ISSN 2618-1576 Vol. 7, Issue 2: 33-39, DOI: 10.25236/FSR.2025.070206

Construction of prediction results for men's singles tennis at the US Open

Fuqian Wang[a], Zhuojian Li*

School of Physical Education, Xiangnan University, Chenzhou, Hunan, China
[a]2820419479@qq.com
**Corresponding author*

Abstract: *The US Open, one of the four Grand Slam tennis tournaments, is played on hard courts alongside the Australian Open, while the US Open typically takes place in August or September during the summer in New York. This study employs literature review and statistical methods such as SPSS to analyze 10 data points from 125 men's singles matches announced on the official website of the 2024 US Open. First, each indicator is coded for preprocessing. Then, KMO test and Bartlett's test for sphericity are used, along with total variance, rotated component matrix, and coefficient matrix, to derive the predictive model. The run test is conducted to check if the data meets randomness, and the selected data undergoes Shapiro-Wilk and Spearman correlation tests. The data are then analyzed in conjunction with the previously obtained model, and finally, the prediction results are tested. This study clarifies that winning break points and the probability of winning the second service point significantly impact actual match scenarios, emphasizing the importance of stability throughout a game.*

Keywords: *US Open; tennis; men's singles*

1. Introduction

The US Open (U.S. Open) is the final Grand Slam tennis tournament of the year, typically held at the USTA Billie Jean King National Tennis Center in Flushing Meadows, Queens, New York, from late August to early September [1]. Its unique blend of urban landscape and sports makes it a model of modern commercial operation. Originating in 1881 as the "National Association for Amateur Athletic Competition," the US Open was initially known as the "National Championships." It is the last of the four Grand Slams to be played each year and became an open tournament in 1968. The event features hard courts, which contrast sharply with the other three Grand Slams 'court types (grass, clay) [2]. Since its inception, the tournament has undergone several venue changes, moving to its current location in 1978 and retaining it ever since. The US Open was one of the first to introduce "Hawk-Eye" technology to assist in officiating, enhancing fairness and technological sophistication [3]. Renowned for its high level of competitive strength and substantial prize money, the US Open attracts top players from around the world. Through diversified event marketing (such as sponsorships and media partnerships) and professional management, it has developed a unique entertainment-oriented tournament culture, including music performances and interactive activities during the event [4]. The US Open is not only a stage for competition but also an important vehicle for global tennis culture dissemination. Its official website construction and live streaming technology (such as multi-angle replays) are at the forefront, further expanding the tournament's influence. Additionally, the US Open has driven the development of the American tennis industry and strengthened the foundation of local tennis culture through campus tennis programs [5].

2. Materials and methods

2.1 Research subjects

The study was based on 125 men's singles matches and 10 indicators of players announced for the 2024 US Open.

ISSN 2618-1576 Vol. 7, Issue 2: 33-39, DOI: 10.25236/FSR.2025.070206

2.2 Research methods

2.2.1 Literature method

Read relevant literature and books about the factors affecting tennis victory through the library. Collect and analyze data from the US Tennis Championships (Wimbledon Championships) to obtain relevant data of the competition, which provides sufficient theoretical basis for this study.

2.2.2 Mathematical statistics method

Using Excel to categorize and statistically analyze the 10 variables published on the official website of the 2024 US Open, including ACEs, double faults, first serve, first serve percentage, second serve percentage, break points, unforced errors, net distance, net distance per foot, and match outcomes. The data was organized and analyzed using spss2023, employing principal component analysis and linear regression to predict the results of the 2024 US Open men's singles matches.

2.2.3 Data sources and indicators

The research indicators for this topic were determined by collecting necessary data from the official website of the US Open, reviewing relevant literature to select appropriate indicators. The statistical indicators are as follows: ACEs (X1), double faults (X2), first serve (X3), first serve success rate (X4), second serve success rate (X5), break point conversion rate (X6), unforced errors (X7), net approach distance (X8), net approach distance per point (X9), and match result (X10).

Table 1 Statistics of winning factors for men's singles at the 2024 US Open

code	metric	Indicator unit
X1	ACEs	number
X2	Double fault on serve	number
X3	First serve	percentage
X4	The probability of winning the first serve point	percentage
X5	Probability of winning the second service point	percentage
X6	Win a break point	percentage
X7	Unforced error	number of times
X8	Coverage distance	foot
X9	Coverage distance / point	foot
X10	Results	Win 1, lose 0

As shown in Table 1, the units of indicators X1 and X2 are in numbers, while those of X3, X4, X5 and X6 are in percentages. X7 is in terms of frequency. The units of the above indicators X1 to X7 all conform to international standards and are used uniformly in different competitions. However, the units of X8 and X9 are in feet, which is unique to the US Open. Other competitions have different measurement units. The unit of the competition result indicator is win = 1 and lose = 0. Such a setting is for better analysis of the influence relationship between other indicators and X10, facilitating the construction of the model.

3. Data preprocessing and testing

3.1 KMO, test and Bartlett spherical test

Bartlett The test is used to check whether the correlation coefficient matrix between variables is an identity matrix (i.e., no correlation among variables). Its null hypothesis (H_0) is: the correlation coefficient matrix is an identity matrix (diagonal elements are 1, off-diagonal elements are 0). If the test is significant (p-value <0.05), then H_0 is rejected, indicating that there is a significant correlation among variables, making it suitable for factor analysis. KMO ≥ 0.9: very suitable; 0.8 ≤ KMO <0.9: suitable; 0.7 ≤ KMO <0.8: generally; KMO <0.5: unsuitable. [6]

Table 2 KMO and Bartlett test

KMO sample appropriateness measure.		.615
Ballotell's sphericity test	Approximate chi-square	616.581
	free degree	45
	conspicuousness	.000

ISSN 2618-1576 Vol. 7, Issue 2: 33-39, DOI: 10.25236/FSR.2025.070206

The KMO test and Bartlett spherical test were used for the 10 obtained indicators. According to Table 2, the measured KMO test coefficient was 0.615 (greater than 0.6), and the significance probability P=0.000<0.050, indicating that the validity of the data was high and it was suitable for factor analysis.

3.2 Factor score model

The loadings (Loadings) reflect the correlation between variables and factors, and the scores (Scores) are the quantified values of samples on factors. The two cannot be confused [7][8].

Table 3 Total variance explained

ingredient	Initial eigenvalue			Extract the sum of squared loads		
	amount to	variance percentage	accumulate %	amount to	variance percentage	accumulate%
1	2.587	25.868	25.868	2.587	25.868	25.868
2	1.741	17.412	43.280	1.741	17.412	43.280
3	1.631	16.309	59.589	1.631	16.309	59.589
4	1.046	10.459	70.048	1.046	10.459	70.048

According to Table 3, the principal component analysis method can be used to extract four common factors from 10 factors. As can be seen from the table, the variance contribution value of the four principal factors is Y1=25.869, Y2=4.3280, Y3=59.589 and Y4=70.048

Table 4. Component matrix after rotation

	ingredient			
	1	2	3	4
X4	.856	-.026	.041	-.116
X10	.823	-.255	.034	.016
X5	.630	-.468	-.149	-.024
X1	.611	.534	-.019	.100
X6	.437	.054	.035	.221
X2	-.012	.810	-.018	-.347
X7	-.191	.785	-.028	.147
X9	-.034	-.136	.887	.006
X8	.061	.116	.876	.089
X3	.057	-.056	.079	.946

The goal of the maximum variance method is to maximize the variance of factor loadings. By rotating the factor axes (maintaining orthogonality), it reassigns the relationship between variables and factors, making the factor structure clearer. performs an orthogonal transformation on the initial component matrix, which can further explain the specific meanings of the four factors. The data in Table 4 shows the close relationship between each indicator and the common factor; the higher the correlation coefficient of the four factor components, the greater the correlation.

Table 5 Component score coefficient matrix

	ingredient			
	1	2	3	4
X1	.295	.343	-.023	.110
X2	.063	.423	.019	-.268
X3	-.005	.030	-.039	.856
X4	.365	.040	.038	-.134
X5	.234	-.212	-.090	-.053
X6	.184	.074	.002	.192
X7	-.028	.426	-.035	.190
X8	.033	.070	.551	.003
X9	-.023	-.080	.567	-.086
X10	.331	-.080	.022	-.026

From Table 5, it can be seen that the indicators of Component 1 are X10, X5, and X4; this component is named "Advantage in Proactivity." The indicators of Component 2 are X1, X2, and X7; this component is named "Technical Stability." The indicators of Component 3 are X8 and X9; this component is named "Endurance Advantage." The indicators of Component 4 are X3 and X6; this

ISSN 2618-1576 Vol. 7, Issue 2: 33-39, DOI: 10.25236/FSR.2025.070206

component is named "Service Advantage." According to the data in Table 5, the model for the scores of this component is:

L1=0.295 * X1 + 0.063 *X2 - 0.005 * X3 + 0.365 * X4+ 0.234 * X5 + 0.184 * X6 - 0.028 * X7 + 0.033 * X8 - 0.023 * X9 + 0.331*X10

L2=0.343 * X1 + 0.423 *X2 + 0.030 * X3 + 0.040 * X4- 0.212 * X5 + 0.074 * X6 + 0.426 * X7 + 0.070 * X8 - 0.080 * X9 - 0.080 * X10

L3=-0.023 * X1 + 0.019 *X2 - 0.039 * X3 + 0.038 * X4- 0.090 * X5 + 0.002 * X6 + 0.002 * X7 - 0.551 * X8 + 0.567 * X9 + 0.022 * X10

Y4=0.110 * X1 - 0.268 *X2 + 0.856 * X3 - 0.134* X4- 0.053 * X5 + 0.192 * X6 + 0.190 * X7 + 0.003 * X8 - 0.086 * X9 + 0.022 * X10

L=(0.25868*L1+0.17412*L2+0.16309*L3+0.10459*L4)/0.70048

4. Model testing

4.1 Tour inspection

To assess whether the outcome of a match is random, we need to use the runs test to evaluate the randomness of the match result data. The runs test in SPSS quantifies the continuity characteristics of data sequences, providing researchers with an intuitive tool for judging data randomness. Its underlying logic is highly consistent with methods used in cryptography, statistics, and other fields for verifying randomness[9][10][11]

Table 6 Tour inspection

	X10
Check value a	.50
Number of cases <test value	124
The number of cases is greater than or equal to the test value	125
Total number of cases	249
number of runs	192
Z	8.446
Asymptotic significance (bilateral)	.000
a. average value	

As shown in Table 6, the match numbered "X10" was selected for the randomness analysis of the players' data in this match. The result shows that $P = 0.00 < 0.05$, which indicates that the outcome of a singles tennis match is not random.

4.2 Spearman correlation analysis

The Spearman correlation coefficient is calculated by converting the raw data to ranks (i.e., positions after sorting), rather than using the raw data values directly. This transformation makes the method independent of the normal distribution assumption [12] of the data

Table 7 Shapiro-Wilk test

	statistics	free degree	conspicuousness
X1	.942	242	.000
X2	.958	242	.000
X3	.994	242	.515
X4	.989	242	.054
X5	.986	242	.016
X6	.977	242	.000
X7	.975	242	.000
X8	.524	242	.000
X9	.966	242	.000
X10	.636	242	.000

In order to conduct the correlation of Spearman test, it is necessary to verify whether the data conforms to the normal distribution. As shown in Table 7, X3 and X4 have $P> 0.05$, which conforms to

ISSN 2618-1576 Vol. 7, Issue 2: 33-39, DOI: 10.25236/FSR.2025.070206

the normal distribution, while other factors do not conform to the normal distribution, so they are defined as "New Data" and imported into the analysis of Spearman correlation.

Table 8 Spearman test correlation

		X1	X2	X5	X7	X6	X8	X9	X10
X1	correlation coefficient	1.000	.227**	.167**	.193**	.133*	.160*	-.057	.303**
	Significance (two-tailed)	.	.000	.008	.002	.036	.013	.376	.000
	Number of cases	250	250	250	250	250	242	242	249
X2	correlation	.227**	1.000	-.354**	.488**	-.002	.054	-.154*	-.182**
	Significance (two-tailed)	.000	.	.000	.000	.975	.401	.017	.004
	Number of cases	250	250	250	250	250	242	242	249
X5	correlation	.167**	-.354**	1.000	-.315**	.232**	-.054	-.068	.601**
	Significance (two-tailed)	.008	.000	.	.000	.000	.403	.289	.000
	Number of cases	250	250	250	250	250	242	242	249
X7	correlation	.193**	.488**	-.315**	1.000	-.059	.214**	-.093	-.298**
	Significance (two-tailed)	.002	.000	.000	.	.355	.001	.148	.000
	Number of cases	250	250	250	250	250	242	242	249
X6	correlation coefficient	.133*	-.002	.232**	-.059	1.000	.000	-.021	.361**
	Significance (two-tailed)	.036	.975	.000	.355	.	.998	.742	.000
	Number of cases	250	250	250	250	250	242	242	249
X8	correlation	.160*	.054	-.054	.214**	.000	1.000	.854**	.015
	Significance (two-tailed)	.013	.401	.403	.001	.998	.	.000	.821
	Number of cases	242	242	242	242	242	242	242	242
X9	correlation coefficient	-.057	-.154*	-.068	-.093	-.021	.854**	1.000	.026
	Significance (two-tailed)	.376	.017	.289	.148	.742	.000	.	.692
	Number of cases	242	242	242	242	242	242	242	242
X10	correlation	.303**	-.182**	.601**	-.298**	.361**	.015	.026	1.000
	Significance (two-tailed)	.000	.004	.000	.000	.000	.821	.692	.
	Number of cases	249	249	249	249	249	242	242	249

After performing a Spearman correlation analysis on the selected 8 evaluation indicators, the correlation coefficients are shown in Table 8. If $\rho > 0.05$, it can be concluded that there is no significant difference, meaning there is no correlation between the indicators; if $\rho < 0.05$, then there is a significant difference, indicating that there is a correlation between the indicators.

4.3 Construction and analysis of regression model

Table 9 Model summary

model	R	R square	Adjusted R squared	Standard estimate error
1	.736a	.541	.524	.346

Using the previously determined 9 evaluation indicators as independent variables and the processed dataset "New Data" as the dependent variable, a regression prediction model was established through regression analysis. After testing, Table 9 shows that the goodness of fit R2 for the regression model is 0.524, indicating a good correlation between the independent and dependent variables

Table 10 Coefficients

model		Unstandardized coefficients		Standardization factor	t	conspicuousness
		B	standard error	Beta		
1	(constant)	-2.386	.345		-6.908	.000
	X1	.001	.005	.008	.150	.881
	X2	.004	.009	.025	.417	.677
	X3	.607	.390	.076	1.558	.121
	X5	1.646	.263	.341	6.254	.000
	X6	.372	.119	.144	3.125	.002
	X7	-.004	.002	-.102	-1.884	.061
	X8	2.676E-6	.000	.044	.788	.432
	X9	.000	.001	.006	.104	.917

As can be seen from Table 10, X5 and X6 have significant indicators and the highest correlation

ISSN 2618-1576 Vol. 7, Issue 2: 33-39, DOI: 10.25236/FSR.2025.070206

with the predicted results, and their correlation is arranged as X5>X6, so the regression prediction equation is obtained: N=-2.386+1.646X5+0.372X6,

4.4 Prediction accuracy

Using the prediction model presented in this paper, we conducted an overall prediction of several representative matches, including one final match, three qualifying rounds, and two quarterfinals at the 2024 US Open. By comparing the predicted outcomes with the actual results, it was found that the predicted results showed a high degree of consistency with the final outcomes. This research indicates that the prediction model is reliable and the results are credible.

5. Discussion

In tournaments like the US Open, winning break points and the probability of winning the second serve can significantly impact the outcome of the match. As a hard court event, the serving side's advantage may reduce the receiving side's ability to break serve, but players 'receiving skills (as mentioned by A. Sanchez-Pay,Jose Antonio Ortega-Soto et al.) might partially offset this disadvantage [13]. Although the scoring rate at break points is one of the key variables in determining victory or defeat (such as the significant P-value for the scoring rate at break points in the 2021 US Open study), the final predictive model shows that the serving score rate, receiving score rate, and return score rate have more comprehensive explanatory power. This suggests that success at break points requires a combination of defensive play in service games, quality of receiving, and other technical performances, rather than existing in isolation.[14]In Grand Slam men's singles matches, the success rate of converting break points has a significant impact on the match result. For example, Knight and O 'Donoghue (2012) found through empirical analysis that the probability of winning a break point is closely related to the match outcome, especially at critical moments (such as "break points"), where psychological pressure and technical stability can directly affect the player's success rate [15]. The second serve win rate (2ndW%) may be influenced by serving strategy and court type. The speed advantage of serves on hard courts may lead to a higher second serve win rate compared to clay courts, but it is necessary to balance stability and aggression [16].I. Prieto-Lage,A. Parames-Gonzalez et al. The second serve success rate is approximately 55% (possibly referring to serving into the service box), and the probability of winning that point may be slightly lower. In men's singles at the US Open, the probability of winning the second serve point is specifically around 50%, but this can vary depending on different studies and courts. Combining other research, it may be approximately 46%-50%[17].

This study focuses on the 2024 US Open men's singles tournament. Although the conclusions and analyses obtained under the specific restrictions of the venue and gender range cannot be fully and directly applied to the other three Grand Slam tournaments (such as Australian Open, French Open and Wimbledon), they still have specific reference value.

References

[1] Jiang Keri. (2018)“Comparative study on the operation mechanism of China-Australia Tennis Open.” Hunan Normal University .

[2] Wang Kaijun and Dong Shengyu. (2008)“Cultural Analysis of the Four Grand Slam Tennis Championships .” Journal of Xi 'an Institute of Physical Education .

[3] Tai Feng. (2008)“Discussion on the application of "Eagle Eye" technology in tennis”. Journal of Jilin Sport University.

[4] Xie Xianghe. (2013)“Analysis on the Evolution of Medal Area Distribution since the Four Grand Slam Tennis Championships”. Journal of Beijing Sport University.

[5] Zheng Sen, Xu Jian. (2019)“Development characteristics and implications of the US Open.” Sports Culture Guide.

[6] Andini, A., Arwiyah, Y., & Pangarso, A. (2017). Analisis Faktor-Faktor yang Mendorong Kinerja Dosen Tetap pada Fakultas Komunikasi dan Bisnis Universitas Telkom. Business Management Journal, 12(2).

[7] Kanlun Chen, Sheng Li et al. (2018)“Evaluating the Operational Performance Level of Chinese Traditional Medicine Listed Companies Based on Factor Analysis.” Modern Management. Hans.

[8] H. Ikeda. (1965)"GENERAL PRINCIPLE OF FACTOR EXTRACTION." . University of Tokyo / Yoko Ikegami.

ISSN 2618-1576 Vol. 7, Issue 2: 33-39, DOI: 10.25236/FSR.2025.070206

[9] Lina Ding, Chunyuan Liu et al. (2019)"A New Lightweight Stream Cipher Based on Chaos." Symmetry.
[10] Fethi Dridi, S. El Assad et al. (2022)"Design, Implementation, and Analysis of a Block Cipher Based on a Secure Chaotic Generator." Applied Sciences.
[11] Renxiu Zhang, Longfei Yu et al. (2021)"A Novel Plaintext-Related Color Image Encryption Scheme Based on Cellular Neural Network and Chen's Chaotic System." Symmetry.
[12] Fernanda Pires Ximenes, Fibria Indriati Dwi Liestiawati. (2024)"Perceptions of Female Administrative Employees on the New Career Advancement Model at Universidade Nacional Timor Lorosa'e." Journal of Law, Politic and Humanities.
[13] A. Sánchez-Pay, José Antonio Ortega-Soto et al. (2021)"Notational analysis in female grand slam tennis competitions." Kinesiology.
[14] Ling Yu, Yanyan Wang. (2023)"Analysis of tennis techniques and tactics based on multiple linear regression model." Applied Mathematics and Nonlinear Sciences.
[15] F. Spieksma. (2016)"Serving in Tennis: Deuce Court and Ad Court Differ." Working Papers of Department of Decision Sciences and Information Management.
[16] Yixiong Cui, Miguel-Ángel Gómez et al. (2018)"Performance profiles of professional female tennis players in grand slams." PLoS ONE.
[17] I. Prieto-Lage, A. Paramés-González et al. (2023)"Match analysis and probability of winning a point in elite men's singles tennis." PLOS ONE.