Journal of the Royal Statistical Society Series A, 2022, 1–13
doi: DOI HERE
Advance Access Publication Date: Day Month Year
Paper

Using Shot Location Transitions to Learn Tennis Player Strategies

Tyler W. Ward,[1] Garritt L. Page ,[2]* Gilbert W. Fellingham[2] and Alejandro Jara[3]

[1]Systems & Technology Research, Arlington, Virginia, USA, [2]Department of Statistics, Brigham Young University, Provo, Utah, USA and [3]Department of Statistics, Pontificia Universidad Católica de Chile, Santiago, Chile

*Corresponding author. page@stat.byu.edu

FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

It is common that tennis players set up a winning shot via a calculated previous shot. To study this phenomena we employ data from five Grand Slam matches that inclue Roger Federer, Rafael Nadal, Novak Djokovic, and Juan del Potro on the three major surfaces (hard court, clay, and grass) in order to infer shot location based on a tennis court that has been divided into six zones in a novel way. Exploratory analysis alone shows player-specific shot location preferences based on receipt location before a given shot. However, we propose a Bayesian hierarchical model and testing scheme that allows for greater understanding of player shot location preference through posterior distributions on each shot location probability estimate. Specifically, we use a Multinomial-Dirichlet model to estimate expected shot locations along with corresponding uncertainty quantification. We also utilize simultaneous posterior probability estimation methods to perform sequential testing and identify differences in player behavior over each surface, receipt location, and shot location.

1. Introduction

Interest in sports analytics has exploded in the past few years. This is particularly true for popular team sports like soccer, baseball, basketball, and American football. However, other popular sports like tennis whose number of fans is surpassed only by soccer, cricket, hockey, and boxing are now garnering much more interest (Tendu et al., 2023). For example, in tennis, work dedicated to understanding how well players are able hit the ball in different scenarios and predicting their shot location has appeared (see, e.g., Giles et al. 2020) and Kovalchik et al. (2020) use sport tracking data to estimate the value of a shote in real time.

Work dedicated to understanding strategies employed by tennis players with regards to shot location is in its infancy. Cornman et al. (2017) and Wilkens (2021) predict tennis match winners in hopes of beating current sports betting odds. Wei et al. (2016) focus on predicting point winners using random forests and a type of K-means clustering. Work to assess existing predictive models for tennis has been performed by Kovalchik (2016) who tests the predictive performance of 11 published forecasting models for predicting the outcomes of 2395 singles matches during the 2014 season of the Association of Tennis Professionals (ATP) Tour. Further work has focused on the distribution of rally lengths in pro tennis matches using data from a crowd sourced tennis data collection effort called "The Match Charting Project"[1]; Specifically, work by Lisi and Grigoletto (2019) analyze factors contributing to the length of a tennis match and Lisi et al. (2023) present the zero-one-modified Geometric distribution as a high-performing method for estimating tennis rally lengths. There are also other studies focused on analyzing serve placement for winning aces (Whiteside and Reid 2017), using computer vision methods to predict player movement (Giles et al. 2020), and Kovalchik and Albert (2022) use mixture models to study the spatial distribution of return location. Evidence of growing interest in player behavior is shown by Kaggle competitions exploring datasets related to professional tennis matches, with one participant even building a model predicting the "ExpectedIn" probability of given shots in the 2019 Australian Open final between Rafael Nadal and Novac Djokovic (Mehra 2023). Whiteside et al. (2017) and Ganser et al. (2021) presented classification of shot type (serve, rally forehand, slice forehand, forehand volley, rally backhand, slice backhand, backhand volley, smash, or false positive) based on analysis of inertial measurement unit data using cubic-kernel support vector machines.

[1] https://www.tennisabstract.com/charting/meta.html

In terms of published work predicting within-rally shot location, Wei et al. (2013a) present research using Hawk-Eye data from the 2012 Australian Open, that employs Gaussian Mixture Models to estimate the distribution of shot location. Presented at the 2013 MIT Sloan Sports Analytics Conference, Wei et al. (2013b) focus both on "in-point" prediction and correctly predicting the location of winners and errors. Very recently, Dona et al. (2024) analyze the rally characteristics of players.

The aim of the current research is three fold. First, we aim to show that registering areas where the ball is received within a rally, instead of tracking the exact ball location, can still provide important information that can be used to discover tennis player strategies. Second, we propose a modeling strategy that can be employed to extract the relevant information from these data. Finally, we customize a flexible inference approach detailed in Held (2004), which permits making formal inference statements regarding different strategies that players adopt depending on opponent and court surface. An important advantage of what we propose is that the data registration process can be carried out with simpler and less expensive technology, in comparison to the technology requirements needed to precisely track the exact location of each shot. For instance, systems like Hawk-Eye use multiple cameras to triangulate the ball's position, which involves significant investment in equipment and maintenance. In addition, high-speed cameras, sensors, and sophisticated software are needed and as a result, the data are rarely publicly available. In contrast, registering the general areas where the ball is received can be done manually by spectators or coaches that are either sitting in a good location at the match or from video footage. Admittedly some precision is lost compared to Hawk-Eye, but data collected using our approach is more readily available and, as will be shown, still reliably uncovers meaningful spatial and tactical patterns in player behavior, even when accounting for potential misclassifications introduced during the manual data collection.

The rest of the article is organized as follows. Section 2 contains a description of the data utilized while Section 3 details the statistical model we employ to estimate the probabilities of hitting to particular locations. In Section 4 we detail results from the model fit and the testing procedure. We provide some conclusions in Section 5.

2. Zones, Data Collection, and Matches

We propose partitioning a tennis court into six zones as seen in Figure 1. This partitioning is motivated by the angles at which tennis players tend to hit ground strokes during a rally when they are hitting from the center of their baseline. Hitting the ball at these angles is motivated by the desire to "open the court" which typically results in gaining advantage over ones opponent. Other court partitions have been proposed in the literature (Chan et al. 2022) and can be used within our general framework, but a contribution we make is to show that the angle motivated partition of the tennis court is able to identify strategy patterns. Using zones rather than precise ball coordinates permits answering questions like the following. Given a player received the ball on their right-hand side (Z2), how does the probability they will return the ball across the court to the right-hand side of their opponent (Z2) compare with the probability they will hit the ball down the line straight across the net (Z5)? Do shot propensities change when two players compete on clay versus hard court? Additionally, collecting data in this way will aid in evaluating player predictability and as a result help inform opponent-specific strategies. For example, if the probability that a player hits to Z2, Z3, Z4 and Z5 are approximately the same, then strategizing against this player requires more care than if one zone (say, Z5), has a markedly higher probability than the rest. Registering data in zones is

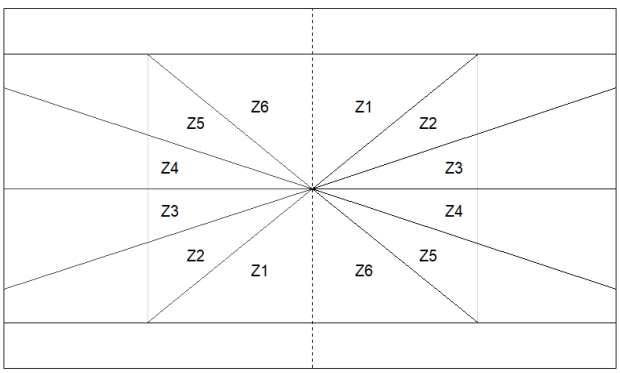

Fig. 1: Tennis court split into six zones on either side of the court. The dashed vertical line corresponds to the net.

less granular than exact, spatially referenced, shot location data. But no special tools are needed to collect these data so, for example, coaches or spectators can collect these data technology free. Further, since shots are are never hit at precisely the same location, it would be challenging to estimate probabilities of shots being hit to a specific location given that it was received from a specific location without aggregating over regions. A possible consequence of collecting data without technology is an increase in misclassifications (i.e., a ball is received in Z1, but is recorded as Z2). However, we found that these classifications are minor (never is Z1 classified as Z6) and our modeling approach seems to be fairly robust to this.

To illustrate the type of patterns we are able to extract/estimate we consider three classic professional tennis matches between Roger Federer and Rafael Nadal: the 2005 French Open Semi-Final (clay), the 2009 Australian Open Final (hard court), and the 2019

Wimbledon Final (grass). In addition, as a means to illustrate the diversity in strategies, we also consider the US Open 2009 finals match between Federer and Juan del Potro along with the Nadal vs Novak Djokovic's 2018 Wimbledon match. The later two are summarized in the online supplmentary material. The data from all the matches were manually collected by individuals with experience in tennis. Data collection consisted of watching each match shot-by-shot and recording the zone where the ball landed as it approached a player (`Ball.lands`), the zone to which the player hit the ball (`Ball.hit.to`). Table 1 shows a small sample of this data for the French Open final between Federer and Nadal.

Table 1. Example Data from 2005 French Open Final

Player	Ball lands	Ball hit to	In out	Rally Number
Federer		Z1	In	1
Nadal	Z1	Z4	In	1
Federer	Z4	Z5	In	1
Nadal	Z5	Z2	In	1
Federer		Z5	In	2
Nadal	Z5	Z2	In	2
Federer	Z2	Z4	In	2
Federer		Z2	In	3
Nadal	Z2	Z3	In	3
Federer	Z3	Z5	In	3
Nadal	Z5	Z5	In	3
Federer	Z5	Z2	In	3
Nadal	Z2	Z2	In	3
Federer	Z2	Z5	Long	3

These data were pre-processed before analysis in the following way. We remove serves. We include all points where a shot had a receipt location (`Ball.lands`) and a hit to location, even if the shot went wide or long and was called out. Shots hit into the net, however, are removed. A partial summary of empirical percentages that resulted from the data collection and preprocessing for Rafael Nadal in the Australian open are provided in Figure 2. Note how Rafael Nadal (a left-handed hitter) hit to Z5 more often than other zones when he received the ball on his far left side in Z5. However, his shot placement choices in this match against Roger Federer (a right-handed player) were much more uniform when he received the ball on his far right in Z2. Thus, there is evidence of trends that can be helpful for estimating shot location probabilities given certain characteristics of a point.

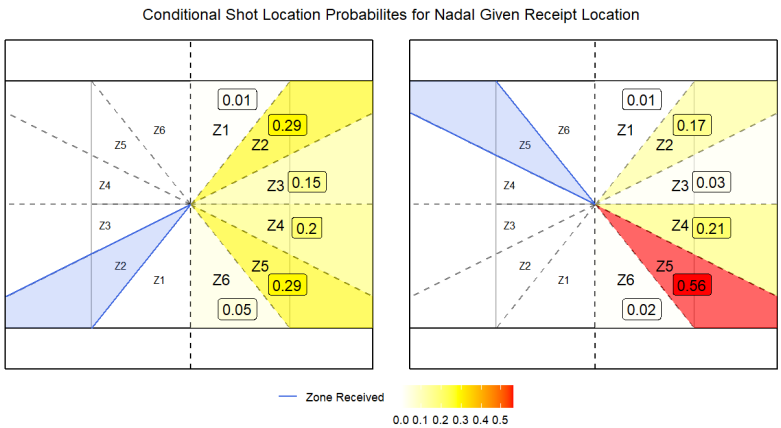

Fig. 2: Empirical conditional shot location proportions for Rafael Nadal from the 2009 Australian Open Final against Roger Federer. The blue zone is that from which Nadal received the ball. Since Nadal is left-handed, raquet swings in Z4-Z6 are typically forehands and for Federer, who is right-handed, they are typically backhands.

An additional feature to note from Figure 2 is the small percentages in zones one and six. Table 2 provides the counts of shots hit to each zone (see Figure 1) by each player in the three Federer vs. Nadal matches we consider (2005 French Open Semi-Final, 2009 Australian Open Final, and 2019 Wimbledon Final). These counts become even more sparse when considering the zone at which the ball was received. To see this, Table 3 displays counts from the French Open match. Notice that for some receipt (`Ball.lands`) locations there were no observations for certain hit locations resulting in very sparse data. Thus, we chose to limit our analysis to consider only

zones two through five. In our notation in Section 3 we specify shot locations of interest $k = 1, \ldots, 4$, with $k = 1$ corresponding with Z2, $k = 2$ corresponding with Z3, $k = 3$ corresponding with Z4, and $k = 4$ corresponding with Z5.

Table 2. Counts of shots hit to each zone by player for a single match played on each surface.

Player	Surface	Z1	**Z2**	**Z3**	**Z4**	**Z5**	Z6	Total
Federer	Grass (Wimbledon)	5	**111**	**100**	**59**	**78**	8	361
	Clay (French Open)	5	**186**	**141**	**77**	**109**	4	522
	Hard Court (Australian Open)	18	**229**	**170**	**118**	**189**	9	733
Nadal	Grass (Wimbledon)	8	**75**	**63**	**76**	**118**	15	355
	Clay (French Open)	9	**129**	**94**	**142**	**134**	14	522
	Hard Court (Australian Open)	9	**160**	**79**	**134**	**317**	19	718

3. Statistical Model and Testing Procedure

We now detail the Bayesian multinomial logistic regression model used to fit the data, followed by a methodology we used to test for differences in the shot location behavior of these two players. The parameters of interest include the probabilities that a ball is hit to each location on the court, given receipt location and court surface being played on. Our model, which includes considerations for receipt location, player, and surface, yielded interesting results that help further understanding about player behavior and expected shot locations, and highlight difference's between Federer and Nadal, discussed in detail in Section 4.

3.1. Model

Before detailing our model, we introduce the notation that will be used. Recall that due to low counts we only consider zones 2, 3, 4, and 5 (as discussed in Section 2). With this is mind, let y_{kspr} denote the number of shots hit to location k ($k = 1, 2, 3, 4$, corresponding to Z2, Z3, Z4, Z5, respectively) on surface s (Hard Court = 1, Clay = 2, Grass = 3) by player p (Federer = 1, Nadal = 2) after receiving in location r ($r = 1, 2, 3, 4$, again corresponding to Z2, Z3, Z4, Z5). Further, let the vector $\boldsymbol{y}_{spr} = (y_{1spr}, y_{2spr}, y_{3spr}, y_{4spr})$ be the corresponding counts associated with the four shot locations. To estimate the probability of a player hitting a shot to a given location during a match we model $\boldsymbol{y}_{spr}$ using the following hierarchical Bayesian Multinomial-Dirichlet model

$$\boldsymbol{y}_{spr} | \boldsymbol{\pi}_{spr} \overset{ind}{\sim} \text{Multinomial}(n_{spr}, \boldsymbol{\pi}_{spr}), \text{ for } s = 1, 2, 3;\ p = 1, 2;\ r = 1, \ldots, 4 \tag{1}$$

$$\boldsymbol{\pi}_{spr} | \boldsymbol{\pi}_{0sr}, m_{sr} \overset{ind}{\sim} \text{Dirichlet}(m_{sr} \times \boldsymbol{\pi}_{0sr}) \tag{2}$$

$$\boldsymbol{\pi}_{0sr} \overset{ind}{\sim} \text{Dirichlet}(\boldsymbol{\alpha}),\ \boldsymbol{\alpha} = (1, 1, 1, 1) \tag{3}$$

$$m_{sr} \overset{iid}{\sim} \text{Gamma}(1, 1), \tag{4}$$

where

$$\boldsymbol{\pi}_{spr} = (\pi_{1spr},\ \pi_{2spr},\ \pi_{3spr},\ \pi_{4spr})\,. \tag{5}$$

In more detail, we model the number of shots hit to each of the four locations (y_{kspr}) using a Multinomial distribution (see Equation (1)), with the fixed total number of shots hit to all locations on surface s by player p from receipt location r defined as n_{spr}, and the vector of probabilities of hitting the four zones ($\boldsymbol{\pi}_{spr}$) as defined in Equation (5). We employ a Dirichlet prior for $\boldsymbol{\pi}_{spr}$ to preserve the constraint that probabilities illustrated in Figure 3 sum to one. The parameter m_{sr} is a constant that permits $\boldsymbol{\pi}_{0sr}$ to directly inform $\boldsymbol{\pi}_{spr}$ in a flexible way.

Note how this model harnesses the hierarchical structure for the two players to "borrow strength". That is, the model uses information across observations from both players to reduce lower level parameters' sensitivity to noise (see Hoff (2009)). This is done as overall probabilities $\boldsymbol{\pi}_{0sr}$ are drawn independently from a Dirichlet($\boldsymbol{\alpha}$) prior, and then multiplied by a constant m_{sr} for a given surface and receipt location, allowing for player specific variation from $\boldsymbol{\pi}_{0sr}$. Therefore, $\boldsymbol{\pi}_{spr}$'s model borrows strength, as all $\boldsymbol{\pi}_{0sr}$ are drawn from the same distribution, but then still have player specific effects after the constant m_{sr} is included. We use a non-informative prior for $\boldsymbol{\pi}_{0sr}$, such that $\boldsymbol{\alpha} = (1, 1, 1, 1)$. This will allow the data to drive inference associated with $\boldsymbol{\pi}_{0sr}$ and $\boldsymbol{\pi}_{spr}$.

In order to compare π_{kspr} between two players, we use the odds ratio which is shown in general form for a given shot location k, surface s, receipt location r in Equation (6). The vector of odds ratios $\boldsymbol{\gamma}_{sr}$, defined in Equation (7), is therefore the four-dimensional vector of odds ratios for Federer (p=1) versus Nadal (p=2) for all of the shot locations ($k = 1, \ldots, 4$) given a certain surface and receipt location.

Table 3. Shot Locations Given Receipt Location for both Nadal and Federer from their 2005 French Open Semi-Final match.

	Ball hit to	Z1	Z2	Z3	Z4	Z5	Z6
Ball lands	Z1	**0**	**2**	**2**	**4**	**2**	**0**
	Z2	**5**	87	73	49	49	**8**
	Z3	**2**	84	63	65	58	**4**
	Z4	**3**	74	45	35	60	**2**
	Z5	**4**	65	49	63	67	**3**
	Z6	**0**	**3**	**3**	**3**	**7**	**1**

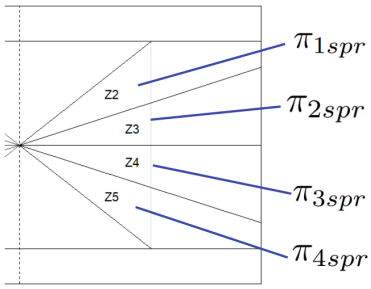

Fig. 3: Illustration of shot location probabilities $\boldsymbol{\pi}_{spr}$ corresponding to each of the considered shot locations, Z2, Z3, Z4, and Z5, respectively.

$$\gamma_{ksr} = \frac{\dfrac{\pi_{ks1r}}{(1-\pi_{ks1r})}}{\dfrac{\pi_{ks2r}}{(1-\pi_{ks2r})}} = \frac{O_{\text{Federer}}}{O_{\text{Nadal}}} \tag{6}$$

$$\boldsymbol{\gamma}_{sr} = \left(\frac{\dfrac{\pi_{1s1r}}{(1-\pi_{1s1r})}}{\dfrac{\pi_{1s2r}}{(1-\pi_{1s2r})}}, \frac{\dfrac{\pi_{2s1r}}{(1-\pi_{2s1r})}}{\dfrac{\pi_{2s2r}}{(1-\pi_{2s2r})}}, \frac{\dfrac{\pi_{3s1r}}{(1-\pi_{3s1r})}}{\dfrac{\pi_{3s2r}}{(1-\pi_{3s2r})}}, \frac{\dfrac{\pi_{4s1r}}{(1-\pi_{4s1r})}}{\dfrac{\pi_{4s2r}}{(1-\pi_{4s2r})}} \right) \tag{7}$$

3.2. Posterior Sampling

Due to the hierarchical structure for the two players, the joint posterior distribution of model parameters is not available in closed form. Therefore, we used Markov chain Monte Carlo (MCMC) simulation to obtain draws from posterior distributions of interest. Specifically, we employed `R`'s `nimble` package by de Valpine et al. (2023) (see also de Valpine et al. (2017)) to draw 20,000 posterior samples using 5 chains of 22000 samples with a 2000 sample burn-in and only keeping every fifth draw (i.e., thinning by five). Code to perform this MCMC sampling for data on a given surface and receipt location is provided in `https://github.com/runstats21/tennisShotLocation`. Raftery-Lewis diagnostics (see Raftery and Lewis (1992)), Gelman-Rubin diagnostics (see Gelman and Rubin (1992)), trace plots, auto-correlation plots, and calculations of effective sample size showed that convergence and mixing using this MCMC method were good.

3.3. Significance Testing

It is of interest to know if Nadal's shot location strategy is significantly different from that of Federer and, if so, which locations are significantly likely to be hit to by either player given a set of conditions. Specifically, we sequentially test the following:

I. Difference in overall behavior (in terms of odds ratio) of Federer (player one) vs. Nadal (player two) considering all four receipt locations and hit locations jointly for a given surface s, namely,

$$H_{0s}: \boldsymbol{\gamma}_{s1} = \boldsymbol{\gamma}_{s2} = \boldsymbol{\gamma}_{s3} = \boldsymbol{\gamma}_{s4} = \mathbf{1}_4 \text{ vs. } H_{1s}: \boldsymbol{\gamma}_{sr} \neq \mathbf{1}_4, \text{ for at least one } r = 1,\ 2,\ 3,\ 4.$$

II. Assuming we reject I, test the difference in four-dimensional vector of shot location odds ratios for each of the four receipt locations separately,

$$H_{0sr}: \boldsymbol{\gamma}_{sr} = \mathbf{1}_4 \text{ vs. } H_{1sr}: \boldsymbol{\gamma}_{sr} \neq \mathbf{1}_4, \text{ for } r = 1, \ldots, 4.$$

III. Assuming we reject I and II, test the difference in individuals' parameters of the odds ratio for one of the four shot locations,

$$H_{0ksr}: \gamma_{ksr} = 1 \text{ vs. } H_{1ksr}: \gamma_{ksr} \neq 1, \text{ for } r = 1, \ldots, 4;\ k = 1, \ldots, 4.$$

As stated in the enumeration above, we perform these tests sequentially, proceeding from the more global significance test to the more local only if all previous tests are significant. We perform these tests in this sequential manner to address issues that accompany multiple testing. In order to perform these tests of significance we employ the approach detailed in Held (2004). The key idea behind this approach is treating the probability statement provided in (8) as evidence for or against the hypothesized vector of odds rations $\boldsymbol{\gamma}_0$ which in the current setting is $\boldsymbol{\gamma}_0 = \mathbf{1}_4$ (four-dimensional vectors of ones)

$$Pr(\{\boldsymbol{\gamma}_{sr} \in \Gamma \colon p(\boldsymbol{\gamma}_{sr} \mid \boldsymbol{y}) \leq p(\boldsymbol{\gamma}_0 \mid \boldsymbol{y})\} \mid \boldsymbol{y}). \tag{8}$$

As defined in (8), the posterior contour probability is the probability of observing odds ratio values that are even less likely than the null value, given our data and prior knowledge. Specifically, the contour probabilities here are the probability of observing a randomly sampled vector of odds ratios, in the four-dimensional parameter space denoted as Γ, such that the posterior density evaluated using the odds ratio (denoted as $p(\boldsymbol{\gamma}_{sr} \mid \boldsymbol{y})$) is less than or equal to the posterior density evaluated using the null values labeled as $\boldsymbol{\gamma}_0$ (denoted as $p(\boldsymbol{\gamma}_0 \mid \boldsymbol{y})$).

Calculating (8) requires that we first derive the joint posterior distribution of $(\boldsymbol{\gamma}_{sr}, \boldsymbol{\pi}_{s2r})$ by way of the transformation $(\boldsymbol{\pi}_{s1r}, \boldsymbol{\pi}_{s2r}) \rightarrow (\boldsymbol{\gamma}_{sr}, \boldsymbol{\pi}_{s2r})$ (Note that $\boldsymbol{\pi}_{s2r}$ is arbitrarily chosen over $\boldsymbol{\pi}_{s1r}$) and then derive the marginal posterior distribution of $\boldsymbol{\gamma}_{sr}$. The later corresponds to $p(\boldsymbol{\gamma}_{sr} \mid \boldsymbol{y}) = \int p(\boldsymbol{\gamma}_{sr}, \boldsymbol{\eta}_{sr} \mid \boldsymbol{y}) d\boldsymbol{\eta}_{sr}$ where $\boldsymbol{\eta}_{sr} = (\boldsymbol{\pi}_{s2r}, \boldsymbol{\pi}_{0sr}, m_{sr})$ is a vector of nuisance parameters. Since the integral is not tractable, Held (2004) recommends using a Rao-Blackwellization approach based on $p(\boldsymbol{\gamma}_{sr} \mid \boldsymbol{\eta}_{sr}, \boldsymbol{y})$ to estimate $p(\boldsymbol{\gamma}_{sr} \mid \boldsymbol{y})$. Thus it is necessary to derive $p(\boldsymbol{\gamma}_{sr} \mid \boldsymbol{\eta}_{sr}, \boldsymbol{y})$ which is given in the following proposition, the proof of which can be found in the Appendix.

Proposition 1 *Under the model specifications detailed in equations* (1)-(4) *we have that*

$$Pr(\boldsymbol{\gamma}_{sr} \mid \boldsymbol{\eta}_{sr}, \boldsymbol{y}) = \prod_{k=1}^{4} \left\{ \frac{\frac{\pi_{ks2r}}{(1-\pi_{ks2r})}}{\left(1 + \frac{\gamma_{ksr}\pi_{ks2r}}{1-\pi_{ks2r}}\right)^2} \right\} \frac{\Gamma(m_{sr}\sum_{k=1}^{4}\pi_{k0sr} + \sum_{k=1}^{4} y_{ks1r})}{\prod_{k=1}^{4}\Gamma(m_{sr}\pi_{k0sr} + y_{ks1r})} \prod_{k=1}^{4} \left\{ \frac{\frac{\gamma_{ksr}\pi_{ks2r}}{1-\pi_{ks2r}}}{1 + \frac{\gamma_{ksr}\pi_{ks2r}}{1-\pi_{ks2r}}} \right\}^{m_{sr}\pi_{k0sr} + y_{ks1r} - 1}. \tag{9}$$

Now, given N MCMC samples from the joint posterior distribution of the parameters in model (1) - (4), the Rao-Blackwell estimate of (8) suggested by Held (2004) is

$$\hat{P}_{\mathrm{RB}}(\boldsymbol{\gamma}_0 \mid \boldsymbol{y}) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{\mathrm{med}_j \prod_{r=1}^{4} p(\boldsymbol{\gamma}_{sr}^i \mid \boldsymbol{\eta}_{sr}^j, \boldsymbol{y}) \leq \mathrm{med}_j \prod_{r=1}^{4} p(\boldsymbol{\gamma}_0 \mid \boldsymbol{\eta}_{sr}^j, \boldsymbol{y})\}, \tag{10}$$

where $\mathbf{1}\{\cdot\}$ is an indicator function and med_j denotes the median of all values indexed by $j = 1, \ldots, N$. Equation (10) estimates (8) for Hypothesis I using median values of MCMC samples from the posterior distribution of interest, which is invariant to any choice of strictly monotonic function applied to any posterior density $p(\boldsymbol{\gamma}_{sr} \mid \boldsymbol{y})$. To estimate (8) for Hypothesis II, we employ the following

$$\hat{P}_{\mathrm{RB}}(\boldsymbol{\gamma}_0 \mid \boldsymbol{y}) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{\mathrm{med}_j\ p(\boldsymbol{\gamma}_{sr}^i \mid \boldsymbol{\eta}_{sr}^j, \boldsymbol{y}) \leq \mathrm{med}_j\ p(\boldsymbol{\gamma}_0 \mid \boldsymbol{\eta}_{sr}^j, \boldsymbol{y})\}. \tag{11}$$

Thus, for a given surface s and receipt location r, we calculate $p(\boldsymbol{\gamma}_{sr}^i \mid \boldsymbol{\eta}_{sr}^i, \boldsymbol{y})$ for each of the i MCMC iterate odds ratio values (i.e., Equation (6)) using the derived closed form defined in (9). This results in an $N \times N$ matrix of full conditional density values, of which each row corresponds to full conditional density values for an individual combination of nuisance parameters j for MCMC iterate i. We then find the N median density values for each row j of this matrix, and count how many of them are less than or equal to the median of density values when the vector of odds ratios was set to $\boldsymbol{\gamma}_0$ (i.e., $\mathrm{med}_j\ p(\boldsymbol{\gamma}_0 \mid \boldsymbol{\eta}_{sr}^j, \boldsymbol{y})$), as specified in (11). Note that these tests can be computationally expensive as the number of MCMC samples N increases.

The only difference between the overall test for a surface s performed using (10) and the tests given surface s and receipt location r using (11) is that we take the product of all four density values for the four receipt locations (Z2, Z3, Z4, Z5) for the overall test before taking the median of each row j. The comparison of the median of these density values with $\mathrm{med}_j \prod_{r=1}^{4} p(\boldsymbol{\gamma}_0 \mid \boldsymbol{\eta}_{sr}^j, \boldsymbol{y})$, the median of the product of density values when evaluated at $\boldsymbol{\gamma}_0$, is then evaluated as indicated in Equation (10). The use of this product in these overall tests is possible because of our assumption of independence for each $\boldsymbol{\pi}_{spr}$ in our model (see Section 3.1).

4. Results

4.1. Significance Tests for Player Differences

In this section we describe results of testing Hypotheses in I, II, and III as outlined in Section 3.3. Specifically, Table 4 shows results from testing I and II. Note that Federer (whose is right-handed) and Nadal (who is left-handed) have significantly different shot location behavior on all three surfaces when considering all receipt locations jointly, with contour probabilities less than 0.0001. Furthermore, Federer and Nadal appear to have significantly different multivariate shot location behavior for all receipt locations except $r = 1$ and 2 on grass, with contour probability estimates less than 0.05. It is also interesting to note that Federer and Nadal appear to have fewer significant differences when receiving the ball on the right side of the court in Z2 and Z3, on average, than when receiving the ball on the left side of the court (Z4 and Z5).

Table 4. Results of multivariate tests for player differences described in Section 3.3. The contour probabilities in a row with "joint" receipt location correspond to Hypothesis I for each of the surfaces. The contour probabilities with each surface for different zones corresponds to Hypothesis II.

Surface	Receipt Location (r)	Contour Probability
Hard Court	Joint ($r = 1, \ldots, 4$)	**<0.0001**
Hard Court	Z2 ($r = 1$)	**0.0049**
Hard Court	Z3 ($r = 2$)	**<0.0001**
Hard Court	Z4 ($r = 3$)	**<0.0001**
Hard Court	Z5 ($r = 4$)	**<0.0001**
Clay	Joint ($r = 1, \ldots, 4$)	**<0.0001**
Clay	Z2 ($r = 1$)	**0.0380**
Clay	Z3 ($r = 2$)	**<0.0001**
Clay	Z4 ($r = 3$)	**<0.0001**
Clay	Z5 ($r = 4$)	**<0.0001**
Grass	Joint ($r = 1, \ldots, 4$)	**<0.0001**
Grass	Z2 ($r = 1$)	0.1051
Grass	Z3 ($r = 2$)	0.3086
Grass	Z4 ($r = 3$)	**0.0002**
Grass	Z5 ($r = 4$)	**<0.0001**

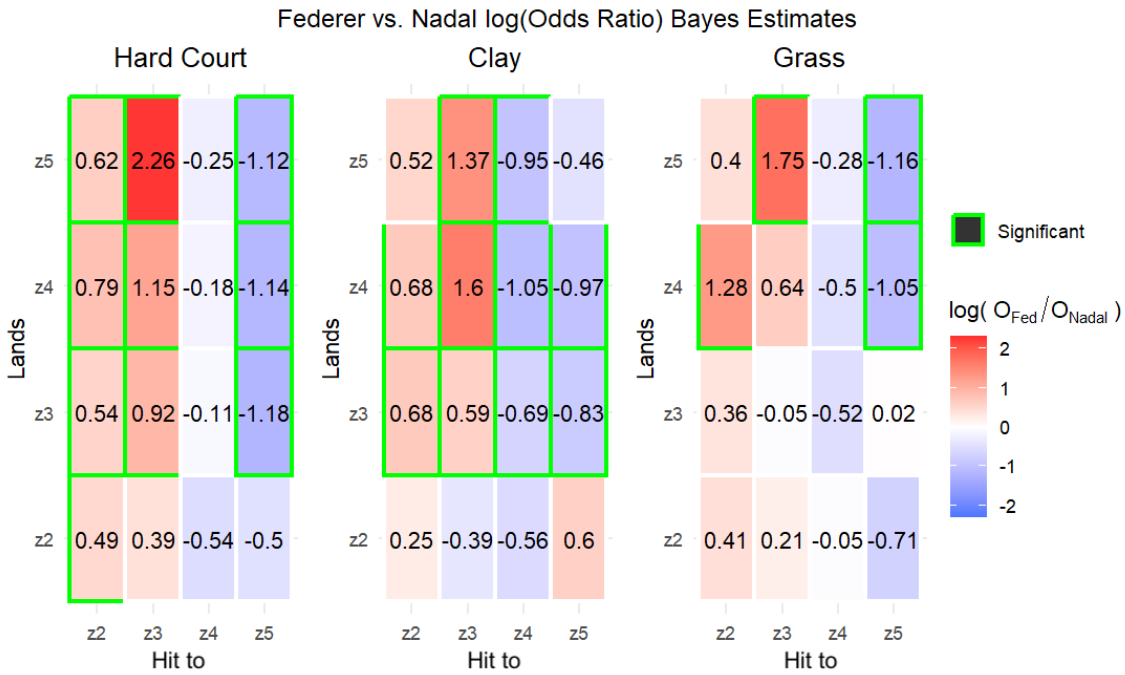

Fig. 4: Individual log(Odds Ratio) estimates and significance test of hypothesis III based on 95% Highest Posterior Density intervals. Recall that a log(Odds Ratio) of 0 indicates equality between probabilities.

Figure 4 shows results for testing Hypotheses III, with odds ratio's displayed on the log scale. Note that only locations where Hypothesis I and II were rejected should be considered for Hypotheses III. That said, no univariate tests were significant (95% Highest Posterior Density (HPD) intervals contained one) for receipt locations with insignificant tests for I and II (i.e., shots on grass when received in Z2 or Z3) which in a sense corroborates the sequential nature of the testing procedure. It is interesting to note that although we observe significant differences in the four dimensional expression of player behavior on clay when received in Z2 (with contour probability of 0.038), we can see in Figure 4 that none of the four shot locations are significantly more likely to be hit to by one player over another given this surface and receipt location. This highlights that this approach of sequential testing is relatively conservative, with our individual tests not finding spurious significance, even with the multivariate test being significant.

In general, it seems that the player differences shown using the log of the odds ratio ($\log(O_{\text{Fed}}/O_{\text{Nadal}})$) in Figure 4 reflects Nadal's desire to keep Federer from hitting his forehand as the log of the odds ratio generally leaning towards Federer being more likely to hit

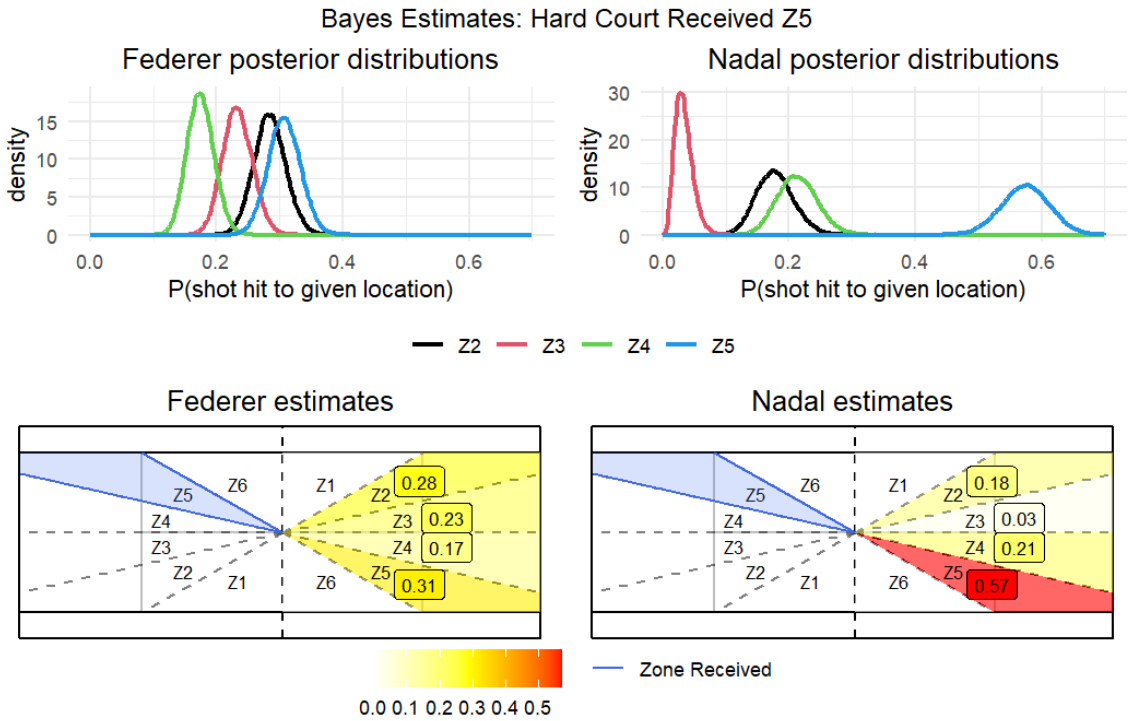

Fig. 5: Bayes estimates (posterior means) and corresponding distributions for both players for Hard Court (s=1) when received in Z5. Federer's distribution of stroke destinations is more uniform, whereas Nadal's is dominated by particular zones.

the opposing right side of the court (Z2 and Z3) with positive values, and Nadal being more likely to hit to the left side of the court with negative log odds ratio values. Additionally, there are many more significant differences between players on hard court and clay than on grass. It is also interesting to see absence of significant difference in players behavior of hitting to Z4 on hard court and grass, compared to significant differences on clay when received in Z3, Z4, and Z5. From this, it seems that players have different strategies across surfaces, especially when on hard court versus clay. Overall, it is clear these players have different behavior on all three surfaces and for most receipt locations, being most similar when receiving the ball in Z2, their far right side.

4.2. Expected Shot Location

Using the posterior MCMC samples, we were able to estimate probabilities of a shot being sent to the four zones (Z2, Z3, Z4, Z5) given receipt location r and surface s for both players. Interestingly, when looking at the posterior distributions for these parameters, we can see that some zones are significantly more likely to be hit to than others, while others have overlapping posterior density curves. For example, Figure 5 shows that the posterior distribution of Nadal's probability to hit to Z5 on hard court after receiving the ball in Z5 is dramatically higher than the distributions for all other zones, allowing for clear distinction of where Nadal will most likely hit the ball. However, overlapping posterior distributions of $\boldsymbol{\pi}_{spr}$ for Federer indicates that it is much harder to predict the location of his next shot when he receives the ball in Z5.

Estimates for each of the probabilities π_{kspr} are shown in Figures 6 and 7. Estimates that have no overlap of their 95% highest posterior density intervals with distributions of the other π_{kspr} parameters are outlined in green (See Table 7 in Appendix B). Posterior differences comparing the π_{kspr} for the shot location with the maximum Bayes estimate compared with the π_{kspr} for the other three zones was also computed, and showed the same behavior.

In Figure 6, we can see that Nadal is expected to hit to Z5 most often against Federer on hard court when receiving the ball in Z3, Z4 and Z5. Though Nadal appears to have similar behavior on grass when receiving the ball in Z5, it is interesting to note the difference in his behavior when playing on Clay. Nadal, known as the "king of clay", appears to be less predictable when playing Federer on this surface. Here the expected probability of Nadal hitting big cross-court shots with this strong left forehand is less than on hard court or grass, matching our intuition as clay is the slowest surface in terms of ball speed after a bounce, implying the need for a wider array of shot locations when trying to defeat Federer. On the other hand, although Federer generally appears to prefer hitting to his opponents right side with strong cross-court right forehand shots, it is interesting to see how he appears less predictable on hard court and grass than clay, especially when receiving the ball in Z4 and Z5 (his left side).

Seeing differences between players in Section 4.1 motivated us to focus the discussion in this section on locations with highest expected posterior probability given surface and receipt location for each player. However, with posterior samples for the 96 different π_{kspr} parameters related to given shot location probabilities, there are many other comparisons that are available that can help glean

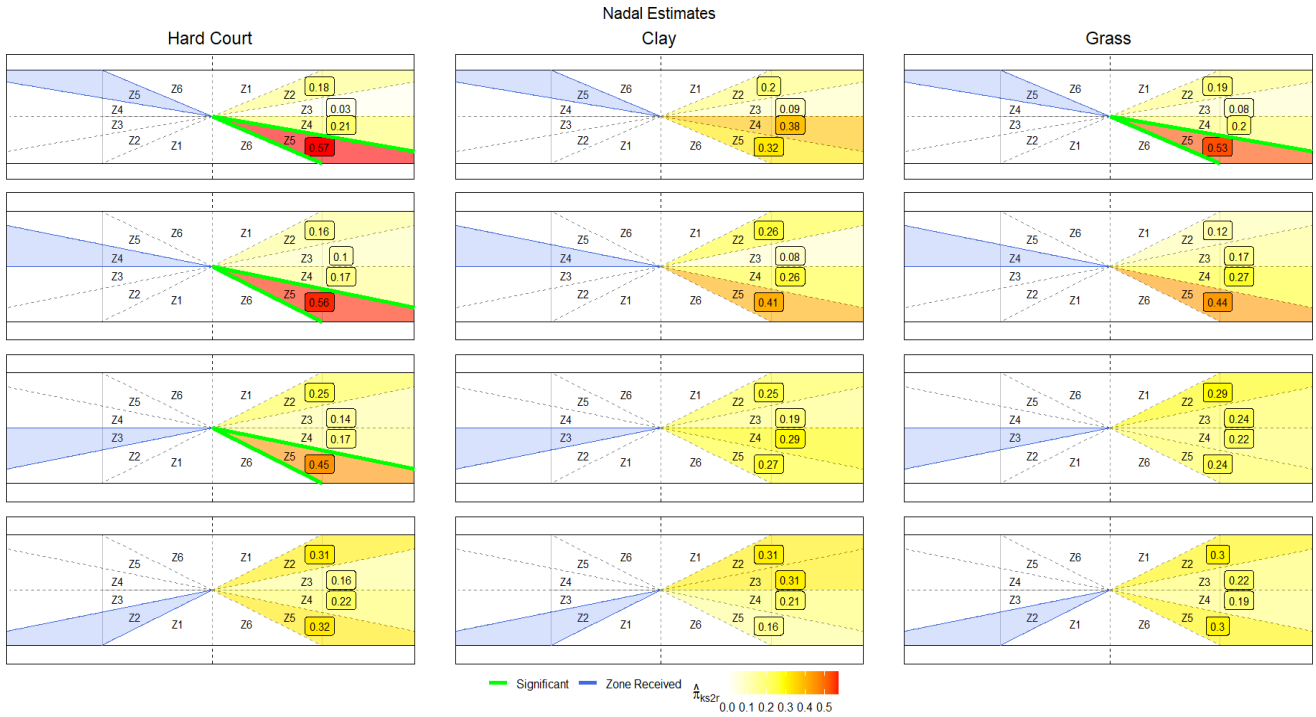

Fig. 6: Posterior means for all Nadal shot location probability parameters (π_{ks2r}), labeled in the legend as $\hat{\pi}_{ks2r}$. Zones significantly more probable to be hit to are outlined in green (non-overlapping 95% HPD credible intervals).

Table 5. 95% posterior probability intervals (i.e., credible intervals) on right vs. left side probabilities ($(\pi_{1spr} + \pi_{2spr}) - (\pi_{3spr} + \pi_{4spr})$). Bold observations indicate significance, with credible intervals not containing zero.

		Federer			Nadal		
Surface	Receipt location	Lwr	Upr	Side	Lwr	Upr	Side
Hard Court	Z2	**0.108**	**0.433**	**Right**	-0.204	0.065	Left
Hard Court	Z3	**0.116**	**0.455**	**Right**	**-0.371**	**-0.077**	**Left**
Hard Court	Z4	-0.069	0.288	Right	**-0.616**	**-0.314**	**Left**
Hard Court	Z5	-0.075	0.144	Right	**-0.695**	**-0.446**	**Left**
Clay	Z2	**0.007**	**0.403**	**Right**	**0.101**	**0.392**	**Right**
Clay	Z3	**0.197**	**0.525**	**Right**	-0.293	0.035	Left
Clay	Z4	**0.198**	**0.524**	**Right**	**-0.517**	**-0.125**	**Left**
Clay	Z5	-0.005	0.312	Right	**-0.585**	**-0.214**	**Left**
Grass	Z2	**0.059**	**0.496**	**Right**	-0.181	0.238	Right
Grass	Z3	-0.029	0.420	Right	-0.132	0.254	Right
Grass	Z4	-0.026	0.400	Right	**-0.613**	**-0.196**	**Left**
Grass	Z5	-0.020	0.334	Right	**-0.655**	**-0.238**	**Left**

information associated with relationships of interest. For example, although the probability estimate for Federer hitting to Z2 on when received in Z3 on hard court is not significantly higher than all other zones, it is significantly higher than the probabilities hit to Z4 and Z5. This establishes the notion that Federer is more likely to hit the ball to the right side of the court than the left. In addition, it is possible to determine the "side preference" for each player using MCMC samples from the joint posterior of $\boldsymbol{\pi}_{spr}$. It is straightforward to collect samples from the marginal posterior distribution of $\pi_{1spr} + \pi_{2spr} - (\pi_{3spr} + \pi_{4spr})$ for each player and use them to estimate 95% credible intervals of the difference just listed. These intervals are provided in Table 5. From this table it appears that Federer is more likely to hit to one of the right side zones of the court than the left, on average. Nadal, however, is more likely to hit to the left zones than the right on hard court, but then varies on which side he is most likely to hit to based on receipt location when playing on clay and grass. These results provide more general information into which side of the court the ball is most likely go, and to some extent highlights that Federer and Nadal have different dominant hands. This can help players/coaches strategize or fans know which side of the court they can expect the ball to be hit. Many other questions likes these can be answered with the methodology described in Section 3.1.

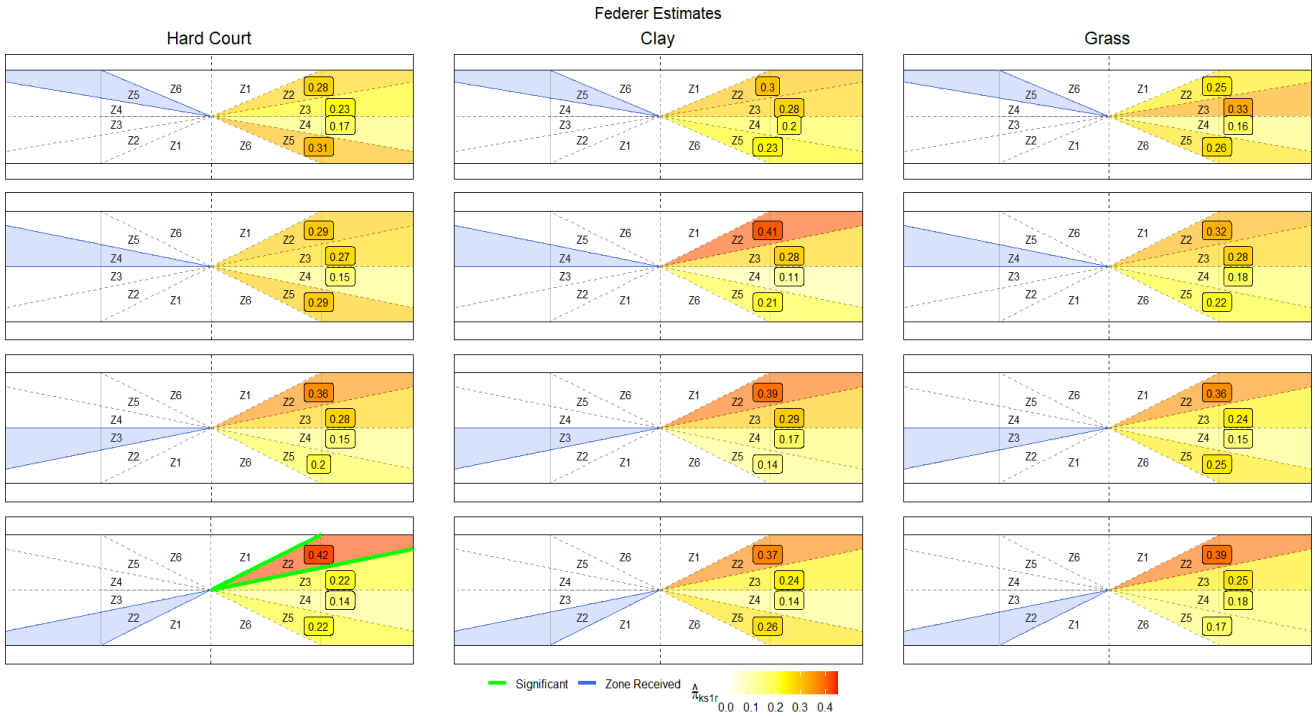

Fig. 7: Posterior means for all Federer shot location probability parameters (π_{ks1r}), labeled in the legend as $\hat{\pi}_{ks1r}$. Zones significantly more probable to be hit to are outlined in green (non-overlapping 95% HPD credible intervals).

5. Conclusion

Using a Bayesian hierarchical model with a Multinomial likelihood and Dirichlet prior, we were able to build a framework to analyze shot location strategy by Roger Federer and Rafael Nadal when playing each other. This framework for player strategy analysis can enhance scouting, coaching, and fan accessibility to the sport of tennis through quantification of player differences and estimation of shot location probabilities. Both players appeared most predictable when hitting to their forehand side on hard court, with Nadal significantly more likely to hit to the opposing far left side of the court (zone five) than any other zone when received in any location other than his far right side. This pattern falls in line with how Nadal described his strategy against Federer in a 2025 interview on the "Served with Andy Roddick" podcast[2]. Federer is significantly more likely to hit to his opponents far right side when he receives the ball on his far right side (zone two). We also found that Federer and Nadal vary in their differences over different court surfaces, especially in clay versus hard court. Though we are able to answer several questions using the posterior samples generated by our model, such as where Federer or Nadal are most likely to hit the ball given receipt location on a certain surface type, and even which side of the court (left vs. right) they are expected to hit the ball, there are many more questions that can be answered using this framework as interest demands, which can be examined in expansions on this work.

We described the results for three classical matches between two of the best tennis players in history. The supplementary material contains details about two additional matches with Nadal competing against Novak Djokovic and Federer against Juan del Potro. There we show that Federer and Nadal's strategy changes when facing different opponents. Future efforts will be dedicated to applying this methodology to a wider range of matches. This will permit studying how the variation in, for example, Federer and Nadal's trends change over time, especially as players with different hand dominance, experience, and style of play compete. Along with allowing us to catalog the diversity of strategies employed by tennis players, including data from additional matches between Federer and Nadal could provide enough data hit to zones one and six to expand our model from four zones (Z2, Z3, Z4, Z5) to six zones (Z1, Z2, Z3, Z4, Z5, Z6) on each side of the court or even partition the court into a more fine grid of zones. Inference can also be improved using hierarchical structure to borrow strength across surfaces and/or receipt locations. Additionally, inclusion of hit type (e.g. top spin vs. slice) and previous hit location information can be examined with the hopes to further improve estimation of shot location probabilities.

A. Proof of Proposition 1

In the proof of Proposition 1, we drop the index referring to surface s and received location r for clarity's sake.

Derivation of $p(\boldsymbol{\gamma}, \boldsymbol{\pi}_2 | \boldsymbol{\pi}_0, m, \boldsymbol{y})$

[2] https://www.youtube.com/watch?v=fMBaJtc7H6g

As specified in equation (9), $p(\boldsymbol{\gamma}|\boldsymbol{\pi}_2, \boldsymbol{\pi}_0, m, \boldsymbol{y}) = \frac{p(\boldsymbol{\gamma}, \boldsymbol{\pi}_2|\boldsymbol{\pi}_0, m, \boldsymbol{y})}{p(\boldsymbol{\pi}_2|\boldsymbol{\pi}_0, m, \boldsymbol{y})}$. With $p(\boldsymbol{\pi}_2|\boldsymbol{\pi}_0, m, \boldsymbol{y})$, known to be Dirichlet distributed, we now need to find $p(\boldsymbol{\gamma}, \boldsymbol{\pi}_2|\boldsymbol{\pi}_0, m, \boldsymbol{y})$ to complete the derivation of the full conditional of interest $p(\boldsymbol{\gamma}|\boldsymbol{\pi}_2, \boldsymbol{\pi}_0, m, \boldsymbol{y})$. Here we write out the full derivation of $p(\boldsymbol{\gamma}, \boldsymbol{\pi}_2|\boldsymbol{\pi}_0, m, \boldsymbol{y})$ using a transformation of variables. Specifically, we find the distribution of the $\boldsymbol{W} = \boldsymbol{\gamma}, \boldsymbol{\pi}_2|\boldsymbol{\pi}_0, m, \boldsymbol{y}$ where here we denote γ_k, the ratio of the odds of Federer (player one) hitting to location k compared to the odds of Nadal (player two) hitting to location k on a given surface and receipt location, as $\frac{\pi_{1k}}{(1-\pi_{1k})} / \frac{\pi_{2k}}{(1-\pi_{2k})}$, using the known distribution of $\boldsymbol{X} = \boldsymbol{\pi}_1, \boldsymbol{\pi}_2|\boldsymbol{\pi}_0, m, \boldsymbol{y}$. Note that $X = \boldsymbol{\pi}_1, \boldsymbol{\pi}_2|\boldsymbol{\pi}_0, m, \boldsymbol{y} = (\boldsymbol{\pi}_1|\boldsymbol{\pi}_0, m, \boldsymbol{y}) \times (\boldsymbol{\pi}_2|\boldsymbol{\pi}_0, m, \boldsymbol{y})$, the product of posterior distributions we already know to be Dirichlet, via the conditional independence of $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ given $\boldsymbol{\pi}_0$, m, and $\boldsymbol{y}$. Therefore, using a transformation of variables g, where $g(\boldsymbol{X}) = (\boldsymbol{\gamma}, \boldsymbol{\pi_2}) = \left(\frac{\boldsymbol{\pi}_1}{(1-\boldsymbol{\pi}_1)} / \frac{\boldsymbol{\pi}_2}{(1-\boldsymbol{\pi}_2)}, \boldsymbol{\pi}_2\right) \implies g^{-1}(\boldsymbol{W}) = (\boldsymbol{\pi}_1, \boldsymbol{\pi_2}) = \left(\frac{\boldsymbol{\gamma}\boldsymbol{\pi}_2}{1-\boldsymbol{\pi}_2} / (1 + \frac{\boldsymbol{\gamma}\boldsymbol{\pi}_2}{1-\boldsymbol{\pi}_2}), \boldsymbol{\pi}_2\right)$,

$$f_{\boldsymbol{W}}(\boldsymbol{w}) = f_{\boldsymbol{X}}(g^{-1}(\boldsymbol{w}))\ |\frac{\partial g^{-1}(\boldsymbol{w})}{\partial \boldsymbol{w}}| = f_{\boldsymbol{X}}(g^{-1}(\boldsymbol{w}))\ |\boldsymbol{J}|. \tag{12}$$

Closed form of Jacobian J in transformation $(\pi_1, \pi_2) \xrightarrow{g} (\gamma, \pi_2)$

∂	$g^{-1}(\boldsymbol{w})$							
$\boldsymbol{W}$	π_{11}	π_{12}	π_{13}	π_{14}	π_{21}	π_{22}	π_{23}	π_{24}
γ_1	J_{11}	0	0	0	0	0	0	0
γ_2	0	J_{22}	0	0	0	0	0	0
γ_3	0	0	J_{33}	0	0	0	0	0
γ_4	0	0	0	J_{44}	0	0	0	0
π_{21}	J_{51}	0	0	0	1	0	0	0
π_{22}	0	J_{62}	0	0	0	1	0	0
π_{23}	0	0	J_{73}	0	0	0	1	0
π_{24}	0	0	0	J_{84}	0	0	0	1

Table 6. Jacobian (matrix of partial derivatives) for transformation of variables from $\boldsymbol{X} = g^{-1}(\boldsymbol{w})$ to $\boldsymbol{W}$ via transformation g

$$\begin{bmatrix} J_{11} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & J_{22} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & J_{33} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & J_{44} & 0 & 0 & 0 & 0 \\ J_{51} & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & J_{62} & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & J_{73} & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & J_{84} & 0 & 0 & 0 & 1 \end{bmatrix} \tag{13}$$

where the first four diagonal elements J_{kk} have the general form

$$J_{kk} = \frac{\partial \frac{\frac{\gamma_k \pi_{2k}}{1-\pi_{2k}}}{(1+\frac{\gamma_k \pi_{2k}}{1-\pi_{2k}})}}{\partial \gamma_k} = \frac{\frac{\pi_{2k}}{(1-\pi_{2k})}}{(1 + \frac{\gamma_k \pi_{2k}}{1-\pi_{2k}})^2}\ , \tag{14}$$

and the diagonal elements of the bottom left 4×4 block, which we denote with J_{ik}, where $i = k + 4$, have the form

$$J_{ik} = \frac{\partial \frac{\frac{\gamma_k \pi_{2k}}{1-\pi_{2k}}}{(1+\frac{\gamma_k \pi_{2k}}{1-\pi_{2k}})}}{\partial \pi_{2k}} = \frac{\frac{\gamma_k}{(1-\pi_{2k})^2}}{(1 + \frac{\gamma_k \pi_{2k}}{1-\pi_{2k}})^2}\ . \tag{15}$$

In order to find $|J|$, the determinant of the Jacobian, we use the following property of determinants of block diagonal matrices in Equation (16),

$$\begin{aligned} &\text{Suppose matrix } \boldsymbol{A} = \begin{pmatrix} \boldsymbol{A}_{11} & \boldsymbol{0} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{pmatrix}, \text{ where } \boldsymbol{A}_{11} \text{ and } \boldsymbol{A}_{22} \text{ are square.} \\ &\text{Then, } |\boldsymbol{A}| = |\boldsymbol{A}_{11}||\boldsymbol{A}_{22}|. \\ &\text{Therefore, since } \boldsymbol{J} = \begin{pmatrix} \underset{4\times4}{\boldsymbol{J}_{11}} & \underset{4\times4}{\boldsymbol{0}} \\ \underset{4\times4}{\boldsymbol{J}_{21}} & \underset{4\times4}{\boldsymbol{I}} \end{pmatrix},\ |\boldsymbol{J}| = |\boldsymbol{J}_{11}||\boldsymbol{I}| = |\boldsymbol{J}_{11}| = \prod_{k=1}^{4} J_{kk}\ . \end{aligned} \tag{16}$$

Therefore,

$$|\boldsymbol{J}| = \prod_{k=1}^{4} J_{kk} = \prod_{k=1}^{4} \frac{\frac{\pi_{2k}}{(1-\pi_{2k})}}{(1+\frac{\gamma_k \pi_{2k}}{1-\pi_{2k}})^2} \ . \tag{17}$$

Closed form of $p(\boldsymbol{\gamma}, \boldsymbol{\pi}_2 | \boldsymbol{\pi}_0, m, \boldsymbol{y})$

Using the closed form of the Jacobian expressed in Equation (17) above, the closed form of $f_{\boldsymbol{W}}(\boldsymbol{w}) = p(\boldsymbol{\gamma}, \boldsymbol{\pi}_2 | \boldsymbol{\pi}_0, m, \boldsymbol{y})$ can therefore be written,

$$\begin{aligned} &p(\boldsymbol{\gamma}, \boldsymbol{\pi}_2 | \boldsymbol{\pi}_0, m, \boldsymbol{y}) = |\boldsymbol{J}| f_{\boldsymbol{X}}(g^{-1}(\boldsymbol{w})) \\ &= \left(\prod_{k=1}^{4} \frac{\frac{\pi_{2k}}{(1-\pi_{2k})}}{(1+\frac{\gamma_k \pi_{2k}}{1-\pi_{2k}})^2} \right) \frac{\Gamma(m\sum_{k=1}^{4} \pi_{0k} + \sum_{k=1}^{4} y_{1k})}{\prod_{k=1}^{4} \Gamma(m\pi_{0k} + y_{1k})} \prod_{k=1}^{4} \left(\frac{\frac{\gamma_k \pi_{2k}}{1-\pi_{2k}}}{1+\frac{\gamma_k \pi_{2k}}{1-\pi_{2k}}} \right)^{m\pi_{0k}+y_{1k}-1} \\ &\times \frac{\Gamma(m\sum_{k=1}^{4} \pi_{0k} + \sum_{k=1}^{4} y_{2k})}{\prod_{k=1}^{4} \Gamma(m\pi_{0k} + y_{2k})} \prod_{k=1}^{4} (\pi_{2k})^{m\pi_{0k}+y_{2k}-1} \end{aligned} \tag{18}$$

where y_{1k} is the number of times player 1 (Federer) hits to location k, y_{2k} is the number of times player 2 (Nadal) hits to location k, and m is the constant drawn from a Gamma$(1,1)$ that allows for player-specific variations from the overall probability estimate (which we denote with π_{0k}) that a shot is sent to location k.

Closed form of $p(\boldsymbol{\gamma} | \boldsymbol{\pi}_2, \boldsymbol{\pi}_0, m, \boldsymbol{y})$

Using the derivation in Equation (18), we can find the full closed form of $p(\boldsymbol{\gamma} | \boldsymbol{\pi}_2, \boldsymbol{\pi}_0, m, \boldsymbol{y}) = \frac{p(\boldsymbol{\gamma}, \boldsymbol{\pi}_2 | \boldsymbol{\pi}_0, m, \boldsymbol{y})}{p(\boldsymbol{\pi}_2 | \boldsymbol{\pi}_0, m, \boldsymbol{y})}$. Specifically, the density $p(\boldsymbol{\pi}_2 | \boldsymbol{\pi}_0, m, \boldsymbol{y})$ can be found in the form of $p(\boldsymbol{\gamma}, \boldsymbol{\pi}_2 | \boldsymbol{\pi}_0, m, \boldsymbol{y})$, namely

$$p(\boldsymbol{\pi}_2 | \boldsymbol{\pi}_0, m, \boldsymbol{y}) = \frac{\Gamma(m\sum_{k=1}^{4} \pi_{0k} + \sum_{k=1}^{4} y_{2k})}{\prod_{k=1}^{4} \Gamma(m\pi_{0k} + y_{2k})} \prod_{k=1}^{4} (\pi_{2k})^{m\pi_{0k}+y_{2k}-1} \ . \tag{19}$$

Therefore, after performing the division $\frac{p(\boldsymbol{\gamma}, \boldsymbol{\pi}_2 | \boldsymbol{\pi}_0, m, \boldsymbol{y})}{p(\boldsymbol{\pi}_2 | \boldsymbol{\pi}_0, m, \boldsymbol{y})}$ from Equation (9), the form of $p(\boldsymbol{\gamma} | \boldsymbol{\pi}_2, \boldsymbol{\pi}_0, m, \boldsymbol{y})$ has the form

$$p(\boldsymbol{\gamma} | \boldsymbol{\pi}_2, \boldsymbol{\pi}_0, m, \boldsymbol{y}) = \left(\prod_{k=1}^{4} \frac{\frac{\pi_{2k}}{(1-\pi_{2k})}}{(1+\frac{\gamma_k \pi_{2k}}{1-\pi_{2k}})^2} \right) \frac{\Gamma(m\sum_{k=1}^{4} \pi_{0k} + \sum_{k=1}^{4} y_{1k})}{\prod_{k=1}^{4} \Gamma(m\pi_{0k} + y_{1k})} \prod_{k=1}^{4} \left(\frac{\frac{\gamma_k \pi_{2k}}{1-\pi_{2k}}}{1+\frac{\gamma_k \pi_{2k}}{1-\pi_{2k}}} \right)^{m\pi_{0k}+y_{1k}-1} . \tag{20}$$

This finishes the proof.

B. Supplementary Tables

Table 7 shows 95% Highest Posterior Density Intervals (HPDI) for the zone with the highest posterior mean (i.e., the "expected shot location") for each shot location probability parameter for Federer vs. Nadal.

Table 7. Expected Shot Locations and Highest Posterior Density Intervals (HPDI)

Surface	Receipt Zone	Federer (95% HPDI)	Nadal (95% HPDI)
Hard Court	Z2	**Z2 (0.334, 0.5)**	Z5 (0.254, 0.377)
Hard Court	Z3	Z2 (0.285, 0.455)	**Z5 (0.374, 0.524)**
Hard Court	Z4	Z5 (0.215, 0.378)	**Z5 (0.479, 0.648)**
Hard Court	Z5	Z5 (0.258, 0.358)	**Z5 (0.498, 0.647)**
Clay	Z2	Z2 (0.268, 0.464)	Z2 (0.244, 0.384)
Clay	Z3	Z2 (0.305, 0.475)	Z4 (0.219, 0.367)
Clay	Z4	Z2 (0.323, 0.495)	Z5 (0.308, 0.507)
Clay	Z5	Z2 (0.224, 0.37)	Z4 (0.287, 0.483)
Grass	Z2	Z2 (0.279, 0.5)	Z2 (0.204, 0.394)
Grass	Z3	Z2 (0.254, 0.472)	Z2 (0.2, 0.375)
Grass	Z4	Z2 (0.22, 0.419)	Z5 (0.33, 0.555)
Grass	Z5	Z3 (0.243, 0.413)	**Z5 (0.412, 0.646)**

6. Author contributions statement

7. Acknowledgments

The authors thank the editors and anonymous reviewers for their valuable suggestions. A.Jara's research was supported by Fondecyt 1220907 grant. Part of this work was performed during a visit of G.L. Page to Chile also supported by Fondecyt 1220907 grant.

References

T. C. Y. Chan, D. S. Fearing, C. Fernandes, and S. Kovalchik. A Markov process approach to untangling intention versus execution in tennis. *Journal of Quantitative Analysis in Sports*, 18(2):127–145, 2022. doi: doi:10.1515/jqas-2021-0077. URL `https://doi.org/10.1515/jqas-2021-0077`.

A. Cornman, G. Spellman, and D. Wright. Machine learning for professional tennis match prediction and betting. *Working Paper, Stanford University*, 2017.

P. de Valpine, D. Turek, C. Paciorek, C. Anderson-Bergman, D. Temple Lang, and R. Bodik. Programming with models: writing statistical algorithms for general model structures with NIMBLE. *Journal of Computational and Graphical Statistics*, 26:403–417, 2017. doi: 10.1080/10618600.2016.1172487.

P. de Valpine, C. Paciorek, D. Turek, N. Michaud, C. Anderson-Bergman, F. Obermeyer, C. Wehrhahn Cortes, A. Rodríguez, D. Temple Lang, W. Zhang, S. Paganin, J. Hug, and P. van Dam-Bates. *NIMBLE: MCMC, Particle Filtering, and Programmable Hierarchical Modeling*, 2023. URL `https://cran.r-project.org/package=nimble`.

N. E. Dona, P. S. Gill, and T. B. Swartz1. What does rally length tell us about player characteristics in tennis? *Journal of the Royal Statistical Society Series A*, 00:1–17, 2024.

A. Ganser, B. Hollaus, and S. Stabinger. Classification of tennis shots with a neural network approach. *Sensors*, 21:5703, 08 2021. doi: 10.3390/s21175703.

A. Gelman and D. B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472, 1992.

B. Giles, S. Kovalchik, and M. Reid. A machine learning approach for automatic detection and classification of changes of direction from player tracking data in professional tennis. *Journal of sports sciences*, 38(1):106–113, 2020.

L. Held. Simultaneous posterior probability statements from monte carlo output. *Journal of Computational and Graphical Statistics*, 13 (1):20–35, 2004. doi: 10.1198/1061860043083. URL `https://doi.org/10.1198/1061860043083`.

P. Hoff. *A First Course in Bayesian Statistical Methods*. Springer Texts in Statistics. Springer New York, 2009. ISBN 9780387924076. URL `https://books.google.com/books?id=V8jT2SimGR0C`.

S. Kovalchik, M. Ingram, K. Weeratunga, and C. Goncu. Space-time von cramm: Evaluating decision-making in tennis with variational generation of complete resolution arcs via mixture modeling, 2020. URL `https://arxiv.org/abs/2005.12853`.

S. A. Kovalchik. Searching for the goat of tennis win prediction. *Journal of Quantitative Analysis in Sports*, 12(3):127–138, 2016. doi: doi:10.1515/jqas-2015-0059. URL `https://doi.org/10.1515/jqas-2015-0059`.

S. A. Kovalchik and J. Albert. A statistical model of serve return impact patterns in professional tennis, 2022. URL `https://arxiv.org/abs/2202.00583`.

F. Lisi and M. Grigoletto. Modeling and simulating durations of professional tennis matches by resampling match features. *Journal of Sports Analytics*, 7, 06 2019. doi: 10.3233/JSA-200455.

F. Lisi, M. Grigoletto, and M. Briglia. On the distribution of rally length in professional tennis matches. *Preprint. https://www.researchgate.net/publication/369375375*, 03 2023.

A. Mehra. Expectedin model. *Kaggle*, Nov 2023. URL `https://www.kaggle.com/code/aahanmehra/expectedin-model`.

A. E. Raftery and S. M. Lewis. [Practical Markov Chain Monte Carlo]: Comment: One Long Run with Diagnostics: Implementation Strategies for Markov Chain Monte Carlo. *Statistical Science*, 7(4):493 – 497, 1992. doi: 10.1214/ss/1177011143. URL `https://doi.org/10.1214/ss/1177011143`.

B. Tendu, C. Apindi, and A. Simwa. Top 20 most popular sports in the world revealed. *Sports Brief - Sport News, Transfers, Scores and Results, https://sportsbrief.com/other-sports/16715-revealed-top-popular-sports-world/*, 2023. URL `https://sportsbrief.com/other-sports/16715-revealed-top-popular-sports-world/`.

X. Wei, P. Lucey, S. Morgan, and S. Sridharan. Predicting shot locations in tennis using spatiotemporal data. *In Digital Image Computing: Techniques and Applications (DICTA), 2013*, pages 1–8, 11 2013a. doi: 10.1109/DICTA.2013.6691516.

X. Wei, P. Lucey, S. Morgan, and S. Sridharan. "sweet-spot": Using spatiotemporal data to discover and predict shots in tennis. *In MIT Sloan Sports Analytics Conference*, 01 2013b.

X. Wei, P. Lucey, S. Morgan, M. Reid, and S. Sridharan. The thin edge of the wedge: Accurately predicting shot outcomes in tennis using style and context priors. *In MIT Sloan Sports Analytics Conference*, 2016.

D. Whiteside and M. Reid. Spatial characteristics of professional tennis serves with implications for serving aces: A machine learning approach. *Journal of sports sciences*, 35(7):648–654, 2017.

D. Whiteside, O. Cant, M. Connolly, and M. Reid. Monitoring hitting load in tennis using inertial sensors and machine learning. *International journal of sports physiology and performance*, 12(9):1212–1217, 2017.

S. Wilkens. Sports prediction and betting models in the machine learning age: The case of tennis. *Journal of Sports Analytics*, 7(2): 99–117, 2021.