identify the crucial indicators for winning anticipation and examine the stability of these indicators (see Table. 3 and Fig. 5). We then test the model on different matches of the same player and other players (see Table. 4, 5). We also explore whether a model built with an early portion of the match effectively predicts the result of later matches (see Fig. 6).

After that, we build a set of **linear regression models** that predict player momentum and identify flow changes. We recognize and assess the reliable indicators based on this model (see Table. 6 and Fig. 7). We also test the model against data of other matches with two accuracy testing methods and identify that method 2 of this model yields the optimal accuracy (see Table. 7, 8).

We also include discussions about whether the flow is randomized based on the first model we have. By forming nearly random winners with our sample data and creating visualizations to compare the two graphs, we conclude that the flow does have some patterns, and is not random and is predictable (Fig. 4).

In short, the models we describe in this report catch some patterns and indicators about the flow of momentum in tennis matches. We also state the conditions that our models work on. We also attach a 1-page memo to explain the finding we have for momentum to all tennis coaches and athletes.

[**Keywords**: tennis, momentum, logistic regression, linear regression, indicators, match]

Contents

1. Introduction

As famous sports, tennis matches are always attractive for the fans of this sport, and some of them may find that the matches possibly have "flows", or "momentum", which favors different players at different times. The momentum has become a factor concerning the athletes and coaches, which is likely to affect the player's performance and it can be used as an indicator for coaches to help athletes act better in matches, but there is still not convinced study and evidence which prove the existence of momentum.

In this report paper, we will construct models for athletes' performance to depict the flow of momentum during selected matches, particularly matches from Wimbledon 2023, and determine if there are indicators of momentum that exist.

2. Assumptions

To simplify the question, we made the following assumptions about tennis athletes

- The difference between athletes' physical and mental conditions is negligible.
- The opponent's actions do not affect the performance of other players, except for unforced errors and distance run.
- The environment of matches does not have any influence on players.

3. Modelling for Flow in Matches

The first thing we do is develop a mathematical model that defines the flow when points occur during a whole match. For the simplicity and fairness in our expression, we use "player 1" and "player 2" refer to the two athletes in a match.

3.1 The initial model

The basic model we have is a linear equation that calculate the intensity, which we will use flow for the rest of this part, of the whole match. The pick of factors, such as serve, is according to the tennis metrics that coaches usually used to evaluate the performance in a selected match [1]. For performance flow at a given point number n and variables from Table 1, the F_n is calculate by,

$$P_n = \begin{cases} 1, & n = 1 \\ 0.9 \cdot P_{\{n-1\}} + S + 0.1 \cdot A + 0.05 \cdot V - 0.15 \cdot D, & n \geq 2 \end{cases}$$

***Table 1**: Variable description for the initial model*

Symbols	Description	Possible Values
S	The player wins the point or not	0 or 1
A	The point wins on an ace or not	0 or 1
V	The player serves during this point or not	0 or 1
D	The point is lost on a double fault or not	0 or 1

For the first point number in the whole match, we set its flow to 1 as an initial value, to ensure that there will not be a negative performance flow during the match. For cases with $n \geq 2$, we calculate the flow using both the player's performance during this point and the flow we calculate for the $n-1$ point. The coefficient 0.9 represents the penalty in flow for the player if the player does not win points during this round.

However, the visualization is not visually intuitive. There are so many turning points that make the graph appear overly convoluted. By applying the LOESS [2] method to find a smooth local regression line, the graph appears in a nearly symmetric pattern. The visualization of the match with match ID "2023-wimbledon-1301" is,

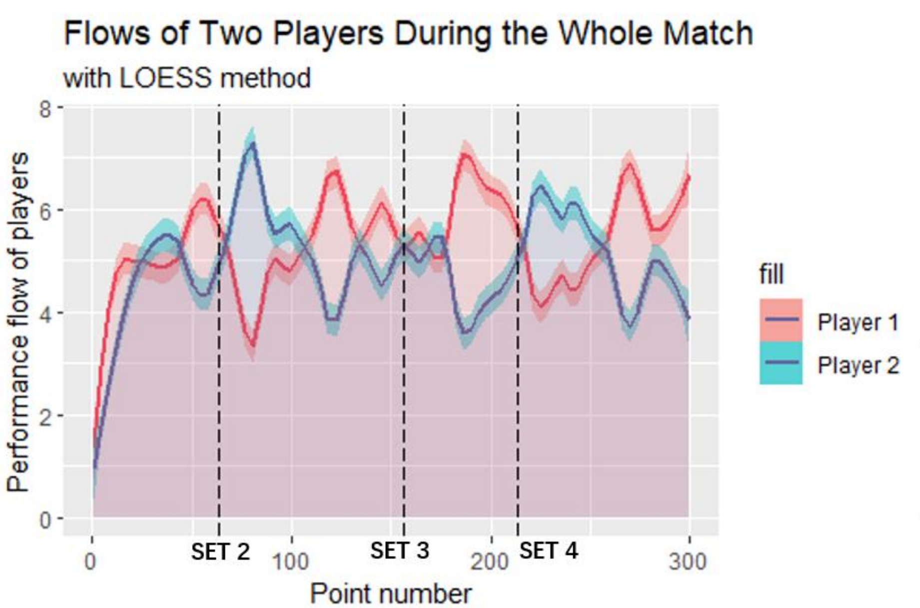

***Figure 1**: Visualization of 1301 match using the initial model*

We assume that there may exist an equilibrium between the two players. In order to prove that, we treat the factors for serving, ace, and double fault as noises since they only constitute a small portion of the performance flow. Therefore, the equation we have is,

$$S_n = P_{n_1} + P_{n_2} = 0.9 \cdot S_{n-1} + 1$$
$$\therefore S_n = 10$$
$$P_{n_1} = P_{n_2} = 5$$

The result shows that the performance flow of the two players will approximately reach an equilibrium when their values are equal to 5 at the same time, and makes the pattern of flow not explicit for both two players.

3.2 The improvement on initial model

To modify the initial model, we take unforced error, number of serve, and the breaks between different sets into consideration.

Athletes may experience anxiety due to unforced errors, leading to a negative impact on their performance [3]. For example, an athlete may be too anxious after he makes unforced errors in sequence. Therefore, we use this as a negative term in our calculation.

There are previous studies that show that there is predictable serve effectiveness for professional athletes [4], which for first serve is 75% and the second serve is 55%, and we use it as a positive impact factor in our calculation.

The long break time between different sets can also be a deduction in the performance flow of athletes [5], as we also considered this into the factor a, which equals 0.3 instead of 0.9 when there is a set break.

For performance flow at a given point number n, with variables from Table 2, the F_n is calculated by,

$$P_n = \begin{cases} 1, & n = 1 \\ a \cdot P_{n-1} + (1+b) \cdot S + 0.1 \cdot A - 0.15 \cdot U, & n \geq 2 \end{cases}$$

***Table 2**: Variable description for the improved model*

Symbols	Description	Possible Values
a	Deduction factor	$a = 0.9$, if there is no set change $a = 0.3$, if there is set change
b	Serve number factor	$b = 0.75$, if serve number is 1 $b = 0.55$, if serve number is 2

		$b = 0$, otherwise
S	The player wins the point or not	0 or 1
A	The point wins on an ace or not	0 or 1
V	The player serves during this point or not	0 or 1
U	The point is lost on an unforced error or not	2, if there is a double fault 1, if there is another unforced error 0, otherwise

By applying this model and the same smooth method on the match with match ID "2023-wimbledon-1301", we get this visualization,

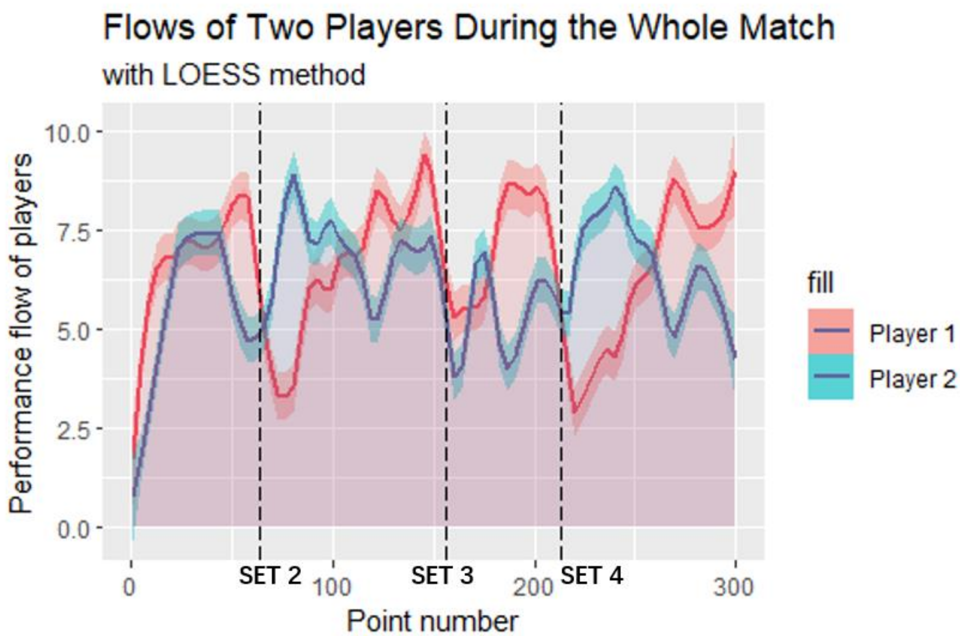

***Figure 2**: Visualization of 1301 match using the improved model*

Comparing this graph with the previous one, it is clear that the performance of the two players varies from time to time. For another match with match id "2023-wimbledon-1302", the graph of the whole match is,

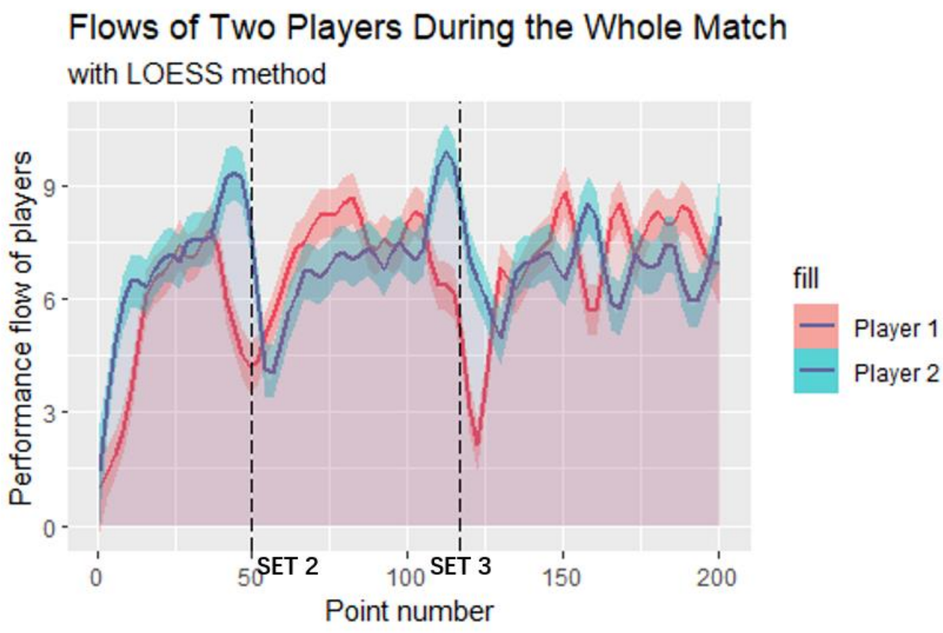

Figure 3*: Visualization of 1302 match using the improved model*

This graph shows that player 1 in this game is overall competitive with player 2, but lost his momentum at the beginning of the third set, and lost this match eventually.

3.3 Interpretation of Visualizations

The interpretation of visualization is easy: the player on the top is the one have leads in points. Take the Figure 3 as an example. At each end of each set, player 2 is on the top of player 1, indicating that player 2 wins each set, and wins the game at the end of the third set.

Still, with Figure 3 as an example, we can also make a detailed interpretation of this visualization. At the beginning of the first set, the two players are equivalent in the points they win. However, at the end of the first set, there is a big drop in the performance of Player 1, and Player 2 wins more points than Player 1 as he wins the set as a result. Then at the beginning of the second set, the two players are still evenly matched, and same with the previous set, player 1 has a drop and play 2 wins this set. At the beginning of the third set, player 1, who might be too stressed mentally, has a big drop in performance. Even though the performance of he and Player 2 are balanced later, Player 1 still loses this set and the whole match. From this visualization, we can conclude roughly that Player 1 may need more training in physical conditions since he can hardly follow his opponent in the latter set.

The visualization made from the mathematical model we developed can not only depict the flow of the whole match but also can serve as a conclusion of patterns for a certain player, while the latter can be used as a tool for analyzing and developing the training plan targeting the shortcomings for this athlete.

4. Momentum: Random or Predictable

Since the time the concept of momentum was introduced in tennis, there is still no common agreement on whether it exists for some patterns or not. Some may argue that the flow or momentum is nonsense; the probability of winning points is equal between the two players. However, from the models we constructed earlier in this paper, this statement, which considers the flow as totally random, may not be true.

To prove that the momentum is not random, we use the match with match ID “2023-wimbledon-1301” for illustrations. We first generate random point winners based on some conditions, which can make the randomized point winner much more realistic as it is possible to happen in a tennis match,

- If there is an unforced error or ace, use the winner in the original data.
- Otherwise, generate a random winner between the two players.

For parameters other than point winner in our model, we use the original data in the calculation. Therefore, we get a result in performance flow with a nearly randomized point winner, to test if the flow is actually random during a match. The original and randomized visualization is in Figure 4.

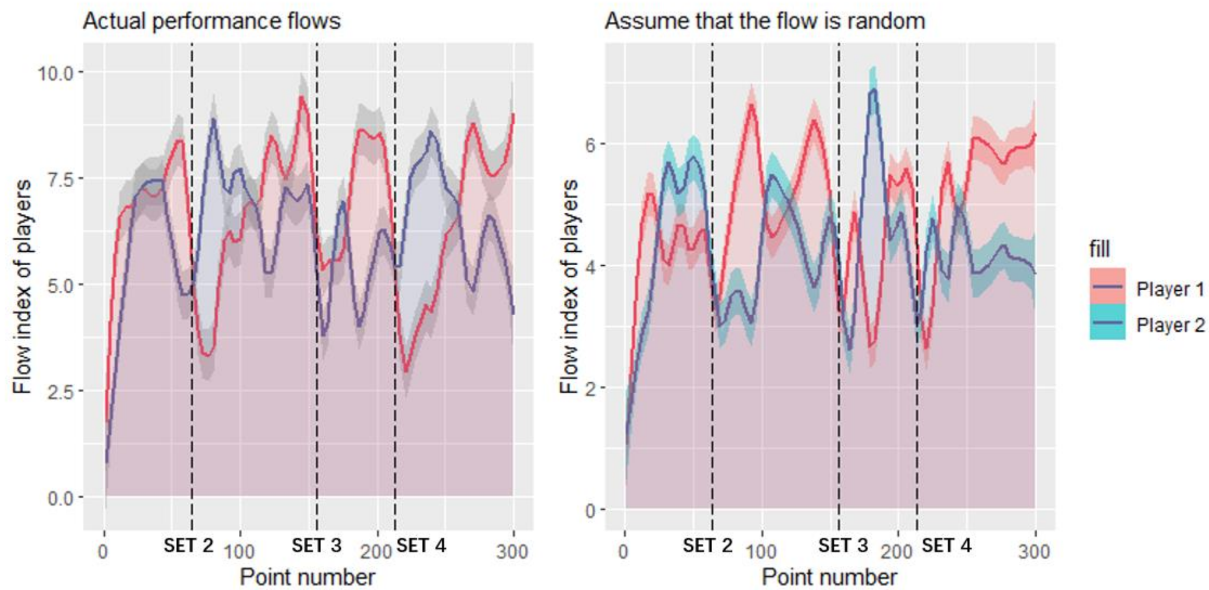

***Figure 4**: the comparation between two visualizations*

From Figure 4, the two visualizations may seem similar as they have the same x index and y index, and have similar colors. However, in the actual visualization, we can see that there are patterns in both Player 1 and Player 2, while in the randomized graph, it is difficult to conclude any patterns that exist in these two players. For example, Player 1 in actual data easily lost points at each beginning of the set, and Player 2 who usually played well at the beginning performed poorly as the point number increased. These patterns somehow show that the flow or momentum is not totally randomized.

5. Prediction Model

To predict player's performance and the indicators of change in performance, we build two regression models based on selected factors and with different target values. Both models are directional, taking independent variable values and producing predictions for a single player.

5.1 Independent Variable Selection and Data Processing

For our prediction models, we hypothesized that the following factors may be indicators of a player's future performance:

- Elapsed time, distance run, and rally count are chosen because as the match progresses, the performance of a player may be affected by these factors due to physical energy expenditure.
- We considered whether the player is the server since the server has a higher probability of winning a point [5].
- Number of winning games, number of points, as well as the results of the preceding two points are included since historical performances may be influential to the result of the current point.
- Exceptional performances, including playing an unforced error or ace, may also be impactful in a player's mentality [3].

To make the data more intuitive in analysis, we do the following data processing to a few of the factors:

- We convert the time in the form "HH:MM:SS" into minutes for easier calculation.
- We subtract the component's distance run from the current player's distance to symbolize the difference in physical exertion.
- We calculate the ratio of the player's winning games and points over the total number of games and points such that the values are more representative in terms of the whole match.
- The values of certain unforced errors are doubled when the player plays a double fault. We also subtract the unforced error values of the opponent since the opponent's mistakes boost the player's confidence and give the player a higher probability of winning [3].

5.2 Logistic Regression Model

Our first prediction model aims to predict whether a player could win the next point based on his performance on the last shot. Since the target variable is binary, either winning or losing, we choose logistic regression as the prediction model since it is ideal for classification.

5.2.1 Model explanation

According to [6], the logistic regression function is defined as:

$$h_i = \frac{1}{1 + e^{-(\beta_0 + \sum_{j=1}^{n} \beta_j x_{ij})}}$$

with m total observations with indices i and n total independent variables with indices j. In the context of our problem, i is the point number and j is the number of independent variables chosen. h_i represents whether the player wins the i^{th} point, x_j represents the j^{th} chosen independent variable, and β_j represents the corresponding coefficient, with β_0 being the intercept. The coefficients, β_j, can be interpreted as the weight or importance of their corresponding independent variables, with a greater magnitude in β_j indicating a more crucial indicator.

The loss function for the above regression model is defined as:

$$\min \ -\sum_{i=1}^{m}[y_i \log\left(\pi\sum_{j=1}^{n}\beta_j x_{ij}\right) + (1-y_i)\log\left(1-\pi\sum_{j=1}^{n}\beta_j x_{ij}\right)$$

The newly introduced variable y_i is the truth class label at the i^{th} point, either 1 or 0 to represent winning or losing.

5.2.2 Indicators identification

The initial model is built using data of player Carlos Alcaraz in the match with match ID "2023-wimbledon-1301". The coefficients for independent variables are listed in Table 3, rounded to the third significant digit.

These coefficients suggest that *game-winning ratio*, *server*, and *previous point result* are the most important indicators of the next point's outcome: a higher game-winning ratio or being the server symbolizes a higher likelihood of winning the next point, while winning the last point leads to a lower likelihood of winning the next. On the other hand, *elapsed time*, *difference in distance run*, and *result of point before last play* negligible roles in determining future point outcome.

Table 3: *Independent Variables and Their Weights of Carlos Alcaraz in Match 1301*

Independent Variable	Coefficient
Elapsed time	0.00246
Difference in distance run	0.0159
Rally count	0.103
Server	0.744
Game-winning ratio	0.819
Point-winning ratio	−0.499
Previous point result	−0.685
Point before last result	−0.0793
Unforced error	0.285
Ace	0.161

To assess the reliability of the indicators, we use the same modeling method to build prediction models of the other matches played by Carlos Alcaraz. Then, we compared and visualized the coefficients of each independent variable from these models, as shown below in Figure 5. The weights of numerous independent variables significantly fluctuate between positive and negative values, including *ace*, *game-winning ratio*, and *rally count*. These factors fail to become indicators due to their instability in correlation types. *Distance run* and *elapsed time* cannot be valid indicators either due to their steady yet trivial coefficient values: *difference in distance run* is barely influential to their performances due to athletes' high physical endurance, while the *elapsed time* affects the stamina of both athletes.

Among the remaining variables, the most stable factor is *server*, while the ranges of other variables are significantly larger than their coefficient values, making them less reliable as indicators.

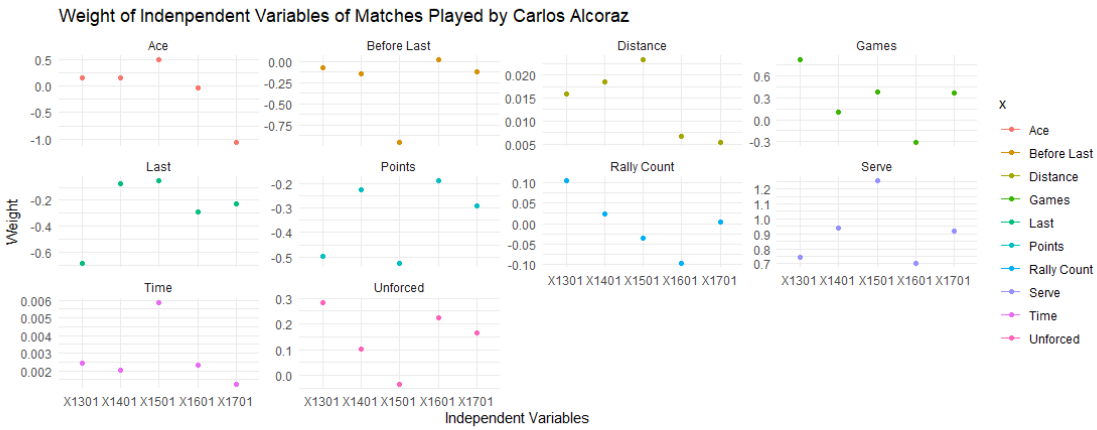

Figure 5: *Weights of Independent Variables of Matches Played by Carlos Alcaraz*

5.2.3 Prediction

To test our model, we use the model built based on Carlos Alcaraz's performance in match "1301", test it over other matches played by the same player, and get the prediction accuracy in Table 4. The accuracies of most of these matches are between 55% and 60%, which are generally low and unreliable for prediction. The accuracy of match "1601" is significantly lower than the other ones, which may be in consequence of the outstanding performance of the opponent which we did not take into consideration.

Table 4 : *Score-Based Model Accuracy of Matches Played by Carlos Alcaraz*

Match ID	1401	1501	1601	1701
Accuracy	0.594	0.590	0.481	0.541

We also tested the model over a random sample of other matches not played by Carlos Alcaraz to test our model's generalizability, and the accuracies for the matches are shown in Table 5. Overall, these accuracies do not differ much from those in Table 4 and remain relatively low in values, which indicates the low reliability yet fine generalization of our prediction model.

Table 5: *Score-Based Model Accuracy of Randomly Chosen Matches*

Match ID	1313	1504	1402	1404
Accuracy	0.557	0.588	0.535	0.607

We then wondered if we could use an early portion of the match data to predict the latter results of the match, and if so, what percentage of the match do we need for best prediction? To find the pattern behind it, we take the data from matches played by the athlete Carlos Alcaraz. For each match, we use the first 10% to 90% of the data with intervals of 10% and then use the model we get to predict the remaining portion of the same match and obtain accuracy. Then, we take the average for the same percentage between different matches, plot the accuracy of each match and the average accuracy, and get Figure 6. From the figure, we can see that the prediction accuracy for individual matches highly depends on the match itself, while on average, the accuracy steadily increases from 10% to 60% but then fluctuates for higher percentages. A possible explanation for the fluctuation is that the high percentage of data for model training leads to overfitting, making the model fail to generalize over the remaining data.

Therefore, to predict the outcome of the same match, it is optimal to take 60% data for modeling.

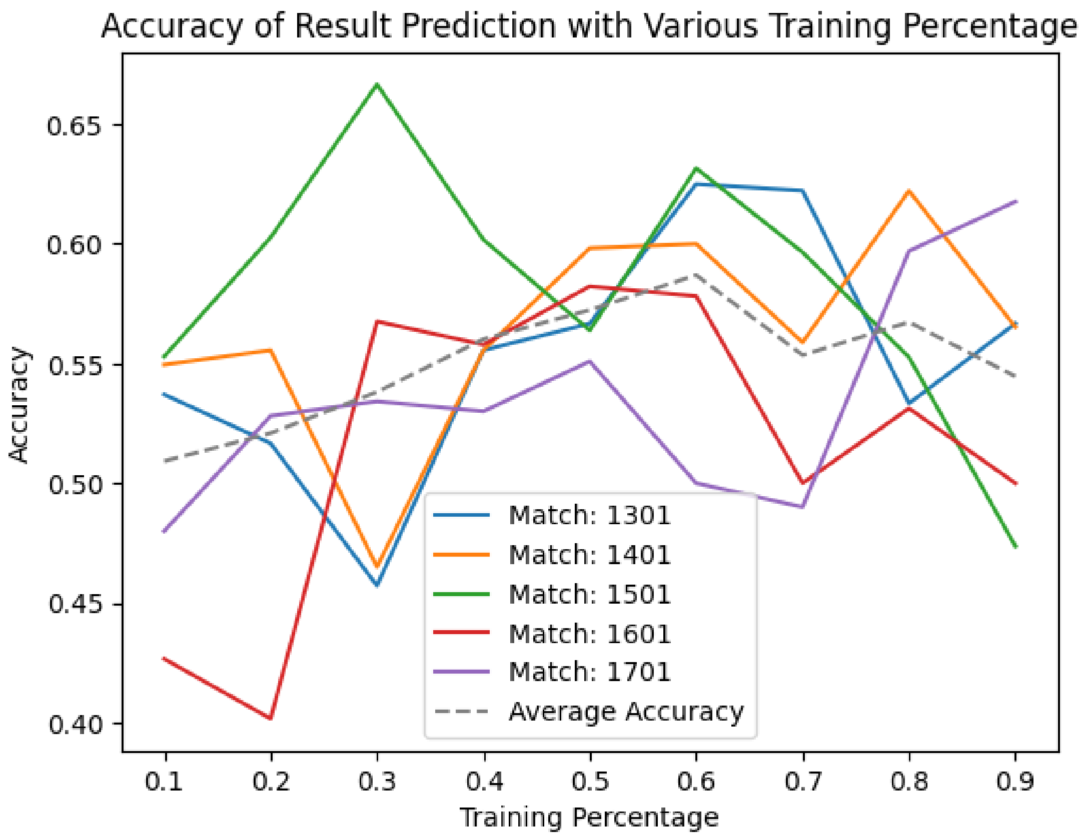

Figure 6: *Score-Based Model Accuracy for Same-Match Prediction*

5.3 Linear Regression Model (Flow Value Based)

After building the above prediction model based on a binary target variable, we shift focus to performance flow as the target variable, calculating the flow using performance flow function.

Since performance flow is a scalar value, we use linear regression to construct our model and minimize the mean squared error to determine the coefficients of independent variables.

5.3.1 Model explanation

The linear regression model is defined as:

$$h_i = \beta_0 + \sum_{j=1}^{n} \beta_j x_{ij}.$$

The notations are the same as those in the logistic regression model, except that h_i now represents the performance flow of the player at the i^{th} point.

The loss function for the above model minimizes the mean squared error between the predicted flow and actual flow, which is formulated as:

$$\min \frac{1}{m} \sum_{i=1}^{m} (y_i - \hat{y}_i)^2.$$

$\hat{y}_i$ represents the predicted performance flow of a player, while y_i represents the calculated flow.

5.3.2 Indicators

This model is also primarily built using data of player Carlos Alcaraz in the match with *match ID* “2023-wimbledon-1301”. The weights of possible indicators are listed in Table 6, rounded to the third significant digit. *Game-winning ratio*, *point-winning ratio*, and results of the preceding two points become the most influential indicators, while *elapsed time* and *difference in distance run* remain the inconsequential factors.

Table 6: *Independent Variables and Their Weights of Carlos Alcaraz in Match 1301 using Linear Regression*

Independent Variable	Coefficient
Elapsed time	0.00825
Difference in distance run	0.00344
Rally count	0.00631
Server	0.149
Game-winning ratio	4.53
Point-winning ratio	3.98
Previous point result	1.54

Point before last result	1.21
Unforced error	−0.0767
Ace	−0.261

To evaluate the stability of the indicators for this model, we compared the coefficients of this model with those of models for the other matches played by Carlos Alcaraz, and plotted similar charts as the logistic regression model comparison, as shown in Figure 7.

Point-winning ratio, *ace*, *distance run*, *rally count*, *serve*, *elapsed time*, and *unforced error* are untrustworthy indicators either due to their capricious correlation types or insignificant magnitude. *Previous point result, point before last result*, and *game-winning ratio*, on the other hand, show stable strong positive correlations with the flow and are reliable indicators of the player's performance.

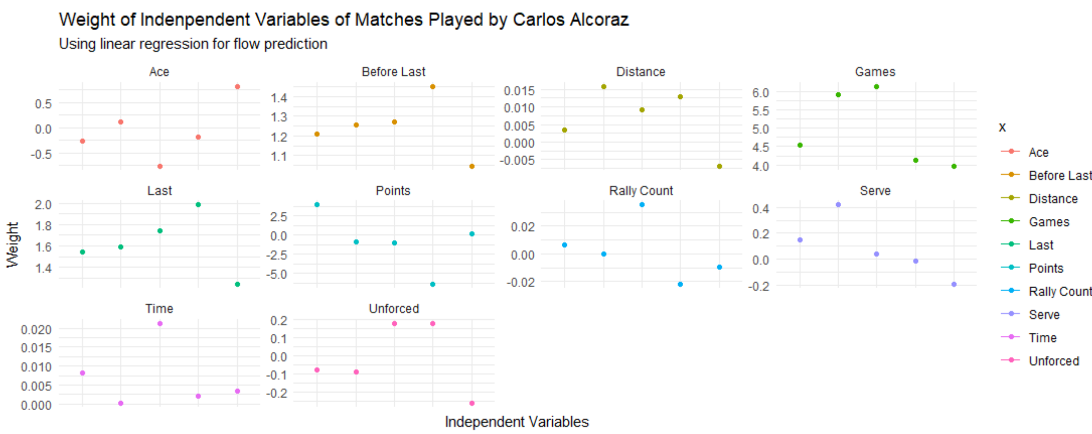

Figure 7: *Weights of Independent Variables of Matches Played by Carlos Alcaraz using Linear Regression*

5.3.3 Prediction

First, since we aim to discover the change in flow, or the instances when the flow of play changes from favoring one player to the other, we build models for each player in a match to compare the predicted momentum differences between players and identify the changes in flow.

After getting the predicted momentum of two players using the test data's independent variables, we compare the momentum values for each tennis point and recorded the player with a higher momentum. After that, we use the following two ways to test the accuracy of our models:

1. Set the player with a higher momentum the winner of the point, and compare the predicted winner with the actual winner of the test match.
2. Calculate the performance flow of the players in the test match, compare their flow, and

record the player with a higher value at each point. Then, we compare the predicted list and the calculated list of players with higher performance values.

We again use the data of match "1301" to build the flow prediction model of players Carlos Alcaraz and Nicolas Jarry for testing. The first set of test matches examines the accuracies when Alcaraz played against different players, and the results for both methods are displayed in Table 7. We also test the generalizability by picking other random matches for testing and record the results in Table 8.

In comparison to the accuracies from the logistic regression model, the accuracies of the linear regression model are significantly higher in both methods, with a minimum accuracy of approximately 65% and a maximum of 85%, which indicates that prediction based on continuous performance flow is more useful and reliable than the categorical point winning data.

Comparing the two methods' accuracies, method 2 performs at least 4% better than method 1 with one player unchanged and yields better accuracies in general. Such an outcome is plausible because the performance flow can only implicitly reflect the result of the point: if a player loses a ball after a sequence of uninterrupted wins, he would probably persist to be of higher momentum value. The relatively low accuracy of Method 1 may be on account of the inconsistency between the winner and the player with higher momentum.

Finally, the model exhibits sound generalizability since the accuracies are relatively stable regardless of the test data chosen.

Table 7: *Momentum-Based Model Accuracy of Matches Played by Carlos Alcaraz*

Match ID	1401	1501	1601	1701
Method 1 Accuracy	0.719	0.75	0.741	0.658
Method 2 Accuracy	0.799	0.809	0.785	0.781

Table 8: *Momentum-Based Model Accuracy of Randomly Chosen Matches*

Match ID	1313	1504	1402	1404
Method 1 Accuracy	0.720	0.75	0.755	0.751
Method 2 Accuracy	0.830	0.796	0.714	0.772

6. Deficiencies and Future Development

There are still many deficiencies for our models and, based on the assumptions that make this problem much simpler, there are still many factors that we didn't take into account during our modeling process.

6.1 Deficiencies

Our model still has deficiencies in the following aspects.

- For the performance flow model, we cannot use it explicitly to show that the flow is not random. Although it shows some patterns that the total random graph does not have, there is still not enough evidence to prove the flow does exist only using the performance model.
- The accuracy of predictions on points made based on our model developed with logistic regression remains relatively low. The accuracy for our model fluctuates around 50% and 60%, which cannot make clear and explicit predictions about the outcome of the matches.
- We rarely take the opponent's information as a factor in our model; that is, we consider little of the performance of the opponent player, which in reality definitely will affect the outcome of matches, into our model. We consider the difference in unforced error and distance run directly into our model, and implicitly the winning point ratio and winning game ratio as it can be calculated with the data of our players. However, other factors such as the physical conditions or technical proficiencies are not in our consideration. The lack of opponent information may lead to unstable predictions of matches.
- There are still some variables we don't consider in our models, for which we do not have enough data or they do not show high relevance with the result:
    - Break point
    - Speed of serve
    - Depth of return
    - Etc.

6.2 Future Development

For future development, we want to focus more on every athlete; that is, focus more on the analysis of athletes at individual levels, such as the mental conditions and personal skill proficiencies, and provide prediction that is much more reliable for these individuals. We also want to find more data that contains some other factors so that we can combine them to improve our model.

7. Reference

[1] Athletes Performance Academic, Tennis metrics: *https://athleticperformanceacademy.co.uk/wp-content/uploads/2022/06/Tennis-Match-KPIs.pdf*

[2] William S. Cleveland (1979) Robust Locally Weighted Regression and Smoothing Scatterplots, *Journal of the American Statistical Association*, 74:368, 829-836, DOI: 10.1080/01621459.1979.10481038

[3] Judith V. Smith (2000) Psychological Momentum in Elite Athletes, p181-216

[4] Permutt, S. (2011) The Efficacy of Momentum-Stopping Timeouts on Short-Term Performance in the National Basketball Association.

[5] Prieto-Lage I, et al. (2023) Match analysis and probability of winning a point in elite men's singles tennis. *PLOS ONE* 18(9): e0286076.

[6] Lili Zhang, et al. (2022) Improving logistic regression on the imbalanced data by a novel penalized log-likelihood function, *Journal of Applied Statistics*, 49:13, 3257-3277

Memorandum

To: Tennis athletes and coaches

Date: 02/04/2024

Subject: Tennis momentum in matches

TL; DR:

- A model of momentum can be used as a training tool to help
- Some indicators exist and can be used to predict flow in matches
- The flow or momentum can be predicted based on previous match data

As a concerning problem in tennis matches, the existence and indicators of momentum are still topics with no conclusion. Even though the pattern of momentum is not identified by recent research and studies, some mathematical models are effective in either depicting the flow with statistics from matches or predicting whether the flow will change from favoring one player to another.

The first model I would like to introduce can be used to calculate the performance flow with certain statistics from previous matches. The results we get from momentum calculations can be visualized to find patterns in specific athletes' characteristics, which can help to improve the athletes' performances in future matches. Coaches can identify players' shortcomings at different stages of a match and modify training plans for both physical and mental aspects accordingly. For example, if a player performs better at the start of a set than later in the set, they can focus their training on performance persistency throughout a long period; on the contrary, the slow starters with lower momentum at the start of the match can focus their training on getting into the zone faster.

Another model we built can be used to predict the momentum calculated by the above model and determine when the flow of a match changes from favoring one player to the other. The indicators of change in flow, such as who the server is, how many games the player has won, or the results of the preceding two points, are identified by the model, so the players can be trained to act accordingly. The prediction model can also be used to anticipate the opponents' performances in future matches by building the model based on their past performances. Consequently, athletes can study the indicators of their opponents' changes in momentum and recognize these indicators during the match. These recognitions may allow athletes to act accordingly, either play more aggressively when the opponent's momentum is about to drop or act defensively when the opponent is in a favorable condition.