What does Rally Length tell us about Player Characteristics in Tennis?

Nirodha Epasinghege Dona, Paramjit S. Gill and Tim B. Swartz *

Abstract

This paper proposes increasingly complex models based on publicly available data involving rally length. The models provide insights regarding player characteristics involving the ability to extend rallies and relates these characteristics to performance measures. The analysis highlights some important features that make a difference between winning and losing, and therefore provides feedback on how players may improve.

Keywords : Bayesian inference, rally length, ATP and WTA tours, tennis analytics.

*N. Epasinghege Dona is a PhD candidate and T. Swartz (email: tim@stat.sfu.ca) is Professor, Department of Statistics and Actuarial Science, Simon Fraser University, 8888 University Drive, Burnaby BC, Canada V5A1S6. P. Gill is Associate Professor, Department of Computer Science, Mathematics, Physics and Statistics, Irving K. Barber Faculty of Science, University of British Columbia Okanagan, 1177 Research Road, Kelowna BC, Canada V1V1V7. Swartz has been partially supported by the Natural Sciences and Engineering Research Council of Canada. The work has been carried out with support from the CANSSI (Canadian Statistical Sciences Institute) Collaborative Research Team (CRT) in Sports Analytics. The authors thank Jeff Sackmann who conceived The Match Charting Project for collecting and distributing tennis data at a very fine resolution. Thanks are also due to dozens of anonymous volunteers who spent countless hours collecting data. We thank the Associate Editor and two anonymous Reviewers for helpful comments that helped improve the paper.

1 INTRODUCTION

As is the case with many sports, tennis has seen an upsurge of work in analytics. One of the first serious statistical contributions to tennis analytics was the work by Klassen and Magnus (2001) which concerned an investigation of the iid assumption that points are independent and identically distributed. They concluded that there is a positive correlation between successive outcomes and that servers are less likely to win a point in important situations.

Baker and McHale (2014, 2017) used the classical Bradley-Terry model to compare the performance of tennis players. McHale and Morton (2011) employed the same model for the purpose of forecasting tennis matches. Kovalchik (2016) studied 11 forecasting models in tennis for the purposes of match prediction. Ingram (2019) subsequently proposed a point-based hierarchical model for predicting the outcome of tennis matches which outperformed the point-based models studied by Kovalchik (2016). Ingram (2019) introduced a binomial model where the probability of winning a serve is dependent on skill, time and playing surface. A comprehensive text on tennis analytics for Wimbledon is given by Klaasen and Magnus (2014).

More recently, tennis analytics has been assisted by the availability of tracking data. With tracking data, player and ball locations are recorded with high frequency (i.e. spatio-temporal data), and these detailed datasets have contributed to explorations of many sporting problems that were previously unimaginable. Gudmundsson and Horton (2017) provide a review on spatio-temporal analyses used in invasion sports where player tracking data are available. In tennis, there are a growing number of papers that provide statistical analyses and rely on tracking data. Tea and Swartz (2023) investigate serving tendencies which may allow a player to anticipate the nature of the opponent's serve. The approach relies on hierarchical models in a Bayesian framework. Kovalchik and Albert (2022) also used a Bayesian framework where they investigate serve returns by introducing a semiparametric mixture model. Other papers that use tracking data but focus exclusively on the serve include Mecheri et al. (2016) and Wei et al. (2015). The text by Albert et al. (2017) provides a flavour for sports statistics across major sports.

Whereas tracking data analyses are on the increase, tracking data are often proprietary. In this paper, we use publicly available non-tracking data to analyze an aspect of tennis that

does not seem to have been previously investigated. Specifically, we consider rally length as the response variable with the added information whether the last shot was touched or not touched (i.e. a winner). This simple and readily available information allows us to analyze the rally characteristics of players. Various models of increasing complexity are introduced where we consider player characteristics, the identification of the server and a reduction in serve advantage as the rally proceeds.

In Section 2, we describe the data and associated issues with the dataset. In Section 3, three models of increasing complexity are proposed which take into account increasingly realistic features of the sport of tennis. In Section 4, priors are introduced for the models and computation is discussed. We fit the models using large datasets based on the ATP (Association of Tennis Professionals) tour for men and the WTA (Women' Tennis Association) tour for women. We consider separate analyses for first and second serves. Insights are obtained for the various models. For example, we consider overall tennis characteristics and how these vary across the the men's and the women's games, and in first versus second serves. We then investigate serve and rally characteristics with respect to various players of interest. We observe how extending rallies is an important component of success in tennis. Model assessment and simulation techniques are also discussed. We conclude with a short discussion in Section 5.

2 TENNIS DATA

The data analysed in this paper is based on 947,821 and 422,776 serves from men's and women's professional tennis matches, respectively. The data were obtained from the Match Charting Project (https://github.com/JeffSackmann/tennis_MatchChartingProject) maintained by Jeff Sackmann. The data involve matches from 1970 through 2022 and contain shot-by-shot outcomes involving 710 distinct men and 489 distinct women. This public dataset provides information on shot type, shot direction, depth of returns, types of errors, and more. The data were collected by volunteers after watching the video recordings of matches. To the best of our knowledge, there is no other source of publicly-available data of this type. At present, these data appear to be under utilized for statistical modelling of tennis outcomes.

The data covers matches from all the major Grand Slam events, the Davis Cup and many minor tournaments. The best players of this time period are included in the dataset. Since the charting of matches was at the discretion of volunteers, data are highly skewed in favour of later years. Also, better (i.e. winning) players have a higher representation since they frequently reached the latter rounds of tournaments.

We have chosen to use limited and relatively simple data to facilitate modelling. As will be seen in Section 4, important tennis insights can be achieved with such data. The data collected for each point are of the form (T, I) where T is the number of touches leading to the point and I is an indicator function as to whether the server won the point. For example, suppose that the serve is in play and the receiver hit the serve out of bounds. In this case, there are two touches yielding $T = 2$ and $I = 1$. Alternatively, suppose that the serve is in play, the receiver returned the serve and the server then hit the ball into the net. In this case, there are three touches yielding $T = 3$ and $I = 0$. Note there is a technicality in the definition of (T, I). If $(T = 1, I = 0)$, then the event corresponds to a fault on a first serve where no point is awarded. However, $(T = 1, I = 0)$ corresponds to a point for the receiver on a second serve (i.e., a winner).

It may seem that the number of touches T is a nonstandard choice for a data variable. We have selected T over the related variable "rally length" (or rally count) since there appears to be some confusion over the definition of rally length. For example, some people refer to an ace as a rally of length zero while others refer to an ace as a rally of length one. We also found that the variable T is more intuitive for modelling purposes.

Data and code that were used in this manuscript have been made available and can be accessed as supplementary materials.

2.1 Data Management Issues

When analyzing data, it is obviously important that the data are accurate. Therefore, we carried out various procedures to check data accuracy.

In the MatchChartingProject dataset, rows correspond to points awarded. Therefore, we augmented the dataset to include all serves. For example, whenever a second serve occurred, this implied that there was a first serve that resulted in a fault, for which no point was awarded.

Also, it is expected that in a huge dataset involving volunteer coding, some mistakes occur in data entry. For example, in the rallyCount variable, there were five non-numeric characters out of more than 500,000 serves in the ATP data. We simply removed the corresponding rows when this occurred.

In our model formulation, we need to determine the number of touches T. However, in the MatchChartingProject dataset, T is not directly provided and is obtained through the rallyCount variable RC. We set $T = RC$ if the IsRallyWinner variable is TRUE, and we set $T = RC + 1$ if the IsRallyWinner variable is FALSE. According to Jeff Sackman, RC is the number of shots excluding errors.

3 MODELS

In this section, we present three models. The first model is simple, concise, and assumes common characteristics across all players. The model serves as a baseline case for the more realistic models to follow. The second model introduces the realism of individual player characteristics. The third model is a further extension, which distinguishes return characteristics according to whether a player is the server.

None of the models below distinguish first serves from second serves. However, since we have large datasets, we simply use the same model analyzed separately for the two situations with model parameters interpreted according to the serve number. We analyze first and second serves separately in Section 4 for both the men's and women's games.

3.1 Model 1

Although this model is not meant to be realistic, it introduces notation and provides a baseline case for comparison purposes. This model also permits straightforward generalizations for more complex models. Later, we use results from Model 1 to elicit prior information for more realistic Bayesian models.

We introduce a parameter vector (a, f, s, w, r, m) which describes player characteristics. There is an implicit assumption that all players have the same characteristics. Specifically,

we define

$a =$ Prob(player serves an *ace*)
$f =$ Prob(player serves a *fault*)
$s =$ Prob(player serves a ball that is neither an ace nor a fault)
$w =$ Prob(player returns a *winner*)
$r =$ Prob(player *returns* a ball in play which is subsequently touched by the opponent)
$m =$ Prob(player makes a *mistake* by touching but not returning a ball in play)

The parameterization includes service parameters a and f, critical components of tennis where an ace is defined by a serve that is in play, is not touched by the receiver, and therefore accrues a point to the server. The return parameters (w, r, m) are convenient since they describe all possibilities that can occur on a return that is touched, and consequently, $w + r + m = 1$. Quality of return capability is characterized by large w and small m.

We note that the proposed parameters have a different focus than standard statistics reported in the tennis literature. For example, with respect to the parameter a (aces), a commonly reported statistic is aces per match. A problem with aces per match is that players may not have the same number of average serves per match. A player with more serves (i.e. longer matches) will have an inflated aces/match statistic. Also, career aces favour players with longevity. Our proposed parameters and their estimates standardize performance with respect to the number of opportunities.

Consider then a first serve that results in a fault. In this case, neither player wins the point and we have

$$\text{Prob}(T = 1) = f \ . \tag{1}$$

Alternatively, consider a one touch event (i.e., $T = 1$) that may be either a first or second serve, but is not a first serve fault. In this case, we have

$$\text{Prob}(T = 1, I) = a^I \ f^{1-I} \ . \tag{2}$$

For two touches, a conditional probability expansion involving the two events yields

$$\text{Prob}(T = 2, I) = s\ m^I\ w^{1-I}$$

and, in general, for rallies with $t = 2, 3, \ldots,$ touches,

$$\text{Prob}(T = t, I) = \begin{cases} s\ r^{t/2-1}\ r^{t/2-1}\ m^I\ w^{1-I} & t \text{ even} \\ s\ r^{t/2-1/2}\ r^{t/2-3/2}\ w^I\ m^{1-I} & t \text{ odd} \end{cases} \tag{3}$$

Therefore, the likelihood based on the observed data is formed by taking the product over all serves using the expressions (1), (2) and (3). Model 1 has a parameterization that is 4-dimensional.

3.2 Model 2

We extend Model 1 by introducing a parameter vector $(a_i, f_i, s_i, w_i, r_i, m_i)$ for each player $i = 1, \ldots, N$. The vector describes playing characteristics of player i. The generalization is needed since some players, for example, are better servers than other players. For each $i = 1, \ldots, N$, the parameters satisfy the constraints in Model 1.

Without loss of generality, we assume that player i serves to player j. Consider then a first serve that results in a fault. In this case, neither player wins the point and we have

$$\text{Prob}(T = 1) = f_i\ . \tag{4}$$

Alternatively, consider a one touch event (i.e., $T = 1$) that may be either a first or second serve, but is not a first serve fault. In this case, we have

$$\text{Prob}(T = 1, I) = a_i^I\ f_i^{1-I}\ . \tag{5}$$

Extending (3) for specific players, for rallies with $t = 2, 3, \ldots,$ touches, we have

$$\text{Prob}(T = t, I) = \begin{cases} s_i\ r_j^{t/2-1}\ r_i^{t/2-1}\ m_j^I\ w_j^{1-I} & t \text{ even} \\ s_i\ r_j^{t/2-1/2}\ r_i^{t/2-3/2}\ w_i^I\ m_i^{1-I} & t \text{ odd} \end{cases} \tag{6}$$

Therefore, the likelihood based on the observed data is formed by taking the product over all tennis points using the expressions (4), (5) and (6), and by introducing appropriate subscripts i and j for specific players. With N players, Model 2 has a parameterization that is $4N$-dimensional.

3.3 Model 3

Although convenient, the Model 2 does not realistically account for differences in return characteristics according to the number of touches. For example, suppose again that player i serves to player j. It is well known in tennis that the probability of j hitting a winner on the immediate serve return (touch number two) is less than the probability of j hitting a winner on touch number four. The reason is that serves often place the returner in vulnerable situations where it is difficult for the returner to hit a quality shot. By the time player j reaches touches $4, 6, 8, \ldots$, the impact of the serve begins to dissipate.

Accordingly, we augment the return parameters (w_i, r_i, m_i) to $(w_i^{(t)}, r_i^{(t)}, m_i^{(t)})$ where the superscript $t = 2, 3, \ldots$, corresponds to the touch number. For example, we would expect that $w_i^{(2)} < w_i^{(4)} < w_i^{(6)}$ and that $w_i^{(3)} > w_i^{(5)} > w_i^{(7)}$.

With this modelling enhancement, the probabilities (4) and (5) remain the same. However, the probabilities in (6), for touches $t = 2, 3, \ldots$, become

$$\text{Prob}(T = t, I) = \begin{cases} s_i \left(\prod_{k=1}^{t/2-1} r_j^{(2k)} \; r_i^{(2k+1)} \right) \; (m_j^{(t)})^I \; (w_j^{(t)})^{1-I} & t \text{ even} \\ s_i \left(\prod_{k=1}^{t/2-1/2} r_j^{(2k)} \right) \left(\prod_{k=1}^{t/2-3/2} r_i^{(2k+1)} \right) \; (w_i^{(t)})^I \; (m_i^{(t)})^{1-I} & t \text{ odd} \end{cases} \tag{7}$$

We note that Model 3 as expressed in equation (7) increases the parametrization. With N players and with maximum number of touches $t_{\max}$, the extended parametrization in Model 3 is $(2 + 2(t_{\max} - 1))N$-dimensional. This increases the computations but the computing time remains manageable as discussed at the end of Section 4.3.

4 ANALYSES AND RESULTS

The analyses and results are provided according to the three models of increasing complexity. All of our models are developed in the Bayesian framework, and consequently

require the specification of prior distributions. We discuss prior selection and associated computational issues. We also provide a comparison of fit among the three models.

4.1 Analysis of Model 1

We begin with the simplest model which provides overall insights but later serves in prior development for more complex models.

For the Bayesian approach, we consider flat priors given by

$$(a, f, s) \sim \text{Dirichlet}(1, 1, 1) \tag{8}$$

which is assumed independent of

$$(w, r, m) \sim \text{Dirichlet}(1, 1, 1) \ . \tag{9}$$

Note that the Dirichlet distributions in (8) and (9) are 2-dimensional where $a+f+s = 1$ and $w + r + m = 1$. The parameter s is a characteristic of lesser interest and is not specifically investigated. To obtain posterior inferences, we use the programming language *Stan* (Stan Development Team 2023). A main benefit of *Stan* is that a user supplies only the statistical model, the prior specification and the data. The associated Markov chain Monte Carlo (MCMC) aspects of the Bayesian implementation are carried out in the background.

To assess the robustness of Bayesian inferences with respect to prior selection, we calculate classical estimates of the parameters that are based on proportions. Recall that T is the number of touches and $I = 1/0$ corresponds to the server/receiver winning the point. Let n be the corresponding number of serves and let t_k be the number of touches on the

kth serve. Then the sample proportions are given as follows:

$$\begin{aligned}
\hat{a} &= \frac{\#\{T=1, I=1\}}{n} \\
\hat{s} &= \frac{\#\{T>1\}}{n} \\
\hat{w} &= \frac{\#\{T=2, I=0\} + \#\{T=3, I=1\} + \#\{T=4, I=0\} + \cdots}{t_1 - 1 + t_2 - 1 + \cdots + t_n - 1} \\
\hat{m} &= \frac{\#\{T=2, I=1\} + \#\{T=3, I=0\} + \#\{T=4, I=1\} + \cdots}{t_1 - 1 + t_2 - 1 + \cdots + t_n - 1}
\end{aligned}$$

In Table 1, we provide estimated posterior means and proportions for Model 1. This is done for first and second serves using both the ATP and WTA data. We observe that there is agreement between the posterior estimates and the proportions which indicates that the information contained in the data dominates the prior. This is a consequence of having a rich dataset with many observations.

In Table 1, we observe some interesting features regarding first serves. It seems that aces occur in the men's game (0.08) at roughly double the rate observed in the women's game (0.04). This may be explained by the higher average speed of men's serves. The remaining displayed parameter estimates are comparable between men and women. We observe that first serve faults occur roughly forty percent of the time ($f = 0.40$ for men and $f = 0.37$ for women). This high rate may be explained by the desire to hit a first serve that is difficult to return - difficult serves are typically directed near the boundaries and are close to faults. If a player is able to get their racket on the ball, the return rate is high ($r = 0.74$ for men and $r = 0.76$ for women). And if a player is able to get their racket on the ball, the probability m of making a mistake is roughly 2.5 times the probability w of hitting a winner.

For second serves, the occurrences of aces and faults are much different from first serves. Aces are reduced in second serves because the server is less aggressive and wants to ensure that they do not double fault. Similarly, faults are greatly reduced with second serves for both men and women. In searching the literature, we could not find any reports of second serve ace percentage other than that second serve aces are extremely rare. Here,

we estimate second serve ace percentage where there are 1% second serve aces for men and 0.4% second serve aces for women.

When comparing men to women on second serves, the biggest parameter differences involve aces ($a = 0.01$ for men and $a = 0.004$ for women) and faults ($f = 0.09$ for men and $f = 0.13$ for women). The lower ace percentage for women compared to men may again be due to lower serve speed. However, the higher fault percentage for women compared to men is not so readily explained.

		Parameters				
Serve	n	a	f	w	r	m
ATP 1st	692385	0.08 (0.08)	0.40 (0.40)	0.07 (0.07)	0.74 (0.74)	0.19 (0.19)
WTA 1st	307750	0.04 (0.04)	0.37 (0.37)	0.07 (0.07)	0.76 (0.76)	0.17 (0.17)
ATP 2nd	255436	0.01 (0.01)	0.09 (0.09)	0.05 (0.05)	0.80 (0.80)	0.15 (0.15)
WTA 2nd	115026	0.004 (0.004)	0.13 (0.13)	0.06 (0.06)	0.77 (0.77)	0.17 (0.17)

Table 1: Estimated posterior means (sample proportions) for the parameters a, f, w, r, m corresponding to Model 1. The posteriors are based on flat priors. The sample size n for the number of serves is also reported.

Under Model 1, we can investigate how the sport of tennis has changed over time. In Figure 1, we plot the proportions corresponding to the parameters a, f, s, w, r, m as a function of the year. We observe that aces (a) are more probable in the modern game. In turn, the receiver's ability to touch the serve (s) is reduced. What is conspicuous in Figure 1 is the little downward hump during the years 1995-2000 in both the s and the r parameters. These parameters are related in that they express an aspect of return capability on the serve and on subsequent touches, respectively. We have investigated the history of technology changes and discussions of strategic changes in tennis, and we cannot offer a good explanation for this observed phenomenon.

4.2 Analysis of Model 2

Model 2 extends Model 1 by introducing player specific characteristics. That is, we extend the parameter vector (a, f, s, w, r, m) to $(a_i, f_i, s_i, w_i, r_i, m_i)$ for all players $i = 1, \ldots, N$. The modelling philosophy is that individual player characteristics arise from a population

Figure 1: Plots of the estimated parameters a, f, s, w, r, m from Model 1 as a function of time (year).

of player characteristics given by

$$(a_i, f_i, s_i) \sim \text{Dirichlet}(ka, kf, ks) \tag{10}$$

which is assumed independent of

$$(w_i, r_i, m_i) \sim \text{Dirichlet}(kw, kr, km) \tag{11}$$

where (a, f, s, w, r, m) are the posterior means obtained through the previous analysis of Model 1, and $k > 0$ is a specified constant. Hence, the approach is empirical Bayes where larger values of k impose greater knowledge through the prior distributions (10) and (11).

More specifically, larger values of k decrease the prior variance.

Our rationale for setting $k > 0$ is that we want players with longevity to have posterior parameter estimates that reflect their actual playing performance. On the other hand, for players with short professional careers who have served infrequently (say $n_i < 100$), we want their posterior estimates to revert closer to population averages. Accordingly, we have set $k = 200$ after some trial and error. Alternatively, we considered an approach with the hyperprior $k \sim \text{Normal}(200, 400)$. Under this specification, we obtained the posterior mean $\hat{k} = 164.7$.

The use of Dirichlet distributions as priors is natural when the parameters form a probability distribution. However, there is a limited literature on the specification of the Dirichlet hyperparameters. In one example, Lefkimmiatis, Maragos and Papandreou (2009) utilize Dirichlet priors in the context of photon-limited imaging where Poisson processes are modelled.

With more parameters in Model 2 than Model 1, computational demands increase. Running *Stan* with 2000 iterations using the men's first serve dataset requires 4.6 hours of computation on a laptop computer.

We are interested in the variability of the player characteristics for the four datasets (ATP 1st Serve, WTA 1st Serve, ATP 2nd Serve, WTA 2nd Serve). In Figures 2-4, we provide the boxplots of the posterior means of the parameters a_i, f_i and w_i for the four datasets. The main takeaway from these figures is that one can see the distributional form (e.g. variability and skewness) that is not readily apparent from Table 1. For example, from Figure 2, it is interesting that there is more variability in first serve ace percentage amongst men than amongst women.

With an understanding of the distribution of parameter estimates, we wish to observe which players distinguish themselves with respect to the player characteristics. In Tables 2-4, we highlight the top five players with respect to their posterior estimates a_i, f_i and w_i under the four datasets. From Table 2, we observe that the top five list of ATP players who serve high ace percentages (first serve) are familiar names. Reilly Opelka is 6 feet 11 inches tall and this affords a serve trajectory that can maximize serving speed. He is a current player who is not highly ranked (#643 in 2023). It will be interesting to see if his overall game catches up to his devastating serve. John Isner is likewise tall (6 feet 10 inches) and is renowned for his serve. Goran Ivanisevic was a left hander which is unusual and may

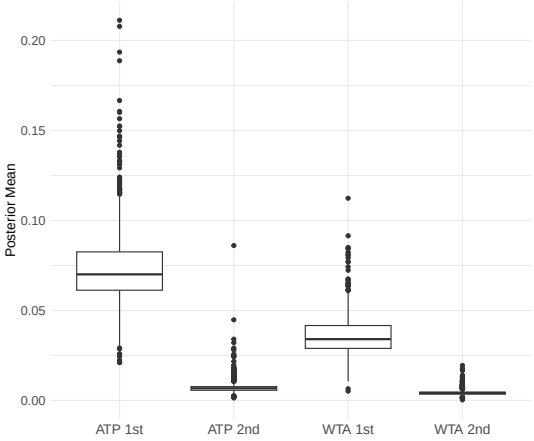

Figure 2: Boxplots of the posterior probability of an ace a_i by players for the four datasets.

have accounted for the difficulty that players experienced when returning his serve. It is noteworthy that four of the top five ATP players are still active in 2023 (Opelka, Isner, Kyrgios) or recently retired (Karlovic). Interestingly, three of the players from the first serve list reside on the second serve list. With second serves, Maxime Cressy is an unusual case with an ace percentage that is comparable to the mean first serve ace percentage. Cressy is renowned for his high risk/reward serve and volley style, and this is much discussed in tennis circles; see Imhoff (2022). It seems that Cressy has discovered something, and that he differs from the population of ATP professionals. In the women's game, Serena Williams sits atop the first serve ace percentage list. Similar to the men, four of the top five players with respect to first serve ace percentage (excepting Lucie Hradecka) are recent players. With WTP second serve ace percentage, there is not much to discuss since all players (even the top ones) have very few aces.

From Table 3, we first remind ourselves that low fault percentage is considered good. Generally speaking, the players in these lists are not specifically known for their low fault percentage. It is remarkable that John Isner appears on this list (ATP first serves) and also in Table 2 for first serve ace percentage. His appearance on both lists confirms that he is an outstanding server. During his playing days, Mats Wilander must have had outstanding serve control; he has low fault percentage on both first and second serves. For the women, we observe that Chris Evert (an iconic player) faulted infrequently on first serve. Given the

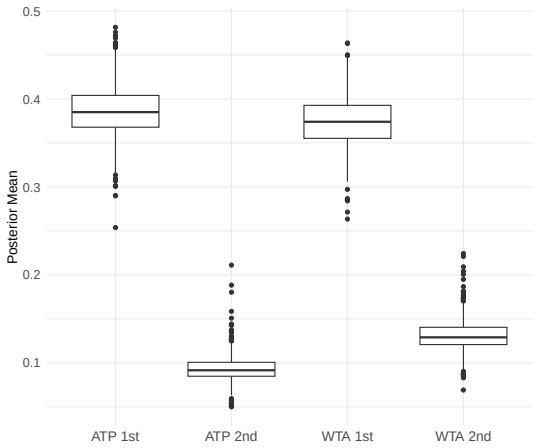

Figure 3: Boxplots of the posterior probability of a fault f_i by players for the four datasets.

ATP 1st Serve		WTA 1st Serve		ATP 2nd Serve		WTA 2nd Serve	
1. R.Opelka	(0.21)	1. S.Williams	(0.11)	1. M.Cressy	(0.08)	1. C.Harrison	(0.02)
2. I.Karlovic	(0.20)	2. M.Keys	(0.09)	2. I.Karlovic	(0.04)	2. J.Ostapenko	(0.02)
3. J.Isner	(0.19)	3. L.Hradecka	(0.09)	3. A.Bublik	(0.03)	3. C.Paquet	(0.02)
4. G.Ivanisevic	(0.18)	4. K.Pliskova	(0.08)	4. R.Opelka	(0.03)	4. S.Lisicki	(0.02)
5. N.Kyrgios	(0.17)	5. C.Vandeweghe	(0.08)	5. N.Kyrgios	(0.03)	5. J.Niemeier	(0.01)

Table 2: Estimated posterior means (Model 2) for the top five players in the four datasets for the probability of serving an ace a_i.

nickname "Ice Maiden", perhaps this reflected her ability not to succumb to the pressure of faulting.

From Table 4, we investigate the ability to hit winners during the rally. It is noteworthy here that most of the players who appear on the top lists are from an earlier era. Maybe this is a consequence of the game being faster today (especially with respect to the serve) and the consequent difficulty of hitting winners from faster shots. It is also noteworthy that Novak Djokovic widely regarded as the GOAT (greatest of all time) in men's tennis does not appear on this list nor in the previous tables. We posit that Djokovic has an all-around game that leads to his excellence. Prominent names that appear in Table 4 include Pat Rafter, Stefan Edberg, Rod Laver, Hana Mandilikova and Anna Kournikova. Since hitting winners during rallies is not a statistic that is routinely collected and analyzed, it good to

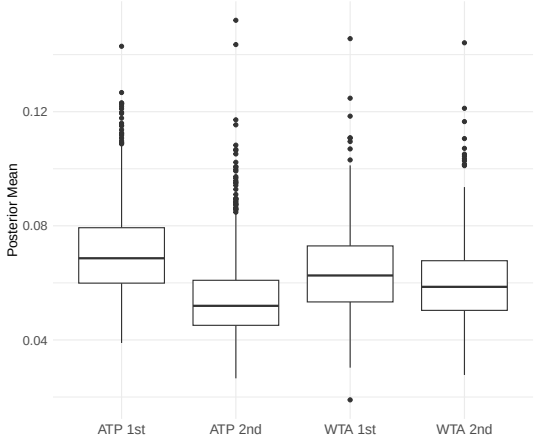

Figure 4: Boxplots of the posterior probability of a winning shot w_i by players for the four datasets.

ATP 1st Serve		WTA 1st Serve		ATP 2nd Serve		WTA 2nd Serve	
1. A.Berasategui	(0.25)	1. S.Errani	(0.26)	1. M.Wilander	(0.04)	1. A.Radwanska	(0.07)
2. M.Wilander	(0.28)	2. C.Evert	(0.27)	2. G.Forget	(0.05)	2. M.Keys	(0.08)
3. S.Baez	(0.29)	3. N.Parrizas Diaz	(0.28)	3. J.Brooksby	(0.05)	3. K.Juvan	(0.08)
4. J.Isner	(0.30)	4. A.Rus	(0.29)	4. A.Chesnokov	(0.05)	4. A.Sanchez Vicario	(0.09)
5. M.Cecchinato	(0.30)	5. M.Niculescu	(0.29)	5. M.Safin	(0.05)	5. K.Muchova	(0.09)

Table 3: Estimated posterior means (Model 2) for the top five players in the four datasets for the probability of committing a fault f_i.

see some excellent all-time players appearing in these lists. However, as mentioned with Djokovic, these individual parameters, studied on their own do not portend success. As we observe later in this section, it is the combination of skills which lead to success.

ATP 1st Serve		WTA 1st Serve		ATP 2nd Serve		WTA 2nd Serve	
1. K.Carlsen	(0.14)	1. K.Scott	(0.15)	1. R.Laver	(0.15)	1. K.Scott	(0.14)
2. K.Curren	(0.13)	2. J.Goerges	(0.12)	2. K.Carlsen	(0.14)	2. J.Ostapenko	(0.12)
3. P.Rafter	(0.12)	3. H.Sukova	(0.12)	3. R.Krajicek	(0.13)	3. J.Goerges	(0.12)
4. S.Edberg	(0.12)	4. S.Waltert	(0.11)	4. D.Brown	(0.12)	4. H.Mandlikova	(0.11)
5. R.Laver	(0.12)	5. J.Ostapenko	(0.11)	5. J.Siemerink	(0.11)	5. A.Kournikova	(0.10)

Table 4: Estimated posterior means (Model 2) for the top five players in the four datasets for the probability of hitting a winner w_i.

We have suggested that the game of tennis may have changed over time. With the various parameters introduced in Model 2 which characterize aspects of the game, we see that longitudinal tennis analyses of individual players are readily possible. For example, consider the ATP mistake parameter m_i on first serves. In Figure 5, we plot the posterior mean of m_i versus the year of entry into professional tennis for the corresponding player. There are several things to observe from Figure 5. First, there are more players in our dataset in recent years. This is a consequence of greater availability of recent recordings for charting. There also seems to be tighter variability in the mistake parameter m_i in recent years.

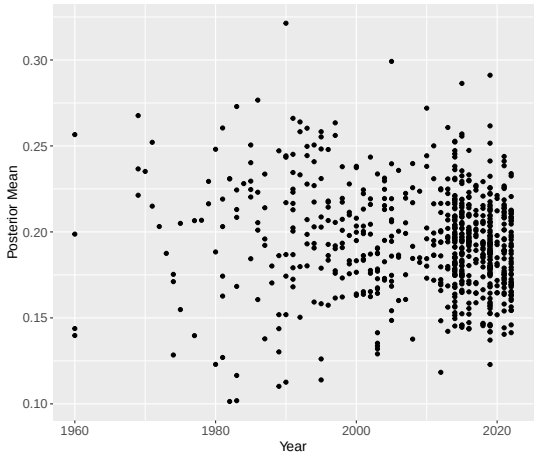

Figure 5: Scatterplot of the posterior mean of m_i for first serve ATP data versus the year that player began professional tennis.

We now investigate how the parameters a, f, w and m correlate individually to success. We collect the Official Pepperstone ATP player ranking points obtained from the ATP website at https://www.atptour.com/en/rankings/singles. There are 100 players on this list and we take the points corresponding to June 6/2022. These points are used to determine player rankings. For example, the 1st, 50th and 100th players on the list are Novak Djokovic, Roger Federer and Alexei Popyrin with 8770, 1030 and 626 points, respectively. We obtain sample correlation coefficients between the player ranking points and their posterior estimates. The results are reported in Table 5. The message again is that individual parameters do not correlate strongly with success. We do note that although the

correlations are small (i.e. close to zero), they tend to correlate in the correct directions. For example, high probabilities of aces a_i are beneficial, and here, the associated correlations are positive with respect to the first serve. We would expect to also see a positive correlation with winners w_i. However, this is not the case since winners occur infrequently (see Figure 4) and there is not much variability in winners across players. Also, as mentioned previously, the analysis of winners needs to be considered in tandem with touch number T. For example, winners on $T = 2$ are unlikely since this corresponds to returning a serve where the receiver is often off-balance. On the other hand, a winner on $T = 3$ is likely since the server retains some advantage from the serve.

Low probabilities of faults f_i and mistakes m_i are beneficial, and here, the associated correlations are negative with respect to the first serve. The correlations involving second serves are less interpretable where fewer aces and faults occur. Again, a limiting factor in the correlation study involving w_i and m_i is that touch number is not considered; this is remedied with Model 3 in Section 4.3.

Parameter	ATP 1st Serve	ATP 2nd Serve
a	0.11	-0.05
f	-0.25	0.01
w	-0.05	-0.14
m	-0.14	-0.22

Table 5: Correlation coefficients between 2022 ATP player ranking points and the posterior estimates of a, f, w, m for first and second serves.

Now, although the previous analyses are informative, it is obviously the case that player excellence is a function of various skills. To investigate the combination of skills, we carry out a regression analysis of the previously mentioned Pepperstone points against the variables $(a_{i1}, f_{i1}, w_{i1}, m_{i1}, a_{i2}, f_{i2}, w_{i2}, m_{i2})$ where the second subscript in the pair of subscripts refers to serve number. Retaining only the significant variables, we obtain the fitted equation

$$y = 7530 + (19832)a_1 - (14168)f_1 - (25817)m_2 \tag{12}$$

where the correlation between the fitted line and the Pepperstone points is an improved

$r = 0.50$. Table 6 provides the full regression results where we note that the parameters a_2 and f_2 are nearly significant.

Term	Estimate	Std Error	P-value
Constant	7530	1984	0.0003
a_1	19832	6799	0.004
f_1	-14168	5628	0.001
w_1	17933	20147	0.376
m_1	4964	12139	0.684
a_2	15401	8122	0.061
f_2	-43920	24273	0.074
w_2	-22918	24976	0.361
m_2	-25817	12223	0.037

Table 6: Full regression results for the fitted model presented in (12).

Equation (12) can assist in player evaluation. For example, suppose that a player has first serve ace percentage $a_1 = 0.08$. The player's points might be expect to rise by 199 points if they could increase their first serve ace percentage to $a_1 = 0.09$. However, it must be kept in mind that there are reasonable restrictions (see Figure 2) as to the extent that one might increase their first serve ace percentage. It is interesting that the fitted equation (12) contains the characteristics a_1, f_1 and m_2. Although the inclusion of a_1 and f_1 appear obvious, it is worth considering why mistakes occurring on the second serve m_2 are important whereas mistakes on the first serve m_1 are not important. Our data reveal that 72% of first serves involve $T = 3$ touches or less. When there are only $T = 3$ touches, this means that mistakes of the type m_1 can only be made on the second or third touches. After a powerful serve, the second touch (i.e. the receiver's return) will be difficult, and therefore, mistakes m_1 will be made at a high rate. Conversely, on the third touch, the server will continue to benefit from the serve as mistakes m_1 will occur at a low rate. These two phenomena will tend to cancel each other out, and this is why m_1 is not statistically significant. Again, the importance of touch number is considered in Section 4.3.

4.3 Analysis of Model 3

The enhancement of Model 2 to Model 3 involves the introduction of return characteristics w, r, m that vary according to touch number. With the player characteristics i, and the touch number t, this leads to parameters $w_i^{(t)}, r_i^{(t)}, m_i^{(t)}$. The motivation is that the effect of the serve dissipates as the rally progresses. We wish to quantify the effect.

We use the same prior distribution as given in (10) and (11) for each touch $t = 2, 3, ...$ We then average posterior estimates of $w_i^{(t)}, r_i^{(t)}, m_i^{(t)}$ over all players i to give varying return characteristics $w^{(t)}, r^{(t)}, m^{(t)}$. For the ATP and WTA datasets, we plot the estimates of the winner probability $w^{(t)}$ with respect to the touch number t in Figure 6. We do this for both first and second serves. For both men and women on first serves, we observe that the server retains an advantage in hitting a winner on touches $T = 3$ and $T = 5$. The advantage then quickly dissipates on subsequent touches for the server where the posterior probability of a winner approaches the prior mean probability. This is expected since there are fewer cases of long rallies. For example, the number of cases where $T = 15$ in the ATP first serve dataset is only $n = 2,245$. We also observe that as the rally continues, the probability of the server and the receiver hitting a winner is the same. On second serves, the lines are flatter. This is expected since second serves are more cautious and provide less of an advantage.

Analogous to Figure 6, we plot the estimates of the mistake probability $m^{(t)}$ with respect to the touch number t in Figure 7. Here, we observe that the server's mistake probabilities (t odd) increase only slightly with touch number. This seems intuitive as mistakes are exaggerated when a player is put in difficult situations, and there are no touch numbers where the server is consistently in distress. We also observe that mistake probabilities decrease for the receiver (t even) as the rally continues with stabilization around $T = 6$. This is the point at which the residual influence of the serve has dissipated. Again, the effects are less pronounced in second serve scenarios.

Under the higher parameterization associated with Model 3, computational demands increase. Running *Stan* with 2000 iterations using the men's first serve dataset requires roughly 18 hours of computation on a laptop computer.

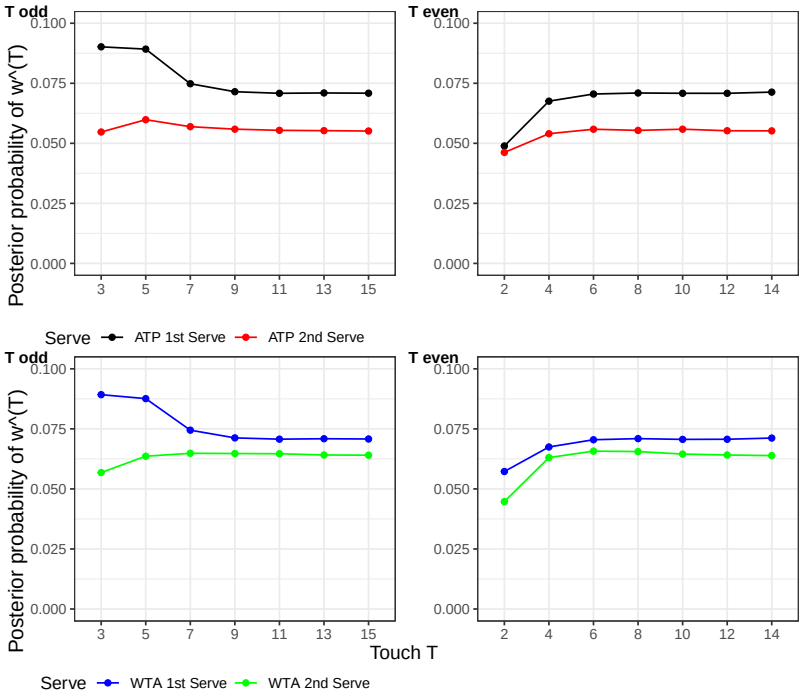

Figure 6: Plot of the posterior estimate $w^{(t)}$ versus t according to Model 3 with respect to the ATP data (top row plots) and WTA data (bottom row plots). The left plots concerns t odd (server) and the right plots concerns t even (receiver).

4.4 Model Assessment

We have introduced increasingly complex models that take into account the physical realities of tennis. However, we would like to see that statistical diagnostics confirm that Model 3 is better than Model 2 which in turn is better than Model 1.

In Table 7, we provide the Widely Applicable Information Criterion (WAIC) as described by Watanabe (2010) from running the three proposed models. WAIC is an appealing statistic since it penalizes model complexity; it therefore provides a balance between fit and parsimony. The WAIC approach requires the storage of log-likelihood values, and as a consequence, the fitting comparison is computationally intensive. We have therefore fit the models using only the most recent 10,000 observations from the ATP dataset based on first serves. Smaller values of WAIC provide evidence of better model fit. From Table 7, as expected, we observe that Model 3 is best, followed by Model 2 and then followed by

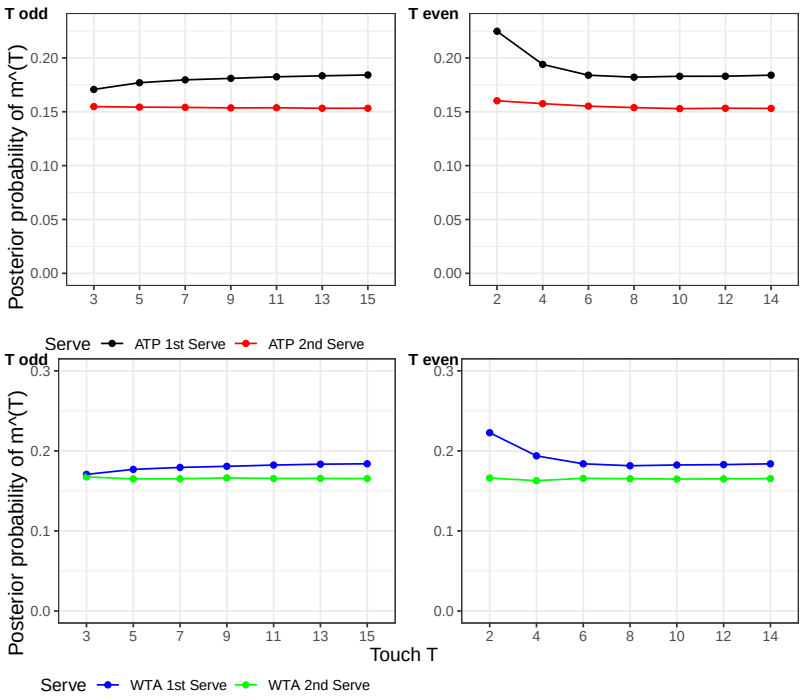

Figure 7: Plot of the posterior estimate $m^{(t)}$ versus t according to Model 3 with respect to the ATP data (top row plots) and WTA data (bottom row plots). The left plots concerns t odd (server) and the right plots concerns t even (receiver).

Model 1. The improvement between Model 3 and Model 2 is greater than the improvement between Model 2 and Model 1. This highlights the importance of recognizing that point probabilities change as the rally progresses.

In an MCMC application, it is important to assess practical convergence of the algorithm. For the most complex model (Model 3), we experimented with different initial values of the parameters. In each case, we obtained the same posterior estimates. In addition, the Gelman-Rubin $\hat{R}$ statistic (Gelman and Rubin 1992) exceeds 0.998 across all parameters and this is indicative of practical convergence. Gelman et al. (1996) consider the assessment of model fit using discrepancy measures involving the posterior predictive distribution.

Model	WAIC
Model 1	53869.7
Model 2	53208.6
Model 3	50813.1

Table 7: WAIC for the three models to assess fit.

4.5 Investigations via Simulation

A feature of our models is that they permit novel investigations of prediction between two players. With Model 3, we have a detailed description of how the game is played at the player level.

To illustrate the utility of Model 3, we have simulated 10,000 matches between Novak Djokovic and Rafael Nadal, two current and prominent players on the ATP tour. Djokovic has often been referred to as the "GOAT" of tennis, the greatest of all time. Nadal has been a main rival of Djokovic during the Djokovic era. We have used the estimated parameters corresponding to Djokovic and Nadal for both the first and second serves, and have coded functions to simulated outcomes at the point level, the game level, the set level and the match level taking into account the scoring system in tennis. In these simulations, we have determined the match winner as the first player to win two sets.

Although investigations are only limited by our imagination, in Table 8, we provide various predictions based on the simulation. We also report what happened in reality based on the 50 matches in our dataset that took place between Djokovic and Nadal during the period 2006-2022. We observe that Djokovic and Nadal are closely matched where a slight advantage is given to Djokovic. In the simulation, Djokovic's point advantage propogates to a greater advantage in games, to a greater advantage in sets, and to a greater advantage in matches. The prediction results also agree closely with actual results. The exception to this involves the prediction of points won. We believe that the points won predictions are too high for both Djokovic and Nadal. The reason is that estimation was based on results against many opponents, and hence, parameter estimates may be thought of as estimates against an average opponent. In the case of Djokovic's points against Nadal, Nadal is not an average opponent. For example, with Djokovic serving, there would be occasions where Djokovic may hit a winner against an average opponent, but Nadal would be able to reach

the shot. We believe that what occurs on touches $t = 1, 3, 5, \ldots$ is more influential in terms of points won than what occurs on touches $t = 2, 4, 6, ...$ due to the decreasing importance of the serve as the touch number t increases.

Prediction Event	Djokovic Prediction	Djokovic Reality
Points won - Djokovic serving	73.4%	62.5%
Points won - Nadal serving	27.8%	39.4%
Games won	51.1%	50.7%
Sets won	54.3%	50.4%
Matches won	55.1%	52.0%

Table 8: Prediction percentages between Djokovic and Nadal based on simulations involving Model 3. The third column is the observed percentage from the 50 actual matches between the two players.

Via the simulation, we also investigate various exotic predictions between Djokovic and Nadal that are not routinely discussed in the media. For example, the simulation model predicts that 24.5% of sets will lead to tiebreaks. We also predict rally length distributions of $P(T = 1) = 0.31$, $P(T = 2) = 0.18$, $P(T = 3) = 0.08$, and $P(T \geq 4) = 0.43$ which we compare against observed proportions 0.32, 0.15, 0.11, and 0.42, respectively. It is evident that Djokovic and Nadal have longer rallies than what one normally sees in men's tennis (Carboch, Placha and Sklenarik 2018).

We note that the simulation procedure also has practical value in "what-if" scenarios for training. A player could predict their probability of winning a match against an average opponent. Then, the player can investigate what might happen if through training, they were able to improve some of their playing characteristics as expressed via the parameters. It would be a simple matter to re-run the simulation, changing some parameters to new values.

5 DISCUSSION

Using a massive dataset of coarse observations in tennis, we have proposed models that describe various features of the sport. These features are explored with respect to both the

men's and women's games, and with respect to both the first and second serves.

An instructive aspect of this research is that players have their own playing characteristics a and f corresponding to the serve, and their characteristics $w^{(t)}$ and $m^{(t)}$ corresponding to touch $T = t$. These probabilities summarize all components of a player's game, and can be viewed with the purpose of evaluation with respect to average player performance.

Also, given these estimated parameters, one can use simulation techniques to investigate all sorts of interesting questions. For example, one may simulate matches between two competitors to obtain the probability of a match lasting beyond two sets. As another example, one may be interested in the probability that a player breaks service in a particular match. With parameters describing inter-related aspects of tennis, the models provide great scope for analysis both at the game level and the player level.

In terms of future directions, it would be interesting to see how characteristics vary across different surfaces (e.g., grass, clay and hardcourt). With fewer matches (and data) on grass, one may extend the proposed hierarchical models where grass parameters are related to hardcourt parameters, for example. More work could also be done on identifying aspects of the game which make players truly exceptional. In this paper, the framework has been proposed to address extended questions such as these that have not been previously investigated. It may also be possible to consider a single model rather than separate models for first and second serves. Here, we imagine the introduction of a correlation structure between each of the model parameters corresponding to first and second serves.

A further avenue for model enhancement concerns the parameters which were estimated based on a player's history against many opponents. In Model 2 and Model 3, these parameters have a subscript i corresponding to the player of interest who is involved in the shot. However, it is apparent that these probabilities are also impacted by i's opponent (see Table 6). It would be useful to incorporate this feature in enhanced future models.

It may also be possible to introduce physical constraints in Model 3 which we believe to be true. For example, we may be able to impose $w_i^{(t)} < w_i^{(t+2)}$ and $m_i^{(t)} > m_i^{(t+2)}$ for t even, and $w_i^{(t)} > w_i^{(t+2)}$ and $m_i^{(t)} < m_i^{(t+2)}$ for t odd. However, the implementation of such constraints in Stan may not be straightforward.

Finally, for simulation purposes, we are typically interested in the current form of a player. It may be useful to extend the model so that more weight is placed on recent matches. Such models be employed to investigate the longitudinal profile of a player.

6 REFERENCES

Albert, J.A., Glickman, M.E., Swartz, T.B. and Koning, R.H., Editors (2017). *Handbook of Statistical Methods and Analyses in Sports*, Chapman & Hall/CRC Handbooks of Modern Statistical Methods, Boca Raton.

Baker, R.D. and McHale, I. (2014). A dynamic paired comparisons model: Who is the greatest tennis player? *European Journal of Operational Research*, 236 (2), 677-684.

Baker, R.D. and McHale, I. (2017). An empirical Bayes model for time-varying paired comparisons rankings: Who is the greatist women's tennis player? *European Journal of Operational Research*, 258 (1), 328-333.

Carboch, J., Placha, K. and Sklenarik, M. (2018). Rally pace and match characteristics of male and female tennis matches at the Australian Open 2017. *Journal of Human Sport and Exercise*, 13(4), 743-751.

Gelman, A., Meng, X.-L. and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6, 733-807.

Gelman, A. and Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, 7, 457-472.

Gudmundsson, J. and Horton, M. (2017). Spatio-temporal analysis of team sports. *ACM Computing Surveys*, 50(2), Article 22.

Imhoff, D. (2022). Cressy's 10-year master plan: to make serve-and-volleying cool again. *Tennis Australia*, Accessed January 9, 2024 at https://ausopen.com/articles/features/cressys-10-year-master-plan-make-serve-and-volleying-cool-again

Ingram, M. (2019). A point-based Bayesian hierarchical model to predict the outcome of tennis matches. *Journal of Quantitative Analysis in Sports*, 15(4), 313-325.

Klassen, F.J.G.M. and Magnus, J.R. (2001). Are points in tennis independent and identically distributed? Evidence from a dynamic binary panel model. *Journal of the American Statistical Association*, 96(454), 500-509.

Klassen, F. and Magnus, J.R. (2014). *Analyzing Wimbledon: The Power of Statistics*, University Press Scholarship Online, https://doi.org/10.1093/acprof:oso/9780199355952.001.0001

Kovalchik, S.A. (2016). Searching for the GOAT of tennis win prediction. *Journal of Quantitative Analysis in Sports*, 12(3), 127-138.

Kovalchik, S.A. and Albert, J. (2022). A statistical model of serve return impact patterns in professional tennis. Accessed October 13, 2022 at https://arxiv.org/abs/2202.00583

Lefkimmiatis, S., Maragos, P. and Papandreou, G. (2009). Bayesian inference on multiscale models for Poisson intensity estimation: Applications to photon-limited image denoising. *IEEE Transactions on Image Processing*, 18(8), 1724-1741.

McHale, I. and Morton, A. (2011). A Bradley-Terry type model for forecasting tennis match results. *International Journal of Forecasting*, 27 (2), 619-630.

Mecheri, S., Rioult, F., Mantel, B., Kauffmann, F. and Benguigui, N. (2016). The serve impact in tennis: First large-scale study of big hawk-eye data. *Statistical Analysis and Data Mining*, 9(5), 310-325.

Stan Development Team (2023). *Stan Modeling Language User's Guide and Reference Manual*, Version 2.32. https://mc-stan.org

Tea, P. and Swartz, T.B. (2023). The analysis of serve decisions in tennis using Bayesian hierarchical models. *Annals of Operations Research*, 325(1), 633-648.

Watanabe, S. (2010). Asymptotic equivalance of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11, 3571-3594.

Wei, X., Lucey, P., Morgan, S., Carr, P., Reid, M. and Sridharan, S. (2015). Predicting serves in tennis using style priors. *Proceedings of the 21th ACM SIGKDD International Conference on Discovery and Data Mining*, 2207-2215.

7 Data Availability

Data and code have been made available in the supplementary files.

8 Funding Statement

The research was not funded by any source.

9 Conflict of Interest Statement

There is no conflict of interest associated with this paper.