One More Shot: Predicting Wins in Women's Professional Tennis

Mikala Lowrance
Southern Utah University ('18)

Racquet technology, tactics, and an increased importance placed on improving player stamina have led the game of tennis to evolve into a more physically demanding sport over the past 20 years. This paper looks to examine rally duration's effect on match outcome of female tennis players at the professional level. The purpose of this research was to assess the relationship between point duration and winning to better understand to what extent a player's stamina as well as momentum shifts effect winning. Data was collected through the Tennis Match Charting Project, in which one hundred forty-one matches where used. Matches were only used from the Australian Open to keep consistency in surface and ball type. Results indicate that the player who is most successful on long rallies of 10 or more shots is both the player in greater physical shape, and is the player that will likely gain the most momentum. From these assumptions, it is speculated that the player with a higher number of long rallies won, will ultimately win the match.

Southern Utah University Working Paper Series

I. Introduction

In professional tennis does the player who wins long rallies serve as a determinant of ultimately winning the match? This paper seeks to determine the effect that winning longer rallies has on match outcome. It is known that longer points tend to occur on game points. Since it is also known that winning game points wins games and thus leads to winning matches, the importance of winning long rallies is clear. Previous studies have investigated the differences in rally length between court surfaces and shots per point at Grand Slam events. However, an analysis of this nature has not been undertaken, therefore resulting in knowledge on the effect of rally length remaining relatively unknown. The results of this research are applicable to tennis as the sport evolves and coaches seek to understand to what importance winning long rallies holds in terms of the grand outcome of the match. While coaches may understand the value of their player being in great physical shape, it is possible they should apply an increased focus on winning long rallies into their player’s minds. Long rallies could serve as a momentum shift or it could signal which player is in better physical shape. If a player has better endurance over another, this would suggest they will likely be the winner of the match. It could then be said that the player winning a greater number of long rallies will be the ultimate winner.

II. Data

Data was collected by the Tennis Match Charting Project which is a crowdsourced data collection source that includes detailed charting of over 3500 professional matches (Sackmann). This data was collected either through a spreadsheet or through an app on a smartphone in which individuals chart various statistics from professional tournaments. Some of the statistics charted from one match include first serve percentage, winners, unforced errors, and forced errors. Once the match is charted, individuals send the matches to the Tennis Match Charting Project who then reorganizes the data into a Microsoft Excel spreadsheet available for use.

The dataset used for this research provides the variables for points won by each player for each rally length. This is additionally separated by if Player 1 or Player 2 is serving. Furthermore, within the dataset there are winners, unforced errors, and forced errors included. The data provides the players name, the tournament, and the year so that additional distinctions in the data can be compared.

For this paper, rally length data statistics were used for the number of points won in each rally length category. The table below explains the how the number of shots hit are counted.

Charting Shot Count

Definition	Description
Rally Start	The first shot of the rally was charted as when the player hit the serve
Rally End	The end of the rally is counted as when the first bounce is outside of the boundaries, hit into the net, or the last ball hit when the other player cannot return the ball
Number of Shots	This is classified only as successful shots. Shots that land outside of the boundaries, or are hit into the net are not calculated into rally length

Additional data was collected in order to include the deciding set binary variable. This is labeled as 1 if the match ended in playing a third set to decide the winner of the match. Furthermore, a match outcome binary variable was included as 1 if player 1 won or 0 if player 1 lost. This data was collected from the Women's Tennis Association (WTA) online website which provides the tournament draw and scores from every professional women's match ever played. From these additional variables added, interaction terms of rally length and if a third set was played were able to be created within the data set.

Difficulty arose while cleaning the data as many matches had to be eliminated due to the match resulting in one of the players withdrawing before it was completed. Additionally, some of the data includes the round of the tournament instead of the player's name; these observations were searched through the WTA to discover the player match up from the given round of the tournament and the outcome in order for the names to be properly replaced, without having to drop all of the match data.

III. Method

Matches from the Women's draw of the Australian Open beginning with year 2000 until 2017 were used in this study in order to keep consistency with court surface and ball type. By using a single tournament, court surface type is able to be controlled for. Surfaces in tennis make large impacts on rally lengths as they speed up the pace of the ball or slow it down. The Australian Open court is a mid-range speed of court, as it is a hard court which provides speeds in between that of grass courts and clay courts, (Cutler, 2016).

In order to discover the possible effect on outcome based on rally length, number of points won by player 1 in each rally group (1-3, 4-6, 7-9, and 10+ shots) were used as independent variables. Total points won in a given rally length of a given match were also included as a control. Since tennis can deliver a wide range of points played, it is important to include total points in order to make the number of points Player 1 won a meaningful number and comparable to the points won by Player 2. Additionally, a binary variable for if a third and deciding set was played was included. Since playing a third set would likely test a player's stamina more so than just a two-set match, this variable was then interacted with each rally length category. The model is represented below:

$$\Pr(winning = 1)_{it} = \beta_0 + \beta_1(rallylength4-6)_{it} + \beta_2(rallylength7-9)_{it} + \beta_3(rallylength10+)_{it} + \beta_4(totalpointsrally4-6)_{it} + \beta_5(totalpointsrally7-9)_{it} + \beta_6(totalpointsrally10+)_{it} + \beta_7(rallylength4-6*decidingset)_{it} + \beta_8(rallylength7-9*decidingset)_{it} + \beta_9(rallylength10*decidingset)_{it} + u_{it}$$

OLS was used in this study to predict match outcome in a linear regression model. Through looking at player *i*, in match *t*, the panel data was able to be used to look into effect of rally length. Furthermore, including the interaction of the deciding set variable and rally length, the effect of rally type in a match that goes to three sets on outcome was able to be examined.

By including the interaction of a third set with each set of rally lengths, the overall impact of playing an additional set is seen. It is important to look at this impact as, theoretically, a third set would make the match 50% longer than just a two set match. Looking at these longer matches makes the importance of a long point much more prevalent. Since our hypothesis places importance on player stamina as a contributing factor of winning, a third set will further clarify the effect of a better player's stamina leading to winning longer rallies and ultimately effecting the overall match outcome.

IV. Results

The linear regression model examines the probability of Player 1 winning the match based upon the number of points won in each rally length category. Controlling for total points played within the match, a change in points won for a rally of 1-3 shots leads to a 6.4 percentage point increase in the probability of winning the match. For a rally that is 4-6 shots that number is a 4.6

percentage point increase, and a 5.5 percentage point increase for 7-9 shot rallies. Ultimately, in this study we are looking to examine the 10+ shot rallies, as they are the longest rally type of our study and deliver results that can be used to look into how a player's stamina is effecting their performance. A single point increase in rallies 10 or more won leads to a 2.8 percentage point increase in probability of winning the overall match.

While results based solely upon points won by player one in a given rally length category do appear to be consistent with the idea of rallies of 10 or more shots being most important, when looking at rally lengths of matches that are determined by a third set, results will greater allow us to see the impact that a player's stamina levels have on outcome. For matches that conclude in a third set, points won in rally lengths of 10 or more appear to matter much more than any other category of rally length. Table 2 shows that when controlling for total points and two set matches, a change in points won for a rally of 10 or more shots leads to a 2.6 percentage point increase in the probability of winning the match. Results indicate that the interaction term of 'rally length 10 or more' and 'deciding set' is the only statistically significant outcome produced of the interaction variables.

Interestingly, winning rallies of 7-9 shots when the match is three sets indicated a 1.5 percentage point decrease in probability of winning the match, at the insignificant level. It is possible this result occurred due to players exerting extra energy to win rallies of 7-9 shots, without as much gain as 10 or more rallies. A higher number of rallies were played in the 7-9 shot category than the 10 or more. Therefore, it is likely that the gain from winning the point of this rally length is cancelled out by energy expended compared to the gain from a 10 shot rally. Winning a 10 shot rally elicits belief in the player's mind, along with fans cheering more for winning a long rally compared to simply a 7-9 shot rally. From these assumptions, it is possible that winning 7-9 shot rallies decreases probability of winning and that this result did not appear just based upon a small sample size.

Regression 1. Binary Outcome Y=1 if Player 1 Won Match (Linear Probability Model)

Variables	Coefficient
Rally 1-3	0.0639***
	(0.0116)
Rally 4-6	0.0459***
	(0.004)
Rally 7-9	0.0551***
	(0.0075)
Rally10+	0.0283***
	(0.0103)
Total Points 1-3	-0.0423***
	(0.0066)
Total Points 4-6	-0.0221***
	(0.0024)
Total Points 7-9	-0.0248***
	(0.0041)
Total Points 10+	-0.011
	(0.0072)
Deciding Set	-0.1076
	(0.2165)
Rally 1-3 * Deciding Set	0.0049
	(0.0103)
Rally 4-6 * Deciding Set	0.0017
	(0.0039)
Rally 7-9 * Deciding Set	-0.0145
	(0.0086)
Rally 10+ * Deciding Set	0.0255***
	(0.0112)

Furthermore, this model of rally length controlling for total points and third set explains 82% of the variation in winning. It is important to note this high R-squared is very relevant in explaining a predictive model of this type.

A massive limitation to this research comes from the collection method of the data. The collection of this data suffers selection bias, as individuals choose which matches they wish to chart. By not including every match of each tournament used, selection bias plays a role as matches chosen to be charted were not selected at random. Additionally, the number of matches varies greatly from year to year, many years have data for only a few matches, while, for example, year 2015 has as many as 38 matches. The other big limitation of this data comes from its small sample size. While one tennis match can deliver a very large number of statistics, collection of one match has proven to be difficult and shows in the number of charted matches available. With only 141 matches used in this study, results could be negatively effected.

Other limitations to this research comes from not knowing when certain rally lengths occurred. If controlling for the score and at which point a given rally occurred at within a game was a possibility, this paper could deliver much stronger results. An additional limitation comes from only looking at a single tournament on the hard court surface. While this allowed for court surface to be precisely controlled for, there are other tournaments besides the Australian Open that play on the same type of hard court surface. By using matches from those additional tournaments, the data set could be much larger.

V. Conclusion

The results of this study are consistent with the idea that long points of third set matches are import in resulting in a winning outcome. Ultimately, stamina and endurance in tennis are key factors in winning long and grueling matches. If the player is capable of winning long rallies of 10 or more shots, even in the scope of a third set match, they are much more capable of having the stamina to lead them to a victory. Small and medium length rallies appear to matter when a match is only two sets long, this is likely because two set matches are arguably less competitive than three set matches. From a less competitive match, it could be assumed that most points are much shorter and hold more importance than a more competitive three set match.

Furthermore, the importance of long rallies in three set matches can explain shifts in momentum. These could prove to be important in understanding wins, as momentum shifts that occur in a third set come in a crucial time towards the close of the match. As players battle in a highly competitive third set match, winning a long rally could provide them with the extra push of desire they need reach the winning outcome.

Future studies further examining the effects of rally length would benefit by looking into the effects of rally length on the following point compared to the overall match outcome. Not only would this dive into a much larger data set, but it would allow for a much clearer look into player performance after a long rally.

Table 1. Summary Statistics – Points by Rally Length

Variable	Mean	Std. Dev	Min	Max
Rally 1-3	6.66	5.13	0	24
Rally 4-6	36.70	14.87	13	88
Rally 7-9	16.08	6.64	5	40
Rally 10+	8.41	5.06	0	30
Total Points 1-3	13.84	10.10	2	50
Total Points 4-6	73.26	26.47	19	161
Total Points 7-9	32.82	13.07	11	88
Total Points 10+	16.33	8.19	2	58
Observations	141			

References

Barnett, T. & Clarke S. R. (2002) Using Microsoft Excel to model a tennis match. 6th Conference on Mathematics and Computers in Sports (G. Cohen ed.). Queensland, Australia: Bond University, pp. 63-68.

Barnett, T., & Clarke, S. R. (2005). Combining player statistics to predict outcomes of tennis matches. *IMA Journal of Management Mathematics, 16*(2), 113-120. doi:10.1093/imaman/dpi001

Burke, K. L., & Houseworth, S. (1995). Structural charting and perceptions of momentum in intercollegiate volleyball. *Journal Of Sport Behavior, 18*(3), 167.

Cutler, H. S., Meaike, J., & Colvin, A. C. (2016). The Effect of Court Surfaces on Injuries in Tennis: A Literature Review. *Medicine & Science In Tennis, 21*(3), 22-27.

Klaassen, F. J. G. M. & Magnus, J. R. (1998) Forecasting the winner of a tennis match. Eur. J. Oper. Res., 148, 257-267.

Klassen, F. J. G. M. & Mangus, J. R. (2001) Are points in tennis independent and identically distributed? Evidence from a dynamic binary panel data model. J. Am. Stat. Assoc, 96, 500-509.

Morris, C. (1997). The most important points in tennis. *Optimal Strategies in Sports*, Ladany, S.P., Machol, R.E. (Eds), North Holland Publishing Company, Amsterdam. pp. 131-140.

Sackmann, Jeff. “Tennis Match Charting Project.” *Tennis Match Charting Project*, Tennis Abstract, www.tennisabstract.com/charting/meta.html.

Sánchez-Moreno, J., Marcelino, R., Mesquita, I., & Ureña, A. (2015). Analysis of the rally length as a critical incident of the game in elite male volleyball. *International Journal of Performance Analysis in Sport*, *15*(2), 620-631.