Problem Chosen	2024 MCM/ICM Summary Sheet	Team Control Number
C		2429225

Unveil the Secrets of Momentum: a Match Predictor

Summary

Tennis, a dynamic sport adored by millions, unfolds with ever-changing and captivating matches, showcasing remarkable athletic prowess. To demystify the fluctuations, and the so-called "**momentum**" in tennis games, our team employs mathematical models, including model fitting, time series evaluation and principal component analysis, to understand the elusive concept of "momentum." By unraveling the dynamics of momentum, we aim to provide valuable insights for coaches and athletes, empowering them to leverage its positive influence and navigate its challenges effectively.

Before we formally started data analysis, we conducted comprehensive data preprocessing. Utilizing tools such as Excel and Matlab, we corrected logical anomalies. To ensure the absence of boundary anomalies, we employed **normality tests** and **the Shapiro-Wilk test**. Subsequently, we utilized weighted differentiation based on statistics to mitigate the impact of the serve. Following preprocessing, we proceeded to data analysis.

First, we assumed that momentum is primarily influenced by the gain and loss of each point in the game and began examining the players' scores. We **categorized scores into common scores and special scores**, each playing a distinct role in affecting momentum. We intended to quantify the impact of this point on the player by expressing it as the ratio of their performance before and after this point. This allows the initial quantification of momentum and roughly distinguish the impact of different scores.

Second, we introduced **time series forecasting** and innovatively combined it with our thoughts and model fitting, creating the **Blended Time Series Evaluation**—a method tailored to this problem. According to the coefficients of **ARIMA** forecasting equation, we allocated different weights to scores at different times. This approach eliminated the effect of time and quantified momentum, displaying greater sensitivity compared to the actual flow of the game, giving **a clear mathematical description of momentum**.

Third, we delved deeper into our research, employing **principle component analysis (PCA)** to identify the main factors from 7 special points classficated into 3 categories by how they occur. **Grey model prediction** was utilized to understand how momentum would fluctuate in the game and predict its development. We conducted **sensitivity analysis** on our model, revealing its ideal conformity with the given datasets and showcasing its robust capabilities. We also tested our model on other datasets, including matches on different types of courts, women's matches, and even other sports events. Though some problems were exposed, but on the whole, our model demonstrated universality.

In conclusion, we compiled our findings and methodologies into a comprehensive memo for coaches and athletes, summarizing our work and providing valuable insights. You can find the complete memo at the end of this article.

Keywords: time series forecasting, ARIMA, principle component analysis, grey model prediction, tennis, momentum, prediction

Contents

1 Introduction

1.1 Problem Background

20-year-old Spanish rising star Carlos Alcaraz defeated 36-year-old Novak Djokovic, winning the 2023 Wimbledon Gentlemen's Championship. The game itself was an extraordinary battle. Djokovic won the first set 6-1. However, the second set was tight, and Alcaraz won 7–6. The two then alternately led to a 2-2 draw. In the end, Alcaraz took control and secured the final game 6-4."For me, it is a dream come true to win Wimbledon. But it is even more special beating a legend like Novak," he told the dignitaries gathered at the Champions' Dinner.[1]

This epic match undoubtedly captured global attention. While marveling at the marvelous performances of the two top players, some may also question the undulating process. How could Alcaraz bounce back and triumph in the challenging second set after the total failure of the first set? And how did Djokovic adjust his tactics when the score was leveled once and twice?

Sometimes, for many points or even games, it is challenging to describe why some players perform better or worse. The incredible swings in the games that showcase the players' performances are often attributed to "**momentum**."

In sports, a player may feel motivated or "powerful" in the game, but quantifying the phenomenon is difficult. Momentum includes enthusiasm, strength, calm, courage, endurance, intelligence, prison, determination and other factors. Additionally, various events during the game can create or change "momentum." Therefore, we need to **quantify "momentum"** to better analyze the game and provide assistance to coaches and athletes. As a former article's purpose, starting from the direction of systematic theory, this paper thinks that tennis service is not only a simple skill, but it should be a complete system, that is, consists of skill, tactics, mentality and stamina in order to renew the conception of tennis service, deepen the recognition of it, found a theoretical system, and guide the practice of it.[9] If the momentum can be quantified, it will be able to directly **evaluate and predict** the direction and outcome of the whole game, thus allowing the quantified to **help the athletes and coaches** to win the game.

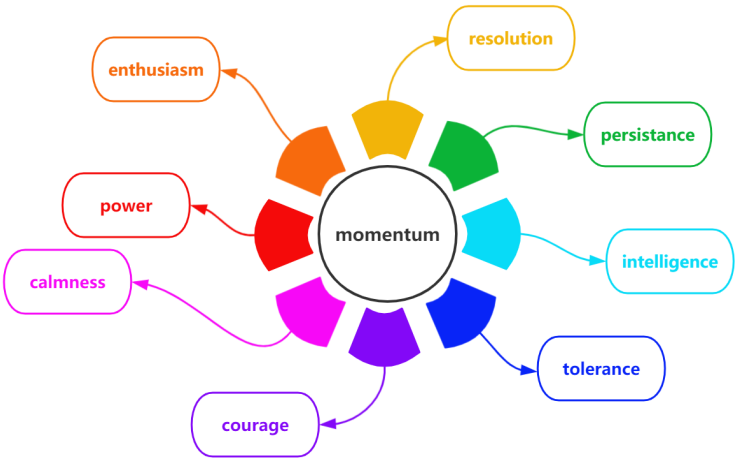


Figure 1: background

1.2 Restatement of the Problem

Considering the background information and the limitations identified in the questions, in order to better provide coaches and athletes with quantitative indicators of "momentum" and practical solutions to answer their doubts, we need to solve the following three questions:

• Problem 1:

Develop a model to **describe** the competition process and various events visually, so as to reflect the "momentum" of the players and the status of the guaranteed break. At the same time, we need to use this model to **evaluate** the coaches' assumption that a player's fluctuations and movement are random.

• Problem 2:

Build a model to **predict** these ups and downs in the competition. Find out the most **relevant factors** and give advice to the players.

• Problem 3:

Test the developed model on other matches. Analyze the prediction effect and reflect, while considering the **universality** of the model.

1.3 Our Work

1. The primary focus of our study is on the momentum in tennis games. Momentum serves as a comprehensive evaluation of players' mentality and competitive state during the match. It quantifies the direction of the competition. Here, we assume that momentum is primarily influenced by **the gain and loss of each point** in the game, prompting us to first examine the players' scores.

2. We categorized scores into common scores and special scores. Regardless of the scoring form, it acts as a morale boost. We designate scores specifically **boosting** morale as special scores. Subsequently, we employed data analysis to quantify the representation of the player's 'momentum' for each score.

3. Our aim is to determine whether performance changes before and after these scores. Evaluating how well players score in the game is the most direct way to assess performance. Acknowledging that the server holds more authority over the current point, we standardized the serve and reception scores before proceeding. We calculated players' average scoring probability when serving for the point in all matches, equally weighting the serve and reception points based on our statistical data. This approach allows for a more intuitive understanding of the players' actual scores.

4. The game unfolds continuously in the timeline, and as time progresses, earlier scores have diminishing impacts on later games. When this impact reduces to a certain extent, it can be ignored. To reveal the difference between performance before and after scoring points, we analyzed players' performance within a specific time before and after centering on this point. Consequently, we transformed the innovative prediction method of **time series analysis** into an evaluation method, fitting the difference curve to quantify the concept of 'momentum.' Further details on our methodology will be presented in the following article.

5. After calculating the momentum based on the formula, we proceed to make predictions using our model. To validate the coaches' claims, we thoroughly examined Wimbledon's official website and

generated a word cloud to identify **the critical elements** in tennis games.

6. To identify key indicators in competitions, we initiated our analysis by examining the fundamental unit of the competitive process. Subsequently, we opted for **principal component analysis** to determine the most relevant factors. Detailed results will be presented in the following sections.

7. **Grey forecasting** is later adopted with the aim of establishing a prediction model. We specifically focus on the momentum flow and make predictions about the match's trend based on data from the ongoing match. Given the incompleteness, trendiness, and changeability of the data, this approach allows us to best ensure the stability and accuracy of our model.

8. Finally, we conducted tests to assess the goodness of fit and generalization of our model. We compared the "momentum" generated by our algorithm with the actual game data to determine whether it accurately reflects the competition's trend. We also collected data from other court surface types, and even from **other sports** like table tennis, fitting the data into our model to assess its generalization capabilities.

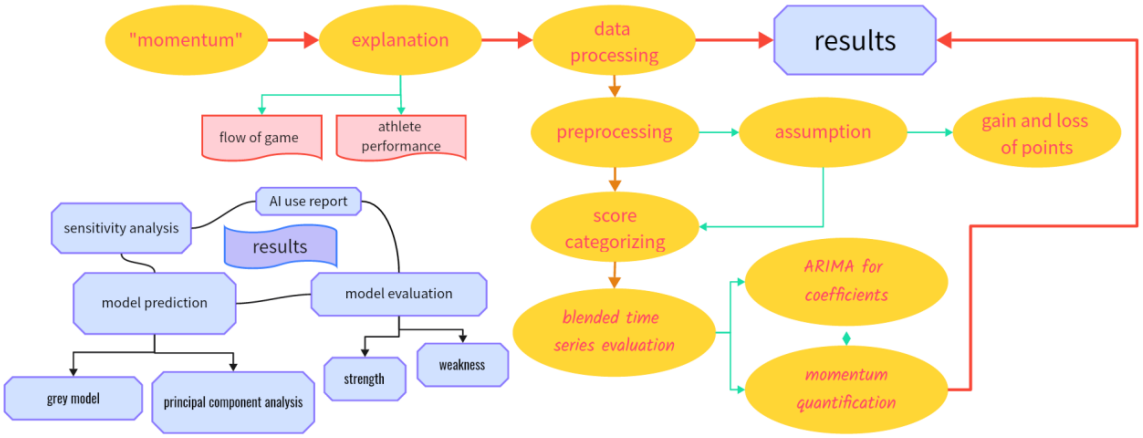


Figure 2: Flow Chart

2 Assumptions and Notation

2.1 Assumptions

1. We assume that momentum primarily changes through the gain and loss of each point in the game. This choice is made because other factors, such as audience reactions, players' own strength, and psychological qualities, are not considered due to the difficulty in evaluation, lack of data support, and their limited impact on the overall results (mostly random and having little correlation with the average data).

2. We assume that **special scores** include the following: Innings, inventory, break point, love game(winning a game without the opponent scoring a point), fault, unf, long rally(10 and above). Scores that do not fall under any special category are referred to as ordinary scores.

3. We assume that both players begin the match with a momentum of 10, and we further assume that the momentum **only increases** (the momentum is relative in doubles competition, so a decrease in one player's momentum is equivalent to an increase in the other player's momentum). We also assume that regular point scored won't increase momentum largely(later data analysis may prove this), while special scores will contribute to momentum increase in a multiplicative manner.Different special scores can occur at the same score (e.g., a long rally), and winning a point with a special score is likely to have a greater impact on momentum than the first point won after a prolonged rally. Therefore, we employ a multiplication rate for simulation purposes.

2.2 Notation

Notations	Definition
ρ	the average probability of the tee scoring across all matches is ρ
ξ_n	the rate ξ of the n^{th} special score
φ_{nm}	time series quantification value before and after ξ_n (m before 1, after 2)
$\psi_i(t)$	momentum ψ of contestant i at time t
a_{ij}	The i^{th} index of the j^{th} evaluation object(scoring key points)
b_j	Information contribution rate of the j^{th} principal components
ζ_n	Composite score of the n^{th} principal components

Table 1: Notations Table

3 Model Design

Within sport , momentum is seen as a bi - directional construct , whereby successfully performed events increase the probability of subsequently performed events being successful and unsuccessful events decrease the probability of subsequent events being performed successfully. This creates the notion of positive and negative momentum. Players , coaches and spectators all consider psychological momentum as a determinant of success.[2] We need to examine the momentum of the simulated tennis game and quantify it. As mentioned earlier, momentum can be viewed as **a comprehensive assessment** of players' mentality and competitive state during the match, providing a quantitative evaluation of the direction of the competition. According to our hypothesis, momentum primarily fluctuates based on the gains and losses of each point in the game. To address this, we innovatively transformed the prediction method of **time series analysis** into an evaluation approach, fitting the difference curve to quantify the concept of 'momentum.' The report outlines the modifications made to achieve this.

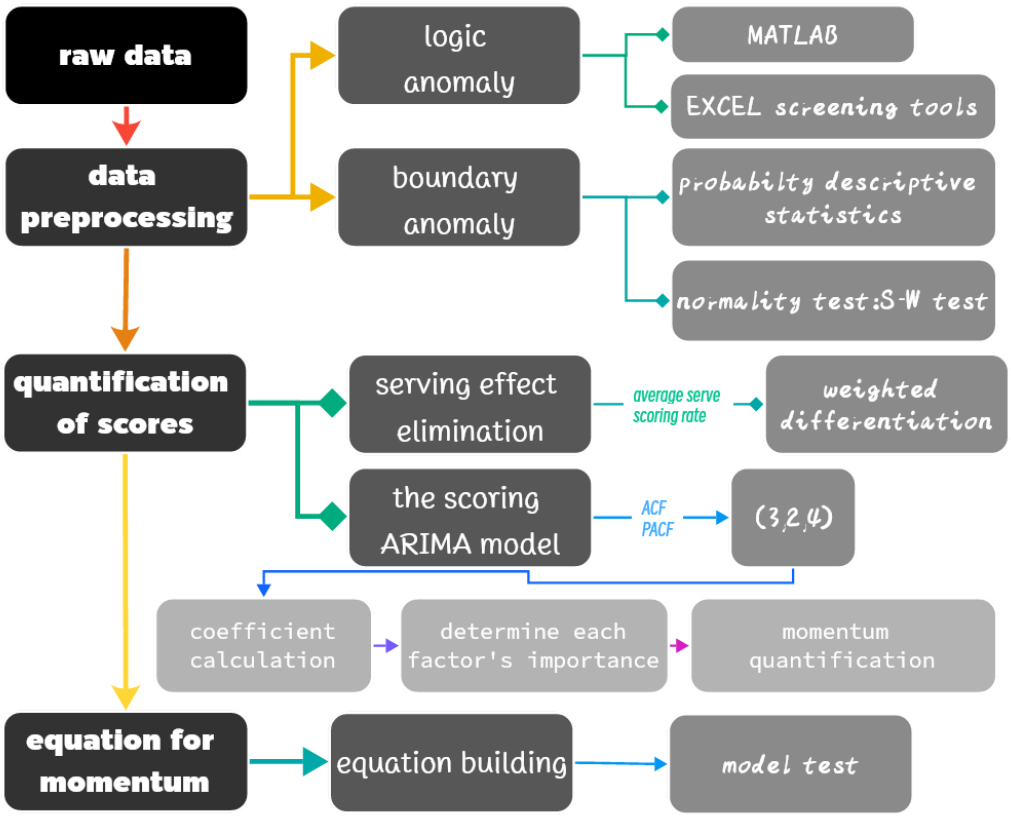


Figure 3: Model Design Flow Chart

3.1 Data Preprocessing

3.1.1 Logical Anomaly

We first preprocess the given data, and after simple logical anomaly processing through the EXCEL screening tool, we determine a set of **logical anomaly** criteria and batch them with MATLAB. The data set contains all the 2023 Wimbledon information. Each row consists of the situation of the game, scoring method, scoring elements, etc. We manually correct the data and use them wisely.(Table 2)

Table 2: Logical Anomaly

Rows	How Many	Problems
2188-2674	two matches	missing rally-count and serve speed data
492-502,1328-1335,5500-5512 $\cdots$	many	the decider changed the statistical method

3.1.2 Boundary Anomaly

Next, we determine the **boundary anomaly**.

In the model building stage, we need to get the average probability of the server scoring in all matches as ρ, because in tennis, the server is often more dominant, so we have to match the serve with

the reception score, so as to get more objective and accurate data. To ensure that the average probability of the resulting server score is objectivity ρ, we determine boundary anomalies for ρ for all matches.

In 31 matches of more than 7000 serves, we plot **the average probability of all serve scores** as ρ (Figure 3) and conducted data analysis (Table 3).

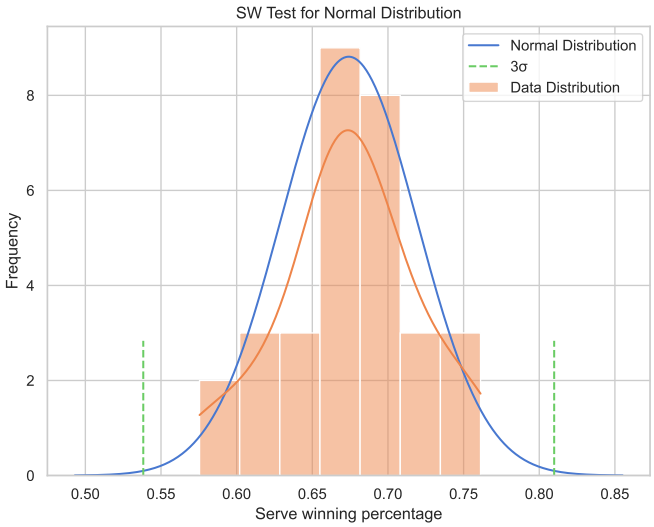


Figure 4: S-W test

Table 3: Overall description of the results

sample capacity	median	average value	standard deviation
31	0.674	0.674	0.046
skewness	kurtosis	S-W test	K-S test
-0.175	0.111	0.974(0.636)	0.107(0.834)

The above table shows the results of probability descriptive statistics and normality test, including median, mean, etc. Next, we tested the normality of the data.

Generally, there are two test methods for normal distribution. One is Shapiro-Wilk test, which is suitable for small sample data ($\leqslant$ 5000); the other is Kolmogorov-Smirnov test, which is suitable for large sample data ($>$ 5000).

If significant (P $\leqslant$0.05), the null hypothesis is rejected (the data meets the normal distribution), the data does not meet the normal distribution, and otherwise the data meets the normal distribution.

Generally, it is difficult to meet the test under realistic research. If the absolute value of sample kurtosis is less than 10 and the absolute value of skewness is less than 3, the normal distribution histogram, PP map or QQ graph can be described as basically consistent with the normal distribution.

According to the test method, probability sample $N \leqslant 5000$, **S-W test** is used, the significance$P$ value is 0.636, the level is not significant, and the null hypothesis cannot be rejected, so the data meet the normal distribution.

Normal distributed probability density function image is symmetric, because can assume that the population follows the mean μ, standard deviation σ normal distribution, then draw a sample point

from the population, the probability on the interval [μ-3σ, μ + 3σ] is about 99.73%, and beyond the possibility of the range is less than 0.3%, is a typical small probability event, so the data beyond the range can be considered as outliers.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{1}$$

Therefore, we excluded outliers with the 3 σ principle, combined with Figure 2, we determined **no boundary outliers**, calculating the average probability of server scoring in all matches as ρ is 0.674.

3.2 Quantification of scores

For the current momentum of the two players ψ_1 and ψ_2, we can predict who they will have the upper hand in the next competition. As described above, we will analyze changes in scores to reflect changes in momentum. So we have to **quantify the score** first.

3.2.1 Eliminate the Effect of Serving

In data preprocessing, we finally obtained an average serve score rate of *0.674* across all matches. So we multiply the server's score *(1-0.674)* and the receiver's score by *0.674*. The data thus obtained is objective. We used the data of the 2023 Wimbledon Gentlemen's Final to represent the whole. This game is more representative, with ups and downs, and the change of potential has a big impact. The two players are equal. The line diagram before the same is shown in Figure 5 left, and the line diagram after the same is shown in Figure 5 right.

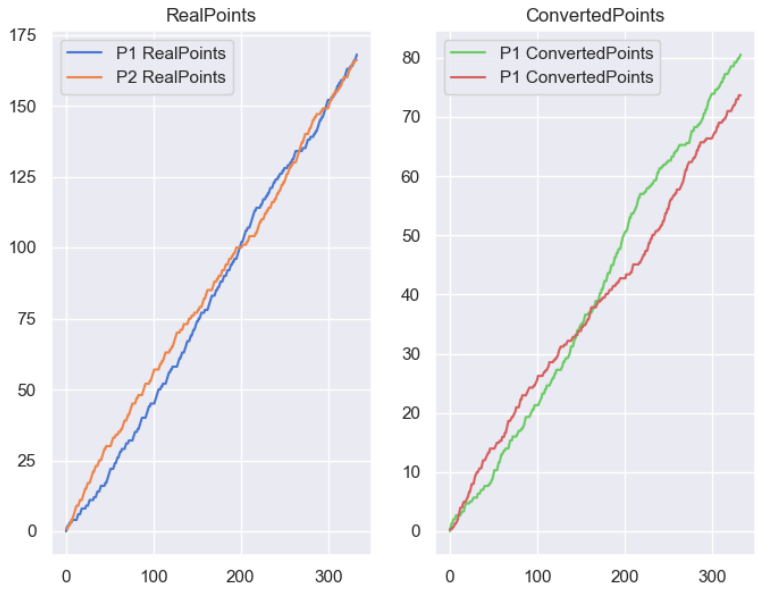


Figure 5: 2023 Wimbledon Gentlemen's Final

Obviously, they are **non-stationary sequences**. For non-stationary sequences, due to its mean and variance instability, we deal with this is to transform it to stationary sequences, so that applying analytical methods on stationary time series, such as building ARMA models to conduct the corresponding study. If a time series is stationary after a differential operation, the sequence is a differential stationary sequence and can be analyzed using the ARIMA model.[3]

There are two methods to test the stability of the sequence, one is the graph test based on the characteristics of timing diagram and autocorrelation graph. The method is simple and widely used, and the disadvantage is subjectivity; the other is the test statistics, and the most common method is the unit root test.

If a sequence is a **pure random sequence**, then there should be no relationship between its sequence values, that is, satisfying $\gamma(k) = 0,\ k \neq 0$ this is an ideal state that can only appear theoretically. In fact, the sample autocorrelation coefficient of a pure random sequence is not absolutely zero, but very close to zero, and fluctuates randomly around zero.

Pure random test, also known as white noise test, is generally constructed test statistic to test the sequence of pure randomness, commonly used Q statistic , LB statistics, by sample delay period autocorrelation coefficient can be calculated, and then calculate the corresponding p value, if the p value is significantly greater than the significance level α, said the sequence cannot reject the pure random hypothesis, can stop the analysis of the sequence.

Next, we use the time series model. We first do the stability test and randomness test of the sequence (see Table 4), and then determine the values of p, d, q from the acf and $pacf$ images of the players' score sequence.

Table 4: DickeyFuller Test

d	ADF Statistic	p-value	1%	5%	10%
1	0.15901415	0.96986683	-3.45014107	-2.87025885	-2.57141515
2	-18.553702	2.09107049	-3.45020115	-2.87028523	-2.57142922

3.2.2 The Scoring ARIMA Model

After we have made the images of acf and $pacf$, we have determined the values of p and q. Meanwhile, we analyzed the size of aic and bic, considering the fitting case and avoiding the overfitting case comprehensively, determining that ***p*=3, *d*=2, *q*=4, using the ARIMA model**. (Table 5) Difference operation has the powerful ability to extract deterministic information, many non-stationary sequences will show the properties of stationary sequence when the non-stationary sequence is called differential stationary sequence. The differential stationary sequences can be fitted using the ARMA model. ARIMA The essence of the model is the combination of the difference operation and the ARMA model. The model with the following structure is called the autoregressive moving average model, which is abbreviated as the ARMA *(p, q)*.

Table 5: ARIMA

name	3,2,4	3,1,4
aic	78.998	233.865
bic	109.439	264.330

The ARIMA model's interpretable parameters have effectively facilitated our objective of simulating increasing momentum with time decay through ARIMA forecasting. The noticeable decrease in parameters aptly reflects this trend. ARIMA's favorability can also be attributed to its modest requirements for time series data, demonstrating a commendable tolerance to fluctuations and noise. Considering the swift changes in tennis field situations, ARIMA's exceptional anti-interference capability aligns seamlessly with our requirements. Despite its primary function as a forecasting model, we have **creatively repurposed** it as an evaluation model, integrating it into our operational framework.

$$x_t = \phi_0 + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-p} \tag{2}$$

That is, the value x_t of the random variable X_t at time t, is the multivariate linear function of pre-p x_{t-1}, x_{t-2}, $\cdots$, x_{t-p} and pre-q $\varepsilon_{t-1}, \varepsilon_{t-2}, \cdots, \varepsilon_{t-q}$, and the error term is the random interference ε_t of the current period, which is the zero mean white noise sequence. It is considered that x_t is mainly influenced by the sequence values in the past p-period and the error terms in the past q-period.

In particular, when q =0, it is AR (p) model; when p =0, it is MA (q) model.

By the above method and the differential reduction, we obtain the formula for ARIMA.

$$\varphi_{nm} = c_1 + e_1 y_{t-1} + e_2 y_{t-2} + e_3 y_{t-3} + f_1 z_{t-1} + f_2 z_{t-2} + f_3 z_{t-3} + f_4 z_{t-4} + z_t \tag{3}$$

And calculate the coefficient. (as Table 6)

Table 6: coefficient

coefficient	1	2	3	4
e	-0.912313	-0.641710	-0.306325	
f	-0.076221	-0.323578	-0.275385	-0.261056

We use this formula to find the various coefficient to obtain the prediction equation φ_1 as the quantified value of the performance before this point. Similarly, φ_2 was used as the quantified value of the performance after this point. You can get a multiplier, ξ.

$$\xi = \frac{\varphi_1}{\varphi_2} \tag{4}$$

By the above, we calculate ξ for **all special scoring points**. And calculate the ξ of all ordinary scoring points as its expected ξ_0. All of the ξ are shown in Table 7.

Table 7: multiplier

$\xi_0(normal)$	$\xi_1(innings)$	$\xi_2(inventory)$	$\xi_3(breakpoint)$
1.02010015	1.02453810	1.02226810	1.02581506
$\xi_4(lovegame)$	$\xi_5(fault)$	$\xi_6(unf)$	$\xi_7(longrally)$
1.01373430	1.03339993	1.03321996	1.02697614

3.3 The Fitted Equation for the Momentum

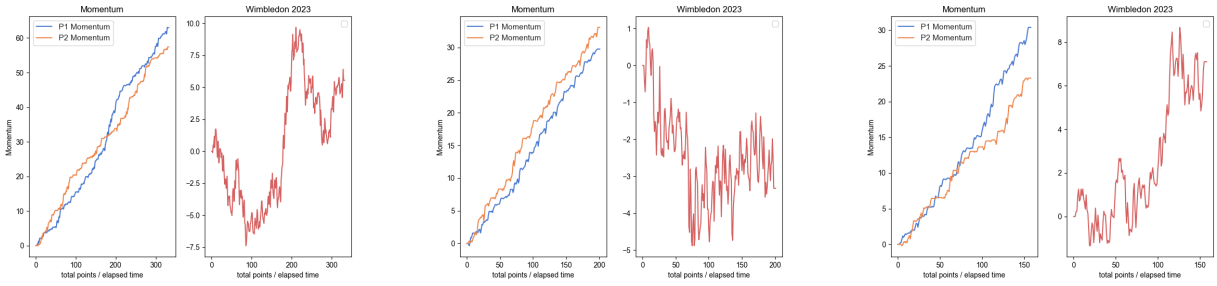
3.3.1 Model Building

After all this work, we can start to assess the momentum of the players. First, similarly, the momentum is also affected by time. This effect has been considered in the process of **plofication**. We just have to consider the time before. Similarly, according to the coefficient already obtained, we can get the current momentum of the players.

$$\psi_i(t) = 10(\prod_{m=0}^{7} \xi_{n_m})^{e_1}(\prod_{j=0}^{7} \xi_{n_j})^{e_2}(\prod_{k=0}^{7} \xi_{n_k})^{e_3} \tag{5}$$

3.3.2 Test of the Model

We brought the calculated coefficients into this formula and plotted them as a fitted curve, as shown in Figure 6.On the left is 2023 Wimbledon Gentlemen's Final, in the middle and on the right are 2023 Wimbledon Gentlemen's Semi-Final.



(a) ModelF (b) ModelSF1 (c) ModelSF2

Figure 6: Fitting Curve

4 Model Prediction

4.1 Coaches' Claim

We can understand the doubts of the tennis coach, he looks at this problem from the perspective of cultivating athletes. After all, the most important thing before the game is to adjust the mentality of the athletes, rather than too much technical and tactical guidance. Making assumptions like this can convince the athletes of their strength without overly straining the details. However, as a coach, we should be concerned about **the relevance of these elements**, and practice them subtly in the guidance. Therefore, we use **principal component analysis** to rank the importance of these technical points to give the coaches a better reference.

We first consulted a large amount of literature on tennis skills and tactics on the official website of Wimbledon, and drew them into a word cloud to find out several important elements, as shown in

Figure 7.[4]

Figure 7: Cloud

As can be seen from **the word cloud map**, in addition to ball, tennis and other key words that must appear. The biggest is the words tactic, evaluation, technical, opponent. And these words are the embodiment of tennis skills and tactics. We find some previous conclusion:1) hair extensions technology is a core part of tennis technology now is the key factor to determine the outcome of the game.2) Now women's tennis should pay attention to efficiency.3)Reducing unforced errors is the main way to improve the winning rate of Women's Tennis match,And put forward the corresponding countermeasures and suggestions,The purpose of this paper is to explore and excavate the quantitative index of winning factors in tennis competition. [8] Therefore, combine with these conclusions, although the hypothesis made by the coach is reasonable, but in front of the big data, **the skills and tactics** are still the magic weapon for the players to win. We will then look at what of the key elements.

4.2 Key Indicators

In past literature[7], Daniel Medwade's winner has been studied. The stage characteristics of the technical process and the application characteristics of technical and tactics are conducive to the construction of the index system of the technical and tactical efficiency evaluation of tennis project.[6]

As can be seen from Figure 7, the opponent is the core, and its various characteristics are the important embodiment of the elements of technical and tactical tactics. But in the end, which one is the real key factor, which needs more data to support, rather than qualitative judgment. So we used the principal component analysis method to do this estimate.

Principal Component Analysis (PCA) stands out as a potent technique for dimensionality reduction and feature extraction. Even after meticulous processing, our data retains certain complexities and intricacies, prompting the need for an effective integration tool. Consequently, our attention turned to PCA, a method adept at significantly reducing data complexity and expediting the analysis process. Notably, PCA also offers optimal visualization and interpretability, enhancing the clarity and practical applicability of our results.

In the ARIMA fitting in the previous section, we know that the fitting is good. At the same time, the results of **big data** tell us some key points and special scoring elements are the key to determine **the trend of the game**, which can provide better help for coaches and athletes. So we used a more objective visual Principal Component Analysis to handle it.[5]

First, we took as many n games as possible, and then calculated the ξ_n-values of the special scores in each match separately to form a matrix.

$$\begin{pmatrix} \xi_{11} & \xi_{12} & \cdots & \xi_{1n} \\ \xi_{21} & \xi_{22} & \cdots & \xi_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \xi_{m1} & \xi_{m2} & \cdots & \xi_{mn} \end{pmatrix} \tag{6}$$

Second, we normalized the raw data. We collect different index units, for example, the service speed unit is km/h, the serve error unit is second, the unit is different; the value of the data collected varies greatly, for example, the serve speed is three digits, while love game is a single digit. Therefore, in order to eliminate the magnitude difference between the indicators, the collected data. There are *m (m=3)* index variables that we conducted the principal component analysis, respectively they are $x_1, x_2, \cdots, x_m$. There are n evaluation objects, and the value of the j index of the i evaluation object is a_{ij}. Each index value a_{ij} is converted into a standardized index value $\tilde{a}_{ij}$. We have

$$\tilde{a}_{ij} = \frac{a_{ij} - \mu_j}{s_j}, i = 1, 2, \cdots, n; j = 1, 2, \cdots, m, \tag{7}$$

in this equation: $\mu_j = \frac{1}{n}\sum_{i=1}^{n} a_{ij}; s_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(a_{ij} - \mu_j)^2}, j = 1, 2, \cdots, m$, so as $\mu_j, s_j$are the sample mean and sample standard deviation of the j^{th} indicator.

Third, the correlation coefficient matrix R is calculated to calculate the correlation coefficient collected between the indicators and construct the correlation coefficient matrix$R = (r_{ij})_{m\times n}$, we have

$$r_{ij} = \frac{\sum_{k=1}^{n} \tilde{a}_{ki} \cdot \tilde{a}_{kj}}{n-1}, i, j = 1, 2, \cdots, m, \tag{8}$$

in this equation: $r_{ii} = 1, r_{ij} = r_{ji}, r_{ij}$is the correlation coefficient of the i^{th} index and the j^{th} index.

Fourth, we calculate the eigenvalue eigenvector: the relational data matrix R obtained in the third step is used to calculate the eigenvalue $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m \geq 0$ of the matrix. m is the

total number of score-related indicators collected and the corresponding eigenvector$u_1, u_2, \cdots, u_m$, where$u_j = [u_{1j}, u_{2j}, \cdots, u_{mj}]^T$, constitutes m new indicator variables composed of the eigenvector:

$$\begin{aligned} y_1 &= u_{11}\tilde{x}_1 + u_{21}\tilde{x}_2 + \cdots + u_{m1}\tilde{x}_m, \\ y_2 &= u_{12}\tilde{x}_1 + u_{22}\tilde{x}_2 + \cdots + u_{m2}\tilde{x}_m, \\ &\vdots \\ y_m &= u_{1m}\tilde{x}_1 + u_{2m}\tilde{x}_2 + \cdots + u_{mm}\tilde{x}_m, \end{aligned} \tag{9}$$

in this equation:y_1 is the first principal component, y_2 is the second principal component,$\cdots$, y_m is the m^{th} third principal component.

Fifth, select principal components: the contribution rate and cumulative contribution rate are calculated according to the eigenvalue calculated in the fourth step, and the principal components are selected to discard the principal components with low contribution rate. The first p most important eigenvectors were selected as the basis vectors of the new coordinate system to constitute the new principal components. The formula for calculating the contribution rate is:

$$b_j = \frac{\lambda_j}{\sum_{k=1}^{m} \lambda_k}, j = 1, 2, \cdots, m \tag{10}$$

The calculation formula of the cumulative contribution rate is:

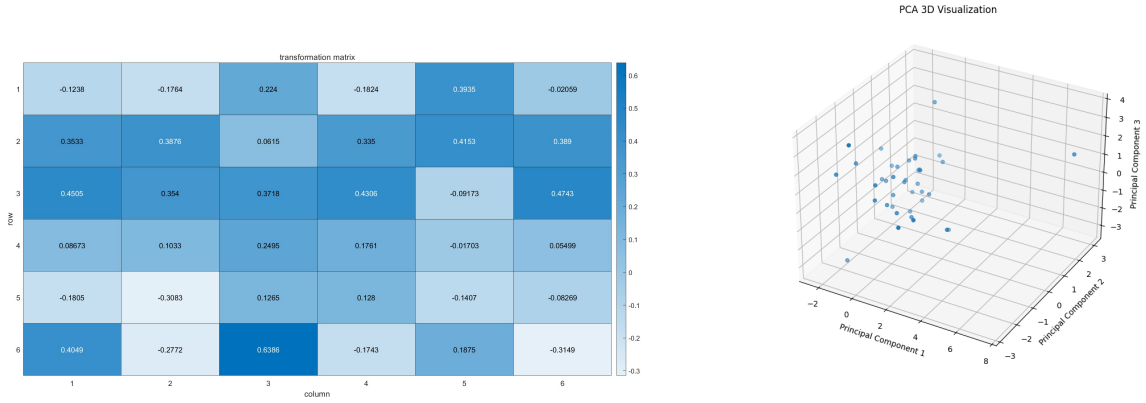
$$\alpha_P = \frac{\sum_{k=1}^{P} \lambda_k}{\sum_{k=1}^{m} \lambda_k} \tag{11}$$

The variance contribution rate of each variable is calculated as:0.44633569, 0.35338399, 0.14438617.

Sixth, calculate the comprehensive score: each sample collected to the original data set of meteorological conditions is projected on the new coordinate system, and the contribution rate of b_j as the principal component is calculated as the weight, so as to obtain the score of the principal component of the sample. The principal component score formula is shown as follows:

$$\zeta = \sum_{j=1}^{P} b_j y_j \tag{12}$$

We obtained the transformation matrix W of the sought variables treated with Python and drew the heat map. Eventually, we drew it in 3D:



(a) matrix W (b) PCA 3D

Figure 8: PCA

4.3 Prediction

We next want to **predict the change in momentum**, and we chose to use **the grey prediction model**. The main feature of grey prediction is that the model uses not the raw data sequence but the generated data sequence. The core system is the gray model (Grey Model, GM), that is, the cumulative generation (or other methods generation) of the original data to get the approximate exponential law and then model the method.

We predict with the GM (1,1) model, which is a first-order differential equation and contains only 1 variable.

First, to ensure the feasibility of the modeling method, we make necessary tests on the known data columns. Data are $x^{(0)} = (x^{(0)}(1), x^{(0)}(2), \cdots, x^{(0)}(n))$, and the rank ratio of the sequence is calculated

$$\lambda(k) = \frac{x^{(0)}(k-1)}{x^{(0)}(k)}, k = 2, 3, \cdots, n \tag{13}$$

Since not all the level ratios $\lambda(k)$ fall within the tolerable coverage $\Theta = (e^{-\frac{2}{n+1}}, e^{\frac{2}{n+2}})$, we do the necessary transformation processing to the sequence$x^{(0)}$ to make it less accessible. That is, take the appropriate constant c and make the translation transformation

$$y^{(0)}(k) = x^{(0)}(k) + c, k = 1, 2, \cdots, n \tag{14}$$

The level ratio of the sequence $y^{(0)} = (y^{(0)}(1), y^{(0)}(2), \cdots, y^{(0)}(n))$

$$\lambda_y(k) = \frac{y^{(0)}(k-1)}{y^{(0)}(k)} \in \Theta, k = 2, 3, \cdots, n \tag{15}$$

Second, data column $x^{(0)} = (x^{(0)}(1), x^{(0)}(2), \cdots, x^{(0)}(n))$, 1 cumulative generation order column (1-AGO)

$$\begin{aligned} x^{(1)} &= (x^{(1)}(1), x^{(1)}(2), \cdots, x^{(1)}(n)) \\ &= (x^{(0)}(1), x^{(0)}(1) + x^{(0)}(2), \cdots, x^{(0)}(1) + \cdots + x^{(0)}(n)) \end{aligned} \tag{16}$$

in this equation:$x^{(1)}(k) = \sum_{i=1}^{k} x^{(0)}(i), k = 1, 2, \cdots, n$.Mean generating sequence of$x^{(1)}$

$$z^{(1)} = (z^{(1)}(1), z^{(1)}(2), \cdots, z^{(1)}(n)) \tag{17}$$

in this equation:$z^{(1)}(k) = 0.5x^{(1)}(k) + 0.5x^{(1)}(k-1), k = 2, 3, \cdots, n$.

The corresponding albino differential equation is given for:

$$\frac{dx^{(1)}}{dt} + ax^{(1)}(t) = b \tag{18}$$

According to the model, the predicted value can be obtained

$$\hat{x}^{(1)}(k+1) = \left(x^{(0)}(1) - \frac{\hat{b}}{\hat{a}}\right) e^{-\hat{a}k} + \frac{\hat{b}}{\hat{a}}, k = 0, 1, \cdots, n. \tag{19}$$

and $\hat{x}^{(0)}(k+1) = \hat{x}^{(1)}(k+1) - \hat{x}^{(1)}(k), k = 0, 1, \cdots, n-1, \cdots$.

After the calculation, we get the data(Table 8):

Table 8: Gray Prediction

Developmental Quotient	Gray action	The posterior difference ratio
-0.002	371.504	0.062

Chart description:

The above table shows the development coefficient, the gray action amount, and the posterior difference ratio. The grey prediction model can be constructed by the development coefficient and the gray action model. The development coefficient indicates the development law and trend of the series, and the gray action amount reflects the changing relationship of the series.

The difference ratio of the posterior can verify the accuracy of the gray prediction, and the smaller the posterior difference ratio is, the higher the accuracy of the gray prediction is. Generally, if the posterior difference ratio C value is less than 0.35, the model accuracy is high, the C value is less than 0.5 means that the model accuracy is qualified, and the C value is less than 0.65 means that the model accuracy is basically qualified; if the C value is greater than 0.65, the model accuracy is unqualified. Judging from the above table, the posterior difference ratio is 0.062 and the model has high accuracy.

5 Model Application

5.1 Fitting Accuracy

We looked for data on the Internet, found more data sets from Wimbledon 2023, and found the data in the momentum column. We draw a chart of the momentum of the 2023 Wimbledon Gentlemen's final. Since the momentum we fit the match was two times, we recovered the difference and compared the difference to the momentum we fit.The left is our predicted model, the right is the real line.

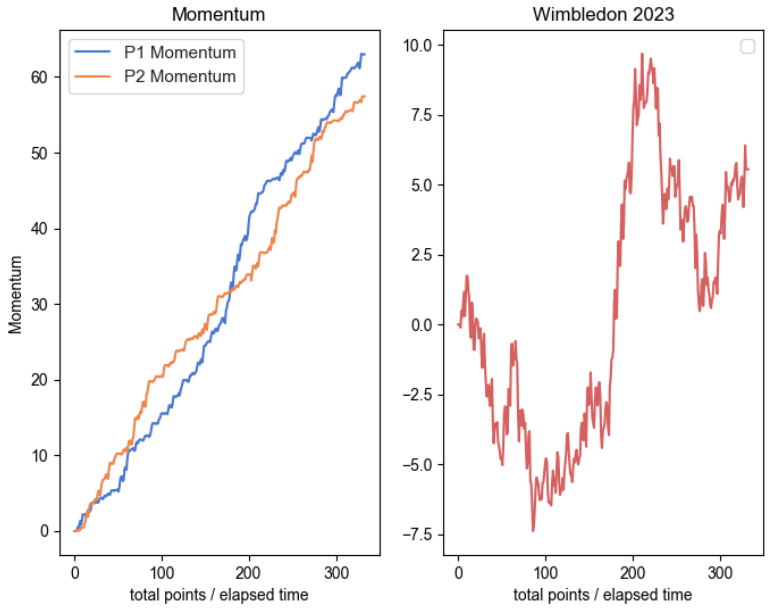
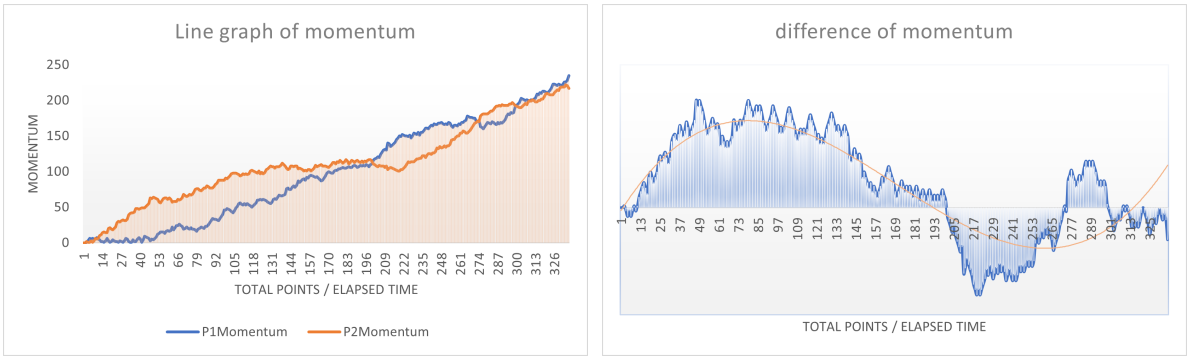


Figure 9: 2023 Wimbledon Gentlemen's Final Model



(a) Line Graph of Momentum (b) Difference of Momentum

Figure 10: 2023 Wimbledon Gentlemen's Final Real

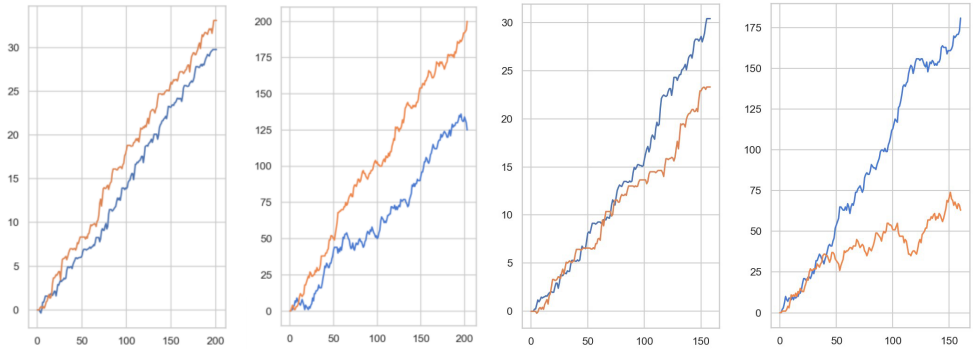
Comparing the two graphs shows that the model we predict roughly conforms to the real situation. But momentum itself is **a subjective concept**, there is no objective quantitative standard. We define

momentum as increasing, less momentum in real data as losing points. This creates a certain amount of error. But momentum itself is a relative concept, and if we compare the momentum between the two, we may get **more objective** results, see Figure 10.

We comparing the two charts:

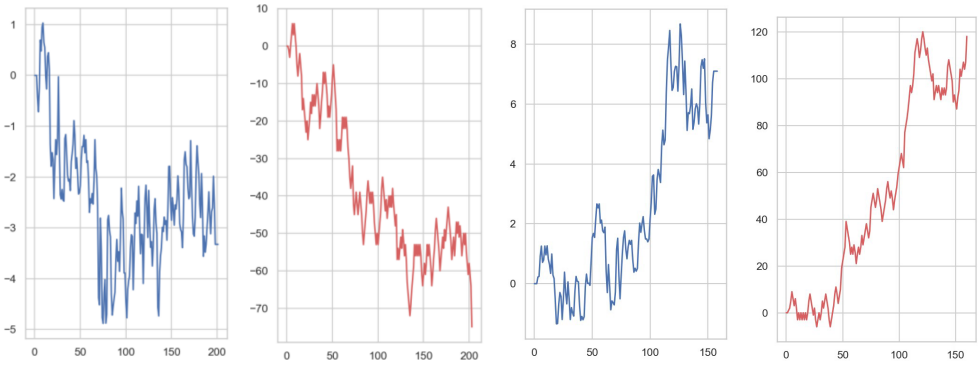
- Their changing trends are roughly the same. Volumes and fluctuations are basically consistent with fluctuations, which shows that our modeling is still successful.
- The ordinate of the two is different in the setting, so there will show certain differences, but this does not affect our comparison, as long as they are standardized.
- Both of them show a trough, then peak and trough, and always fluctuating trend. This shows the sensitivity and stability of our model.
- Ignoring the intermediate process and looking only at the initial and end values, our model is almost exactly the same as the true value (after normalization). This shows that although there will be some twists and turns and errors in the process, the prediction of the final result is still very accurate. Therefore, in a competitive sport, our model has a large application market.

Although our model in this picture has a very good application, but a picture, a game can not have a very high universality. So we selected two other games for **drawing and comparative analysis**. See the following figures.



(a) Model2 (b) Real2 (c) Model3 (d) Real3

Figure 11: 2023 Wimbledon momentum of each player



(a) Model2 (b) Real2 (c) Model3 (d) Real3

Figure 12: 2023 Wimbledon momentum difference of each player

Let's analyze the above **eight charts**: the four charts in the first row are always very different between the real one and the model. Upon careful observation, we will find that the difference is that the curve of the real inferior player is often much lower than that of the dominant player. This is because the momentum built up by our model has only increased without decreased. Let's also focus on the two sets of graphs in line 2. In these two sets of Fig. The model of pair 3 clearly fits better. All present the first trough and then peak and then wave peak and then trough situation. The relationship between the winners is clearly correct and with little difference. However, the model in pair 2 produced a large difference. Although the final result is the same, the difference of momentum is negative, the relationship between the players is correct. But the difference is quite large. Our model is generally in a trend of first falling and then rising. The real momentum of the data has been falling. This requires a deeper reflection.

We have summarized the following reasons:

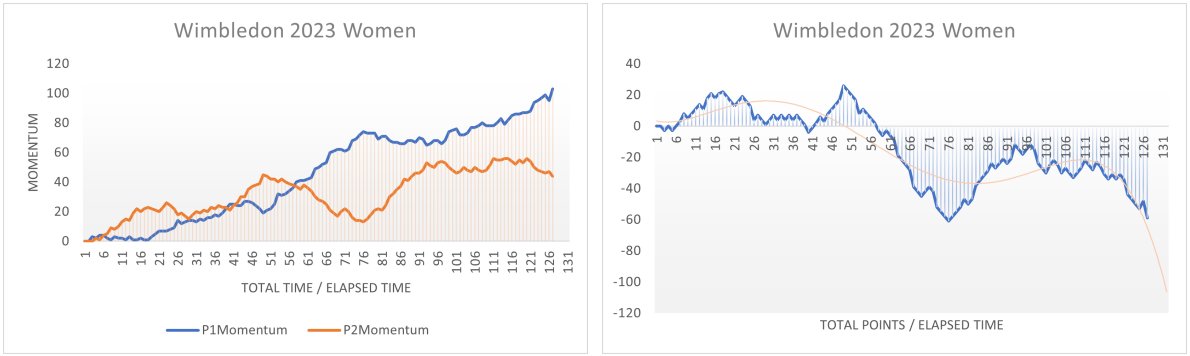
- First, the original data used for our model fitting is only one game and is not universal.
- Second, the special score that we considered may not be comprehensive enough.
- Third, we ignored the interference of off-field factors in our hypothesis, but some cheers, such as the audience, may also affect the excitement of the players and affect the momentum of the game.

5.2 Fitting Universality

We looked up the data of other competitions[10] and found that the models we established showed **good fitting accuracy** in tennis in the same event, and also had good fit in similar events such as table tennis.

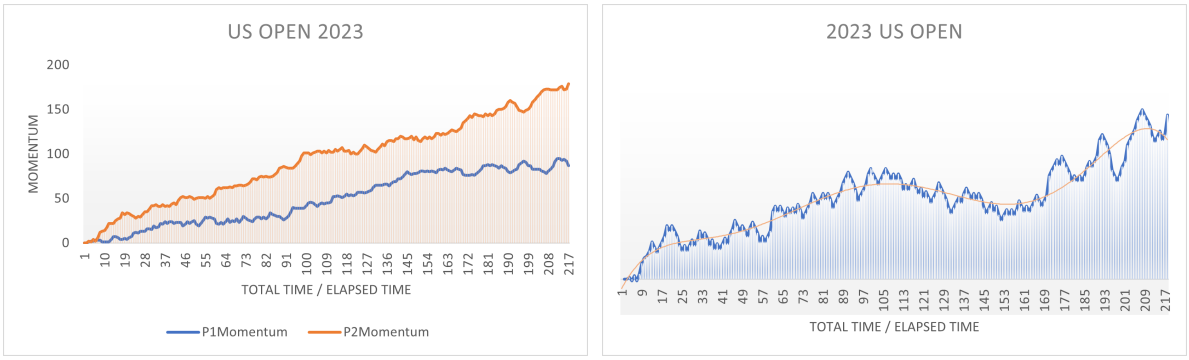
For example, we want to fit a table tennis match with our model. We have to find his special points and ordinary points first. For example, the scoring on the far o, the serving directly, and the receiving directly are special scoring points, which have a great effect on the momentum of a player. But like some marginal balls, playing tennis is only a normal score, with little impact on the momentum. After making such a classification, we will use the model to fit it again, and we will find that the fit is very good.

The next few pictures are the fits of the 2023 Wimbledon women's singles and the 2023 US Open, and it can be found that our model is highly restored.



(a) Line Graph of Momentum (b) Difference of Momentum

Figure 13: 2023 Wimbledon Women's



(a) Line Graph of Momentum (b) Difference of Momentum

Figure 14: Fitting Curve

However, the fitting function we do does not consider the off-site factors and scores the fitting criterion. We found that there were big differences when we applied our model to some **group events**, such as football, basketball, volleyball and other games, etc. There are more factors to consider momentum in collective projects, because the momentum is no longer the person's play, but the vector of a team's momentum. In the table tennis data, we found that he fitted the singles relatively well, but slightly insufficient in the doubles. This is because doubles test their cooperation and offensive ability and pure scoring does not fully reflect their momentum.

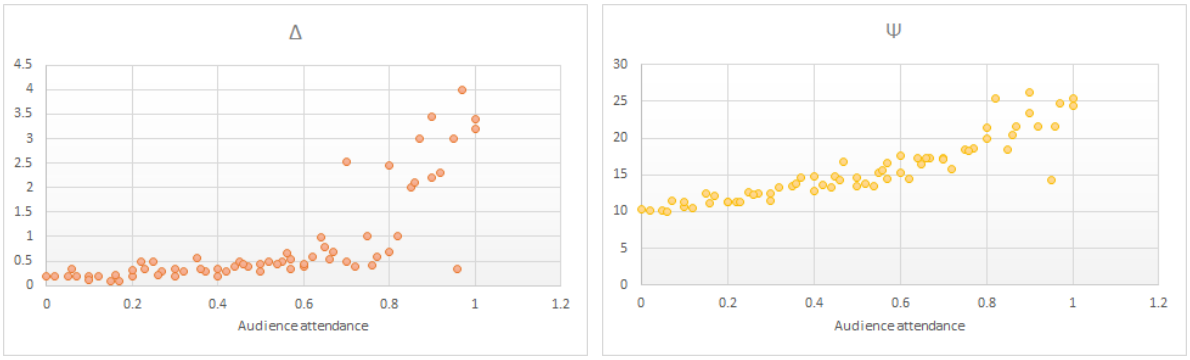
At the same time, in some low-scoring games, such as football, there are very few points in a game. We can not apply our model directly. Because in football, every attack and every successful defense increases the momentum of the team, but that's not reflected in scoring.

In conclusion, our model is completely **score-dependent**, predicting better matches for high-scoring, individual matches, and less effectively for low-scoring teams.

6 Analysis on Model's Sensitivity

In order to explore how momentum are influenced with the audience, we found the Wimbledon figures during the outbreak, when the attendance rate was distributed from 0% to 100%. We use different attendance rates to do specific analysis.

In our model, the factor of the **audience** is completely ignored. However, the shouts and cheers of the audience affected the play and mentality of the players to some extent. When a player is in a slump, if the audience is cheering for him, then his momentum will grow. By doing this sensitivity analysis, we mainly consider the degree of interference of the audience on our model.



(a) Predicted Momentum Differences (b) The Average of the Momentum

Figure 15: Model's Sensitivity with Audience

Judging from Figure a, when the audience attendance rate is less than or equal to 0.8, it has **almost negligible influence** on our model, indicating that our model is less sensitive to the audience and can be applied to most cases. When the attendance rate is greater than 0.8, the impact will increase significantly, because the attendance rate of the popular race must be 100%. Therefore, its audience shouts and cheers will be particularly strong, and the excitement of the athletes will be stronger.

In order to further study **the influence of the audience** on our model, we draw the relationship between the average of momentum and the audience attendance (Figure b). We can find that the average of momentum is positively correlated with the average of the audience attendance, but the change is not obvious, and the attendance will **increase significantly** when it approaches 100%.

Overall, our model has **a very good sensitivity**.

7 Strengths and Weaknesses

7.1 Strengths

1. Can reveal **trends, periodicity and seasonality** in tennis scoring data. Through time series analysis, long-term trends and cyclical changes in tennis scoring data can be identified, helping to predict future changes.

2. Can be used for **prediction and predictive analysis**. Through time series models, future trends can be predicted based on past player performance, helping with coaching decisions and plans.

3. There are many rating indicators for tennis data, so we can use **principal component analysis** to replace the original indicators with a few indicators while retaining most of the information.

4. In our key influencing factors about the momentum of tennis principal component analysis comprehensive evaluation function, the principal component weight for its contribution rate, it reflects the key influencing factors of tennis potential contains the proportion of all information, so determine the weight is **objective**, reasonable, it overcome some evaluation method to determine the weight of defects.

5. Our model is able to capture the random fluctuations of the players. In the competition, the psychology of players will fluctuate greatly with the progress of the competition. Time series

analysis can help us understand the random changes of players and judge their impact on the momentum.

7.2 Weaknesses

1. Some assumptions will need to be met. Our ARIMA is based on some **assumptions**, such as the outside world does not affect the momentum of the players, the momentum is only related to scores, etc. If the data does not meet these assumptions, it may lead to inaccurate analysis results.

2. The meaning of our principal component interpretation of momentum is generally somewhat **ambiguous**, not as clear and exact as the meaning of the original variable, which is the price has to pay in the process of variable dimension reduction.

References

[1] TUE 18 JUL 2023 14:00 BST, A new Wimbledon champion, the first 24 hours, The final shot to take the title was just the start for Carlos Alcaraz

[2] Ben Moss & Peter O'Donoghue,Momentum in US Open men's singles tennis

[3] J. D. Hamilton, Time series analysis, en. Princeton, NJ: Princeton University Press, Jan. 1994.

[4] https://www.wimbledon.com/index.html

[5] Si Shoukui&Sun Zhaoliang, Mathematical modeling algorithm in application (second edition)

[6] Guo Wenxia&Zhao Guangtao, Construction and application of the technical and tactical performance evaluation model of the tennis match

[7] Zhou Wu, Li Zixi, Cao Yingying, Analysis of winning factors by Daniel Medwade, an elite tennis player

[8] MENG Fan-ming;HUANG Wen-ming(Hulunbeier University Institute of Physical Education,Meral 021008,China;Physical Education Department,Fuzhou University,Fuzhou 350002,China),Analysis of Female Tennis Players Winning Factors Based on Decision Tree

[9] Tao Zhixiang Qi Bing Hu Yabin Qiu Rong Dong Health care, Exploration on Tennis Service System

[10] https://github.com/JeffSackmann/tennis_slam_pointbypoint

Memo for the Coach

This memo serves as an in-depth exploration of the profound impact of momentum on tennis matches, offering valuable insights for athletes to navigate unexpected incidents with strategic precision.

Momentum is a dynamic reflection of a player's athletic prowess, capturing the essence of their control over the competition and their genuine enjoyment of the competitive process. It goes beyond a mere numerical representation, directly influencing a player's capability to convert points and navigate the intricate balance of avoiding mistakes while embracing adventurous return strategies. The ebb and flow of momentum between players is a nuanced dance that shapes the unfolding narrative of the game.

When your athlete executes a stellar shot—a powerful ace or a flawless net point—it signifies a heightened competitive state. Emphasize to your players the importance of not just playing the game but appreciating it. Encourage them to explore diverse tactics fearlessly, ensuring that their momentum remains not just steady but ascends. This holistic approach transforms them into formidable forces on the court, ready to face any challenge.

The process of understanding momentum involves the establishment of intricate models, meticulous calculations, and insightful charting. Focus on pivotal points where one player's momentum eclipses another, as these instances mark significant shifts in the game's dynamics. Coaches play a crucial role in dissecting these moments with athletes, unraveling the underlying reasons behind these reversals and providing a roadmap for strategic adjustments.

Furthermore, coaches should lead discussions on the nuanced art of adjusting mentality and employing optimal strategies in diverse circumstances. This collective wisdom equips athletes to astutely gauge their current state during similar situations, empowering them to stabilize in potential overtaking scenarios, sustain momentum during performance peaks, and gracefully resist disruptions by opponents' attacks. It is a holistic strategy ensuring not just victory but sustained excellence.

In conclusion, by delving into the intricacies of momentum, athletes can elevate their game to new heights. This memo is not just a guide; it's an invitation to appreciate the dance of momentum, providing athletes and coaches alike with the tools to shape their narrative and strive for excellence in every match.

MCM#2429225

2024.2.5