Measuring the wind vane of Success by Momentum based on PCA

Yiyang Liu[1, *, #], Runchi Gu[1, #], Jiali Yang[2]

[1]School of Mathematical Sciences, Nankai University, Tianjin, China, 300071

[2]Department of Film and Dance, Tianjin University of Commerce Boustead College, Tianjin, China, 300192

* Corresponding Author Email: 2113819@mail.nankai.edu.cn

#These authors contributed equally.

Abstract. Tennis is a sport favored and admired by countless number of people throughout the world, which comprises two playing styles: singles and doubles. The paper presents a quantitative definition of the "momentum" and constructs two models regarding "momentum" and performance through analyzing various factors in the matches. We divide the given data into the training set and the testing set, and extract useful information under each point. Then, through analyzing various factors comprising serve counts, situation of ace and so on, conduct the dimensionality reduction utilizing PCA. Then we classify the performance of players per point into three key factors: endurance performance, serving performance and scoring performance and give the corresponding evaluation standard. Finally we utilize our models and analysis of the result, we present several related suggestions for coaches and players.

Keywords: Principal component analysis (PCA), Entropy weight method (EWM), Analytic hierarchy process (AHP), swings in matches

1. Introduction

Players stand on opposing sides of the rectangular court, whose surface is made of various materials such as grass, clay or hard [1]. In this sport, one team (one or two player) is assigned to be the server and the other to be the receiver throughout every point of the game. Swings in tennis matches frequently occurred with one player winning the first set by a relatively large margin but losing the second set reversely [2]. Hence, a fun phenomenon in tennis that can account for the situation is the "momentum", which generally means the psychological or physiological advantage for players at one point in a match [3].

In the paper we provide a precise means of calculation for the "momentum" in tennis matches and judge whether the phenomenon have an effect on swings in play. Then we establish a mathematical model to predict the swings in tennis matches utilizing the data set of Wimbledon Gentlemen's singles matches. Take into account of the "momentum" swings in past matches and offer suggestions to a player regarding preparation for the next opponent [4]. Then we generalize the model so as to adapt to different situations such as tournaments, Women's matches, court surfaces and similar sports.

2. Preliminary

2.1. Assumptions

1.Previous winning is always considered as a positive impact to players, omitting the fact that they may become over confident. Simultaneously, previous losing is always considered as a negative impact, omitting the fact that players may display intense revenge mentality and become more motivated.

2.Omit the influence on consecutive winning, direction of serve, depth of serve and depth of return since these four factors are unquantifiable factors in terms of the "momentum" of the player. To

exemplify it, in terms of the direction of serve, no matter whether the direction is body & center, body & wide, body, center or wide, there is always an opportunity to create difficulty for the opponent [5].

3.The player's mindset is seldom affected during games and between sets, otherwise the model will be hardly achieved considering its complexity.

4.The performance evaluation model is constructed through factors of subjective judgement and hence it is reasonable to construct the model utilizing AHP.

2.2. Notations

The primary notations used in this paper are listed in Table 1.

Table 1. Notations

Symbol	Description
x_1/X_1	Serving player before/after standardization(ditto)
x_2/X_2	Serve counts
$x_3/X_3(Y_4)$	Score situation of the last point
$x_4/X_4(Y_5)$	Situation of ace
x_5/X_5	Situation of hitting untouchable winning shot
x_6/X_6	Category of the untouchable shot
$x_7/X_7(Y_6)$	Situation of double fault
$x_8/X_8(Y_7)$	Situation of unforced error
$x_9/X_9(Y_8)$	Player's position to the net
$x_{10}/X_{10}(Y_9)$	Score situation while at net
$x_{11}/X_{11}(Y_{10})$	Winning of the breakpoints
$x_{12}/X_{12}(Y_{11})$	Miss of the breakpoints
x_{13}/X_{13}	Number of shot during the point
x_{14}/X_{14}	Distance run during the point
x_{15}/X_{15}	Speed of the serve
Y_1	Serving quality
Y_2	Excitement level of receiving points
$Y_3(Z_1)$	Energy expenditure (Endurance Performance)
M_i	"Momentum" per point
AM	Arithmetic mean of the "momentum" per game
RM	"Relative Momentum" per game
Z_2	Serving performance
Z_3	Scoring performance

3. "Momentum" Prediction Model in Tennis Matches

We conduct the qualitative analysis on the meaning of the "Momentum" which is the level of psychological or physiological impact on the player concerning the previous point. Model I, however, presents a quantitative analysis on the "Momentum" and predicts its trend in other tennis matches.

3.1. Factor Extraction Regarding the "Momentum" of Players

Before processing the data listed in the "Wimbledon_featured_matches", the first step is to select the useful columns of the data set. After careful consideration, we utilize the scenario of 11792 points (combining player one and player two) in the first 25 matches as the training set to construct our prediction model, and following matches (26-31) as testing set to test the ability of the model. To simplify the process, define the factors with positive impact on the player's mindset (which is likely to proliferate his performance) as 1, the negative impact as -1 and those factors with neutral (neither boost nor reduce) impacts as 0. Hence, the factor extraction process can be listed as follows.

1.If the serving player is player one himself, it can be deduced that it is a very positive impact, and hence shall be marked as 1. If the serving player is the opponent player two, the factor shall be marked

as -1. Moreover, severing once is a positive impact and twice is negative. Hence the former one is 1 and the latter one is -1.

2.It is sufficient to mark the previous victory of himself as 1 and victory of his opponent as -1. Note that the first cell shall be marked as 0 since neither side wins.

3.Mark the impact of player one himself hitting an untouchable winning serve in the previous point as 1, player two as -1 and others as 0. The impact on hitting an untouchable winning shot is likewise.

4.Create a new column "winner_shot_type_influence" based on it to study the influence on the category of untouchable shot. In this scenario, previous forehand and backhand shots of player one is marked as 1, player 2 as -1 and others as 0.

5.Mark the impact on previous lost point of player one due to the missing of two serves as -1, player two as 1 and others as 0. The impact on unforced error is likewise.

6.If player one makes it to the net, mark the factor as -1, player two as 1 and others as 0. During the previous game, if player one won the point while at net, mark it as 1, player two as -1 and others as 0.

Known from previous analysis, there are 15 factors which contributes to the "momentum" of one player versus the other in the pursuit of every point, which are the serving player(x_1), serve counts(x_2), score situation of the last point(x_3), situation of ace(x_4), situation of hitting untouchable winning shot(x_5), category of the untouchable shot(x_6), situation of double fault(x_7), situation of unforced error(x_8), player's position to the net(x_9), score situation while at net(x_{10}), winning of the breakpoints(x_{11}), miss of the breakpoints(x_{12}), number of shot during the point(x_{13}), distance run during the point(x_{14}) and finally the speed of the serve(x_{15}). Table 2 shows the specific details.

Table 2. Excerpt of the Factor Extraction

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}
1	-1	0	0	0	0	0	0	0	0	0	0	6.000	2	95
1	1	-1	0	0	0	0	-1	0	0	0	0	5.253	1	118
1	1	1	0	0	0	0	0	0	0	0	0	13.800	4	120
1	1	-1	0	1	0	0	-1	1	0	0	0	51.108	13	130
1	1	1	0	1	1	0	0	0	0	0	0	0.649	1	112

3.2. Data Processing Based on PCA

After extracting the useful information from the training dataset, the following step is to identify the correlation between all these factors and dimensionality reduction, which can be achieved through principal component analysis (PCA) [6]. Due to the fact that the values of these 15 factors are distributed dis orderly, a z-score standardization is required to allow the data follow the same distribution state. Hence, the standardized table of the 15 significant factors related to the "momentum" of tennis players is obtained [7-8]. The excerpt is given below, where values are kept three decimal places. It is shown in Table 3.

Table 3. Excerpt of the z-score standardization of the original 15 factors

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}
0.999	-1.156	0.000	0	-0.000	0.000	0	0.000	0.000	0	0	0	-0.576	-0.313	0.895
0.999	0.865	-1.002	0	-0.000	0.000	0	-1.943	0.000	0	0	0	-0.634	-0.638	1.112
0.999	0.865	1.002	0	-0.000	0.000	0	0.000	0.000	0	0	0	0.024	0.338	1.131
0.999	0.865	-1.002	0	1.744	0.000	0	-1.943	2.077	0	0	0	2.892	3.265	1.225
0.999	0.865	1.002	0	1.744	2.126	0	0.000	0.000	0	0	0	-0.987	-0.638	1.056

3.3. Dimensionality Reduction Utilizing PCA

The first step of the principal component analysis is to calculate the covariance of the 15 factors listed above. The result of the covariance is shown as Figure 1.



Figure 1. Covariance matrix of the 15 factors

Through figure 1, it is sufficient to claim that there is a straightforward correlation between the factors $\mathbf{X_1}$, $\mathbf{X_2}$ and $\mathbf{X_{15}}$; $\mathbf{X_5}$ and $\mathbf{X_6}$; $\mathbf{X_{13}}$ and $\mathbf{X_{14}}$. Hence, calculate the eigenvalues of these three pairs respectively and sort them in descending order.

In terms of factors X_1, X_2 and X_{15}, the eigenvalues and their corresponding eigenvectors are listed in the table 4.

Table 4. Eigenvalues and eigenvectors of factors $\mathbf{X_1}$, $\mathbf{X_2}$ and $\mathbf{X_{15}}$

Eigenvalue	**Eigenvector**		
2.2881	0.6297	0.3010	0.7620
0.6450	0.4654	-0.8843	-0.0375
0.0670	0.6220	0.3569	-0.6969

Moreover, the cumulative contribution rate reaches 0.7627 by the first eigenvalue, which is approaching the standard rate 80%.[5] Therefore, consolidate the three factors into one factor, which can be explained as serving quality(Y_1) after analyzing the real meanings of these three factors. The cumulative contribution rate reaches 0.7170 and 0.9008 by the first eigenvalue respectively. Therefore, consolidate the four factors into two new factors, with the former two into the factor named excitement level of receiving points (Y_2) and the latter two into the factor named energy expenditure(Y_3)[9].

4. Prediction of the Testing Set

4.1. Prediction

Since Model I is established from the training set which comprises 25 matches, the remaining 6 matches are served as the testing set to predict the “momentum”. Through inserting these data into the model, the momentum of both player one and player two per point is obtained, the excerpt of which is given in Table 5.

Table 5. Excerpt of Values of the "Momentum" of the Testing Set

a) **Match 26**

time	p1 momentum	p2 momentum
0:00:00	47.56517665	54.54503244
0:00:20	24.00900642	78.66710010
0:00:47	18.84668457	83.49949534
0:01:09	20.10688567	80.78439104

b) **Match 27**

time	p1 momentum	p2 momentum
0:00:00	42.21601229	59.23055250
0:00:19	18.67309452	79.07976767
0:00:44	18.57015178	79.21157207
0:01:14	11.60580633	90.36779951

c) **Match 28**

time	p1 momentum	p2 momentum
0:00:00	42.93892202	59.03486775
0:00:31	15.71411187	80.10359152
0:01:15	63.54262322	38.95868969
0:01:38	18.90042489	77.47156327

d) **Match 29**

time	p1 momentum	p2 momentum
0:00:00	58.94145133	39.35720221
0:00:27	74.86144201	19.66932677
0:01:10	84.22603937	13.50278346
0:01:36	75.06397852	23.73144313

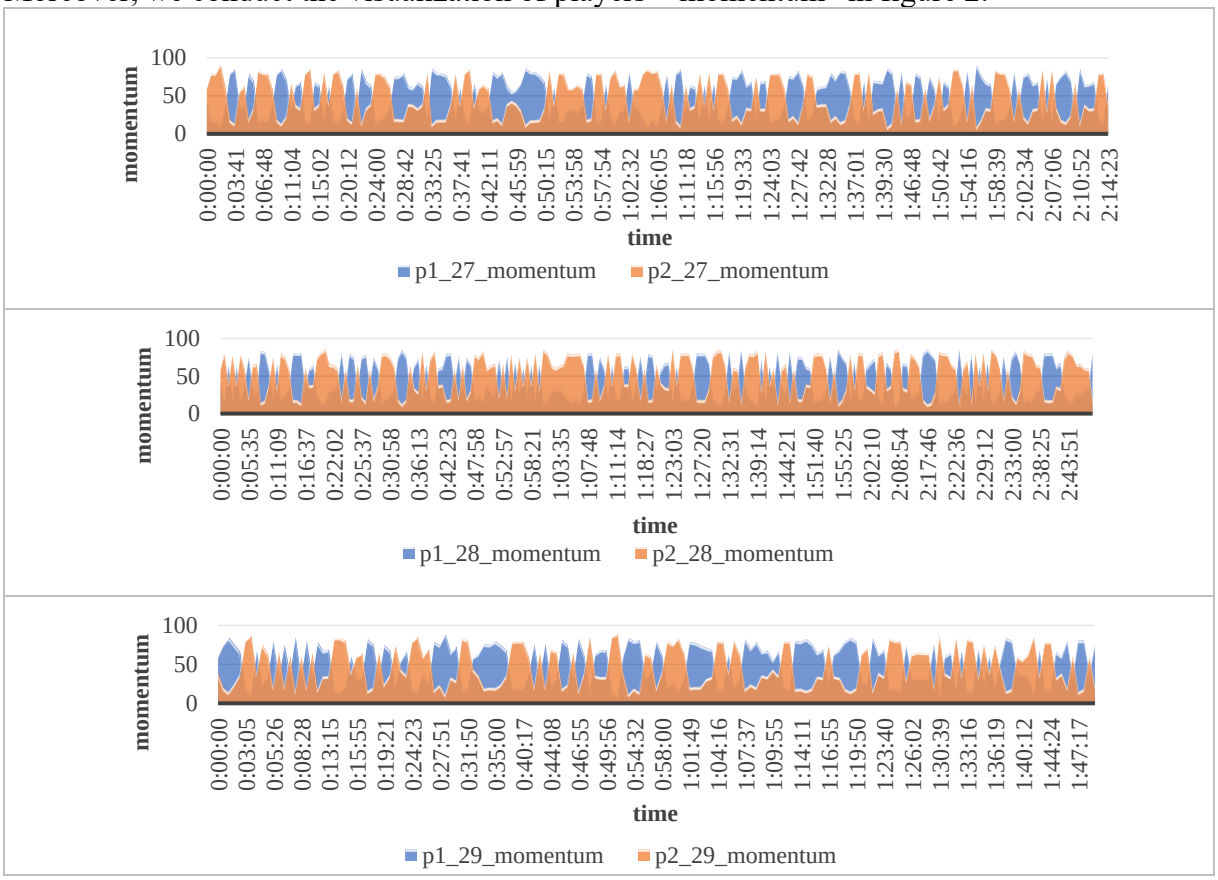
e) **Match 30**

time	p1 momentum	p2 momentum
0:00:00	43.78210348	57.90578855
0:00:35	19.63044726	78.38387647
0:01:28	27.9882901	73.72772756
0:02:00	63.72669098	37.98355567

f) **Match 31**

time	p1 momentum	p2 momentum
0:00:00	42.71395773	59.0813444
0:00:25	21.58071597	80.19872814
0:01:19	59.26619154	38.19702864
0:02:17	59.81977584	38.78444726

Moreover, we conduct the visualization of players' "momentum" in figure 2.



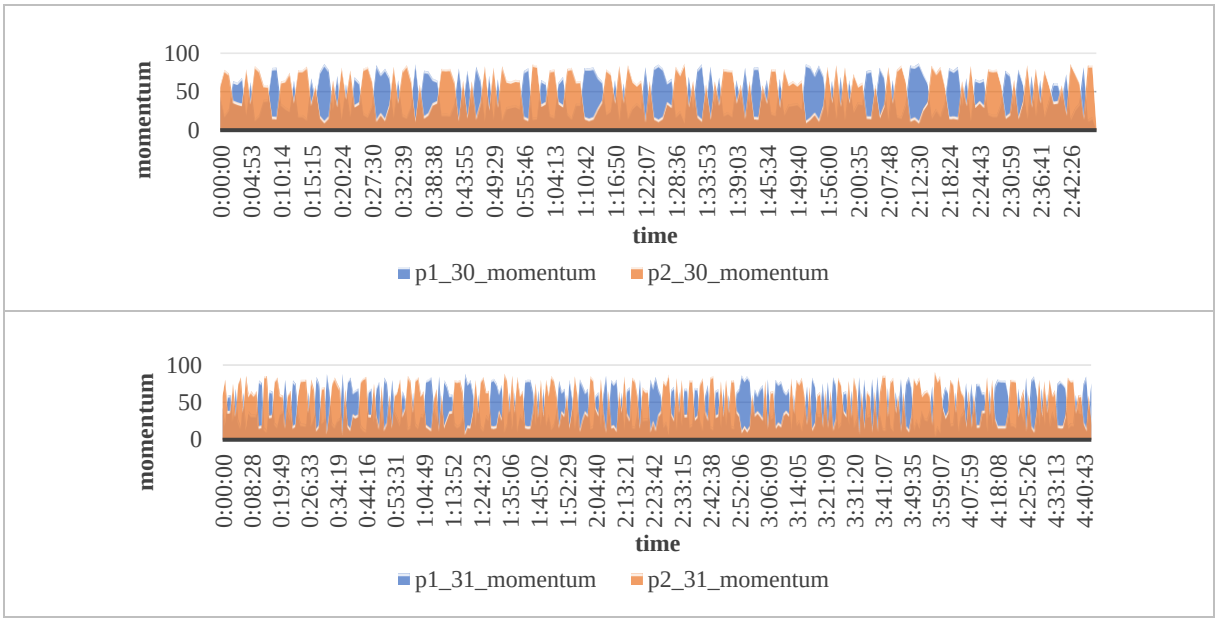


Figure 2. Visualization of the "Momentum" of both players in the Testing Set

4.2. Analysis

In terms of the “momentum” prediction model, we processed 15 factors from the given data set in total. Thus, the model encompasses comprehensive consideration and can be applied to multiple scenarios. The utilization of the entropy weight method is completely objective and hence yields good performance[10]. As per Model I, we did omit some complicated factors such as set-to-set and game-to-game situation, home advantage, international ranks of the players and historical encounters. In the future model, we may need to take them into consideration.

5. Conclusion

We proposed two different methods and models based on the relationship between “momentum” in tennis games and various factors during each point. To ensure the stability of our model, we tested the correlation between the “momentum” and the outcome of victory or defeat and predicted the outcome per game according to the testing set. In addition, we conducted the accuracy test via two methods, among which the best accuracy ratio is $\mathbf{0.959641256}$ and the least accurate one is $\mathbf{0.914893617}$. After comparing the prediction with the actual result, we reached a conclusion that the model has a relatively good generalization ability. Since coaches are not allowed to instruct the players in professional tennis matches, their approaches of taking advantage of the “momentum” are primarily through everyday training. Through assisting players with the suggestions given above, we deem that the “momentum” of players will definitely gain steadily.

References

[1] Anuraj A, Boparai G S, Leung C K, et al. Sports data mining for cricket match prediction [C] International Conference on Advanced Information Networking and Applications. Cham: Springer International Publishing, 2023: 668-680.

[2] Empacher C, Kamps U, Volovskiy G. Statistical prediction of future sports records based on record values [J]. Stats, 2023, 6(1): 131-147.

[3] Zhou, D., Chong, Z., & Wang, Q. (2020). What is the future policy for photovoltaic power applications in China? Lessons from the past. Resources Policy, 65, 101575.

[4] Putukian M, Purcell L, Schneider K J, et al. Clinical recovery from concussion–return to school and sport: a systematic review and meta-analysis [J]. British journal of sports medicine, 2023, 57(12): 798-809.

[5] Susanto E, Bayok M, Satriawan R, et al. Talent identification predicting in athletics: A case study in Indonesia [J]. Annals of Applied Sport Science, 2023, 11(1): 0-0.

[6] Chiu R M Y, Zhang L, Teng J, et al. Motivation, injury prevention, and the incidence of sports injuries: A three-wave longitudinal test of self-determination theory [J]. Scandinavian Journal of Medicine & Science in Sports, 2023.

[7] Shang Z, Chen Y, Wu Y, et al. Subjective and objective video quality assessment of high dynamic range sports content [C] Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023: 556-564.

[8] Casper J M, Pfahl M E, McCullough B. Intercollegiate sport and the environment: Examining fan engagement based on athletics department sustainability efforts [J]. Journal of issues in intercollegiate athletics, 2023, 7(1): 13.

[9] Depken, C. A., Gandar, J. M., & Shapiro, D. A. (2023). Set-level strategic and psychological momentum in best-of-five matches in professional tennis. Applied Economics Letters, 30(6), 735-739.

[10] Sun H C, Lin T Y, Tsai Y L. Performance prediction in major league baseball by long short-term memory networks [J]. International Journal of Data Science and Analytics, 2023, 15(1): 93-104.