Next Shot Prediction in Badminton

by

Xunhang Gao

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Major: Statistics

Program of Study Committee:
Jarad Niemi, Major Professor
Lynna Chu
Zhengyuan Zhu

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation/thesis. The Graduate College will ensure this dissertation/thesis is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2025

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Page

CHAPTER 1. ACKNOWLEDGMENTS

I would like to express my sincere gratitude to Dr. Jarad Niemi for his invaluable guidance, teaching, and mentorship throughout the process of developing this creative component. His patience, insightful scientific perspective, and meticulous attention to detail have profoundly enhanced my understanding and appreciation for statistics and academic research. Additionally, I extend my appreciation to Dr. Zhengyuan Zhu and Dr. Lynna Chu for generously serving on my committee and contributing their valuable insights and expertise.

CHAPTER 2. ABSTRACT

This project conducted a statistical analysis of badminton matches to identify significant factors influencing player performance. Utilizing a comprehensive dataset from stroke-level match data between 2018 and 2021, we explore spatial and tactical behaviors through statistical methodologies, including Poisson point processes, kernel density estimation, logistic regression, and Bayesian inference. The analysis reveals critical differences in serve against genders and evaluates the significant impacts of spatial variables, such as player positioning and opponent location, on distrubution of future shots. Additionally, the project leverages Bayesian methods to quantify the uncertainty in these estimations effectively. The main objective of this project is to provide valuable insights on actionable strategies for player training and competitive play. Future research directions are also discussed, including possible improvements in data accuracy based on potential integration of trajectory modeling and application of more advanced models, such as multinomial logistic regression models.

CHAPTER 3. Introduction

Badminton, characterized by its rapid pace and strategic intricacies, has emerged as a globally popular sport attracting significant academic and professional interest in performance analytics. However, compared to sports like tennis and basketball, systematic and rigorous statistical analyses specific to stroke-level data remain relatively underexplored. Addressing this gap requires sophisticated statistical methodologies to better understand tactical decision-making, player behavior, and performance determinants.

In tennis, the research challenged the assumption that each tennis point is independent of others. Using Feasible Generalized Least Squares (FGLS) (Klaassen and Magnus (2001)), researchers found that point outcomes depended on various factors like player ranking and previous points, showing clear dependencies between points.

Another important tennis study, the analysis of final sets in four years at Wimbledon (Magnus and Klaassen (1999)) used logistic regression to explore how player ranking and service effectiveness influenced match outcomes. They concluded that higher-ranked players consistently had an advantage, and the effectiveness of the serve didn't decrease even in long matches.

Basketball analytics have also seen notable progress. The study of identifying key factors in momentum in basketball games (Chen et al. (2021)) used an elastic net regression method—a combination of ridge regression, LASSO, and ordinary least squares (OLS)—to identify what drives momentum shifts. They found rebounds and turnovers significantly influenced momentum.

In badminton, the research of analyzing the badminton scoring systems (Percy (2009)) examined differences between old and new scoring systems. They used probability theory to find minimal differences between systems in terms of point-winning probability. Additionally, this study introduced Bayesian methods to predict match outcomes before the matches began.

The paper identifying player's performance in badminton using Markov chain (Galeanoa et al. (2022)) looked at player performance by analyzing sequences of shots. The researchers divided the badminton court into 12 zones and tracked where shots landed over three-shot sequences. Using a Markov chain model with states representing winning or losing patterns, they calculated an Expected Pattern Value (EPV), linking specific shot sequences directly to the probability of winning the rally. This helped identify strategic connections between shot patterns and rally outcomes.

Finally, in volleyball analytics, the paper of prediction of the results of volleyball matches applied Bayesian hierarchical models (Gabrio (2021)). These included Poisson log-linear models for predicting points and Bernoulli distributions for match results. They also introduced an inverse-Wishart model to handle correlations among team performance metrics. This research demonstrated the strength and reliability of Bayesian methods in predicting sports outcomes.

These studies show the value of advanced statistical techniques, such as Markov chains, logistic regression, Bayesian modeling, and elastic net regression, in sports analytics. Inspired by their success, especially in badminton and tennis, this project seeks to apply similar methods specifically to badminton. The goal is to better understand how players choose shots, how they serve, and what factors influence their decisions during matches.

The primary objective of this project is to identify and quantify the significant factors influencing shot selection patterns, serve distributions, and rally outcomes in elite badminton matches. Specifically, the project addresses the following scientific questions: What spatial and player-specific factors significantly influence shot distribution patterns? How do player attributes such as handedness and gender affect strategic decisions during serves and rallies? Lastly, how can statistical methods, such as kernel density esitmation, generalized linear model, and Bayesian analysis, enhance the estimation and interpretation of uncertainties in predicting badminton match outcomes?

CHAPTER 4. Badminton Data

The data for this project come from the ShuttleSet dataset, a standardized dataset specifically for badminton matches proposed in the research by Wang et al. (2023). And for the purpose of analysis, we use homography techniques to correct the coordinates in the original data and qualify the distribution of serves, patterns of other shots, and progression in scores by adding various variables based on original data.

4.1 Raw Data

The raw data from ShettleSet dataset includes stroke-level data collected from recordings of 44 badminton matches played between 2018 and 2021 by 27 top-ranking male and female players. And all the 44 matches are selected from the finals, semi-finals, and quarter-finals of the top tournaments, such as Super 1000 series and World Tour Finals. The ShuttleSet dataset consists of two main parts: match-level information and stroke-level match data.

The first part contains basic match-level detail (full match-level information is listed in Table 4.1), including:

- Tournament information: 1) The event name, such as " Fuzhou Open 2018" and " World Tour Finals 2019". 2) The round of each match in the events, such as "finals". "semi-finals", and "quarter-finals". 3) The date of each match, including the year, month, and day.
- Match information: 1) The number of sets (games) played in each match. 2) The duration of each match recorded in minutes. 3) The winner and the loser of each match.

The Table 4.2 presents the number of matches played by each player, and Chou Tien Chen and Viktor Axelsen are the two male player with most data and Carolina Marin is the female player with most data. Our analysis of player-specific playing strategies will mainly focus on these three players.

The second part contains stroke-level details for each match, including:

- Basic rally information: 1) The current score of both players in the match. 2) The winner of each rally, listed in the last stroke (row) of each rally in the data. 3) The winning reason categorized by "out", "touched the net", "not pass over the net", "opponent's ball landed", and "misjudged".

- Spatial information (locations of the pixels in the recording of matches): 1) The hitting locations (landing location of next shot) of each shot. 2) The landing location (the hitting location of next shot) of each shot. 3) The player location when each shot is made by this player. 4) The opponent location when each shot is made. All of these locations are recorded in both specific (x, y) coordinates and areas, ranging from 1 to 9, following analytical convention in badminton presented in the Figure 4.1.

- Hitting information: 1) The types of techniques used for each shot (18 types). 2) Whether the shot was a backhand. 3) Whether the shot was hit around the head.

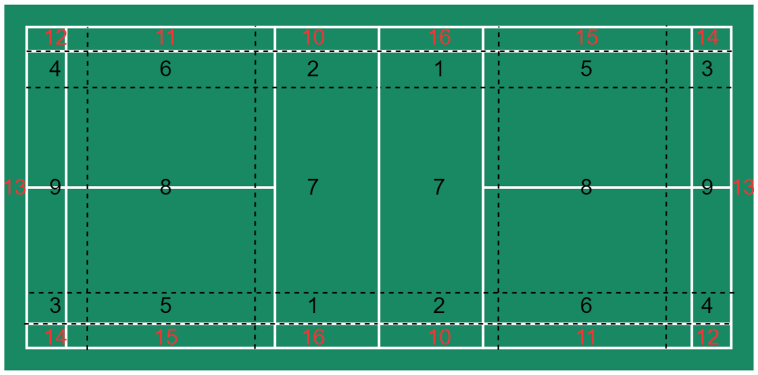

Figure 4.1: The grid layout of the badminton court.

4.2 Data Processing

4.2.1 Homography

The locations of the badminton shuttles in the original dataset is represented by the locations of the pixels in the videos of matches. Due to the perspective issue, the distribution of these locations of pixels form an trapezoid, shown in Figure 4.2, instead of a 1340 cm by 610 cm standard square badminton court. In order to transform the pixels into standard locations, we may use the homography transformation from Dubrofsky (2009).

A point $[x, y]^T$ under Cartesian coordinates in a 2D plane can be transformed into the corresponding point $[x, y, 1]^T$ under homography coordinates by adding a scaling factor $z = 1$ as the third element. And the points $[x, y, z]^T$ under homography coordinates can be converted back to the points $[x/z, y/z]^T$ by dividing the first two elements with the scaling factor z. And any points $[zx, zy, z]$ that is scaled from $[x, y, 1]$ are considered as the same point in homography coordinate.

The projective transformation can be achieved by establishing the following relation,

$$p' = Hp, \tag{4.1}$$

where $p' = [x', y', 1]^T$ represent the homogeneous coordinates of the point in the destination plane, $p = [x, y, 1]^T$ represent the homogeneous coordinates of the point in the source plane, and the H is a 3×3 homography matrix. Under our situation, the p represents the locations of the pixels in the videos, and p' represents the location of the corresponding shuttles in the real-world badminton courts, and H can be solved with at least four distinct pairs of p and p'. By manually marking the locations of the four corners of the badminton courts in the videos, we will be able to solve H combining with the four corners of the courts (i.e., (0, 0), (0, 610), (1340, 0), (1340, 610)) which are consistent since the standard court is 1340 cm in length and 610 cm in width.

Let the four distinct pairs of p and p' representing corresponding corners be $p'_i = [x'_i, y'_i, 1]$ and $p_i = [x_i, y_i, 1]$ for $i = 1, 2, 3, 4$, and

$$H = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix}.$$

The homography matrix H maps each point in the source image to its corresponding point in the destination image:

$$\begin{bmatrix} x'_i \\ y'_i \\ 1 \end{bmatrix} = H \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix}.$$

For each point pair, the above equation translates to two linear equations,

$$x'_i = \frac{h_{11}x_i + h_{12}y_i + h_{13}}{h_{31}x_i + h_{32}y_i + h_{33}}$$
$$y'_i = \frac{h_{21}x_i + h_{22}y_i + h_{23}}{h_{31}x_i + h_{32}y_i + h_{33}}.$$

These can be rewritten in terms of h_{ij},

$$(h_{31}x_i + h_{32}y_i + h_{33})x'_i = h_{11}x_i + h_{12}y_i + h_{13}$$
$$(h_{31}x_i + h_{32}y_i + h_{33})y'_i = h_{21}x_i + h_{22}y_i + h_{23}.$$

And these simplify to

$$\begin{aligned} h_{11}x_i + h_{12}y_i + h_{13} - h_{31}x_i x'_i - h_{32}y_i x'_i - h_{33}x'_i = 0 \\ h_{21}x_i + h_{22}y_i + h_{23} - h_{31}x_i y'_i + h_{32}y_i y'_i + h_{33}y'_i = 0. \end{aligned} \tag{4.2}$$

For each pair of points, we obtain two linear equations (4.2) in the unknowns $h_{11}, h_{12}, ..., h_{33}$. And since this system of linear equations is under determined, it should be transformed into a system of 8 linear equations of the form $Ah = b$, where

$$A = \begin{bmatrix} x_1 & y_1 & 1 & 0 & 0 & 0 & -x'_1x_1 & -x'_1y_1 \\ 0 & 0 & 0 & x_1 & y_1 & 1 & -x'_1x_1 & -x'_1y_1 \\ x_2 & y_2 & 1 & 0 & 0 & 0 & -x'_2x_2 & -x'_2y_2 \\ 0 & 0 & 0 & x_2 & y_2 & 1 & -x'_2x_2 & -x'_2y_2 \\ x_3 & y_3 & 1 & 0 & 0 & 0 & -x'_3x_3 & -x'_3y_3 \\ 0 & 0 & 0 & x_3 & y_3 & 1 & -x'_3x_3 & -x'_3y_3 \\ x_4 & y_4 & 1 & 0 & 0 & 0 & -x'_4x_4 & -x'_4y_4 \\ 0 & 0 & 0 & x_4 & y_4 & 1 & -x'_4x_4 & -x'_4y_4 \end{bmatrix}, h = \begin{bmatrix} h_{11} \\ h_{12} \\ \vdots \\ h_{32} \end{bmatrix}, b = \begin{bmatrix} x'_1 \\ y'_1 \\ x'_2 \\ \vdots \\ y'_4 \end{bmatrix}$$

under the assumption $h_{33} = 1$. To find the homography matrix H, solve for h using the generalized-inverse of A

$$h = A^{-1}b.$$

Then, we can form the homography matrix H under the assumption of $h_{33} = 1$ by rearranging the h,

$$H = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix}.$$

Once we have the matrix H, we can transform any point $p = [x, y, 1]^T$ from the videos to the corresponding point $p' = [x', y', z']^T$ in the destination plane, standard badminton courts, by multiplying H with p,

$$p' = Hp.$$

Finally, normalize the resulting homogeneous coordinates by dividing all elements with the scaling factor z' to get the Cartesian coordinates,

$$p' = [\tfrac{x'}{z'}, \tfrac{y'}{z'}, 1]^T = \tfrac{Hp}{z'}$$

And this $[\frac{x'}{z'}, \frac{y'}{z'}]^T$ provides us with the transformed coordinates in the standard badminton court plane in Firgure 4.3.

4.2.2 Qualifying Significant Characteristics

For further analysis the influences of gender and handedness of players, we added a dataset containing basic information about each player, including player names, gender (male or female), and handedness (left or right-handed).

In this project, we assume the serves have no difference regardless of they were made on left side or right side of the court. In order to analyze the correlation between distribution of serves other factors, we added a variable, "serve type", categorizing the serves by long and short according to the landing location (potential landing location as most of the serves were returned by opponents before actually landed) to the stroke-level data, as there is a clear separation in the distribution of counts of serves based on the Figure 4.4. A summary of serve types according to gender and handedness is in Table 4.3, which shows obvious higher proportion of serving long among female players compared to male players and is further investigated in later parts.

Potentially significant variables of interests added to the stroke-level dataset include 1) a binary variable indicating whether the player is ahead on the score each shot, 2) the lagged locations of shots and players to facilitate the analysis of influences among sequential shots, and 3) a binary variable indicating whether the landing locations of each shot are changed from one side to the other (left and right sides) compared to the location of previous shot.

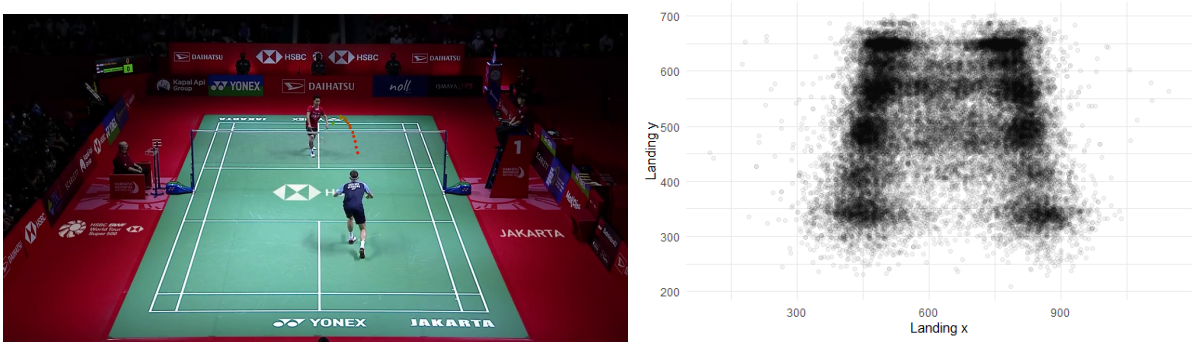

Figure 4.2: The distribution of landing locations (x, y) in raw data, shaped in a trapezoid.

Tournament	Round	Year	Set	Duration (minute)	Winner	Loser
Fuzhou Open 2018	Finals	2018	3	67	Kento Momota	Chou Tien Chen
Denmark Open 2018	Finals	2018	3	77	Kento Momota	Chou Tien Chen
Malaysia Open 2018	Quarter-Finals	2018	2	53	Kento Momota	Chou Tien Chen
Fuzhou Open 2019	Finals	2019	3	83	Kento Momota	Chou Tien Chen
World Tour Finals	Group Stage	2019	2	52	Chen Long	Chou Tien Chen
KOREA OPEN 2019	Finals	2019	2	53	Kento Momota	Chou Tien Chen
Denmark Open 2019	Quarter-Finals	2019	2	54	Chen Long	Chou Tien Chen
Fuzhou Open 2019	Semi-Finals	2019	2	45	Chou Tien Chen	Anders Antonsen
Sudirman Cup 2019	Quarter-Finals	2019	2	36	Chou Tien Chen	Jonatan Christie
Sudirman Cup 2019	Group Stage	2019	2	41	Chou Tien Chen	Ng Ka Long Angus
Indonesia Open 2019	Quarter-Finals	2019	3	76	Chou Tien Chen	Jonatan Christie
Thailand Masters 2020	Semi-Finals	2020	2	37	Ng Ka Long Angus	Shi Yu Qi
Malaysia Masters 2020	Semi-Finals	2020	2	44	Viktor Axelsen	Ng Ka Long Angus
Malaysia Masters 2020	Quarter-Finals	2020	3	65	Ng Ka Long Angus	Jonatan Christie
Malaysia Masters 2020	Quarter-Finals	2020	3	68	Viktor Axelsen	Chen Long
Malaysia Masters 2020	Finals	2020	3	54	Kento Momota	Viktor Axelsen
Indonesia Masters 2020	Quarter-Finals	2020	3	68	Anders Antonsen	Jonatan Christie
Indonesia Masters 2020	Finals	2020	3	71	Anthony Sinisuka Ginting	Anders Antonsen
Indonesia Masters 2020	Semi-Finals	2020	2	43	Anthony Sinisuka Ginting	Viktor Axelsen
All England Open 2020	Quarter-Finals	2020	3	68	Viktor Axelsen	Shi Yu Qi
HSBC BWF World Tour Finals 2020	Quarter-Finals	2020	3	55	An Se Young	Carolina Marin
HSBC BWF World Tour Finals 2020	Quarter-Finals	2020	3	50	Anthony Sinisuka Ginting	Lee Zii Jia
HSBC BWF World Tour Finals 2020	Quarter-Finals	2020	2	34	Evgeniya Kosetskaya	Michelle Li
HSBC BWF World Tour Finals 2020	Quarter-Finals	2020	2	45	Ng Ka Long Angus	Kidambi Srikanth
HSBC BWF World Tour Finals 2020	Quarter-Finals	2020	2	42	Pusarla V. Sindhu	Pornpawee Chochuwong
HSBC BWF World Tour Finals 2020	Semi-Finals	2020	3	42	Carolina Marin	Pornpawee Chochuwong
HSBC BWF World Tour Finals 2020	Semi-Finals	2020	3	60	Anders Antonsen	Viktor Axelsen
Yonex Thailand Open 2021	Quarter-Finals	2021	2	44	An Se Young	Ratchanok Intanon
Yonex Thailand Open 2021	Quarter-Finals	2021	2	45	Mia Blichfeldt	Busanan Ongbamrungphan
Yonex Thailand Open 2021	Quarter-Finals	2021	2	44	Ng Ka Long Angus	Lee Cheuk Yiu
Yonex Thailand Open 2021	Quarter-Finals	2021	3	66	Anthony Sinisuka Ginting	Rasmus Gemke
Toyota Thailand Open 2021	Quarter-Finals	2021	3	81	Anders Antonsen	Sameer Verma
Toyota Thailand Open 2021	Quarter-Finals	2021	2	34	Carolina Marin	Neslihan Yigit
Toyota Thailand Open 2021	Quarter-Finals	2021	3	77	Hans-Kristian Solberg Vittinghus	Lee Cheuk Yiu
Toyota Thailand Open 2021	Quarter-Finals	2021	2	41	Viktor Axelsen	Liew Daren
Toyota Thailand Open 2021	Quarter-Finals	2021	2	38	Ratchanok Intanon	Pusarla V. Sindhu
Toyota Thailand Open 2021	Semi-Finals	2021	2	51	Carolina Marin	An Se Young
Toyota Thailand Open 2021	Finals	2021	2	43	Hans-Kristian Solberg Vittinghus	Anders Antonsen
Toyota Thailand Open 2021	Finals	2021	2	40	Viktor Axelsen	Hans-Kristian Solberg Vittinghus

Table 4.1: Full match-level information

	Men's players	**# of Matches**	handedness
1	CHOU Tien Chen	11	right
2	Viktor AXELSEN	11	right
3	NG Ka Long Angus	7	right
4	Anders ANTONSEN	6	right
5	Kento MOMOTA	6	left
6	Anthony Sinisuka GINTING	5	right
7	Jonatan CHRISTIE	5	right
8	CHEN Long	3	right
9	Hans-Kristian Solberg VITTINGHUS	3	right
10	LEE Cheuk Yiu	2	right
11	SHI Yuqi	2	right
12	KIDAMBI Srikanth	1	right
13	LEE Zii Jia	1	right
14	LIEW Daren	1	right
15	Rasmus GEMKE	1	right
16	Sameer VERMA	1	right
	Women's players	**# of Matches**	handedness
1	Carolina MARIN	5	left
2	An Se Young	4	right
3	Pornpawee CHOCHUWONG	3	right
4	PUSARLA V. Sindhu	2	right
5	Ratchanok INTANON	2	right
6	Busanan ONGBAMRUNGPHAN	1	right
7	Evgeniya KOSETSKAYA	1	right
8	Mia BLICHFELDT	1	right
9	Michelle LI	1	right
10	Neslihan YIGIT	1	right
11	Supanida KATETHONG	1	left

Table 4.2: Number of Matches Played by Each Player (Grouped by Gender)

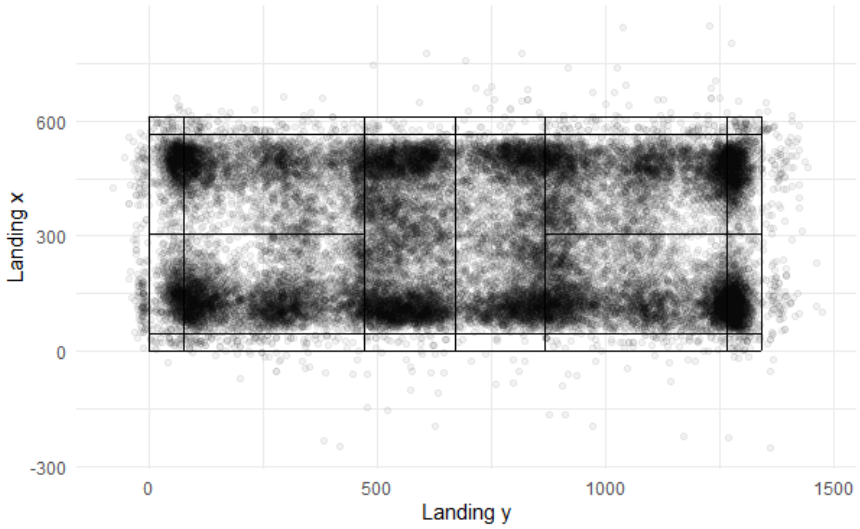

Figure 4.3: The landing locations (x, y) after homography transformation to standard 2D badminton court.

Gender	**Serve Type**	**Count**	**Proportion**
female	long	200	0.091
female	short	317	0.144
male	long	204	0.092
male	short	1475	0.671
Handedness	**Serve Type**	**Count**	**Proportion**
left	long	49	0.022
left	short	320	0.145
right	long	355	0.161
right	short	1472	0.670

Table 4.3: Summary of serve types by gender and handedness.

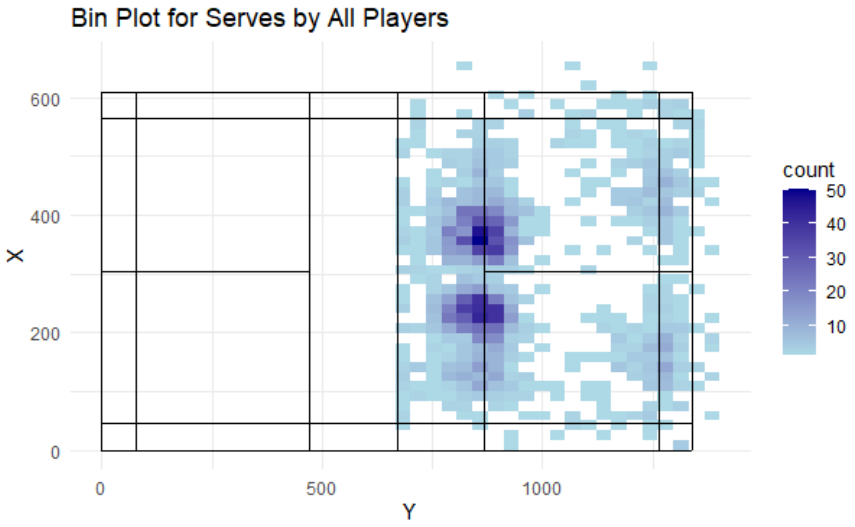

Figure 4.4: The distribution of counts of serves, which is clearly separated into two clusters, one in forecourt and one in backcourt.

CHAPTER 5. Methods

In this project, the objective is to analyze the factors contribute to the variation in patterns of shots and the winning of each point, such as gender, handedness, and spatial information. Then, generate actionable strategies for future training and matches. And these analysis is conducted through kernel density estimation, logistic regression models, and Bayesian inference.

5.1 Poisson Point Process

To analyze the distributions of shot across the badminton court, we model these distributions with Poisson point process (PPP) and estimate these distribution with kernel density estimation (KDE).

A Poisson point process in a 2D area is a collection of points $\{(x_i, y_i) \mid (x_i, y_i) \in \mathbb{R}^2\}$ such that

1. If $A \subset R^2$ is a bounded region, the number of points $N(A)$ in A follows a Poisson distribution,

$$P(N(A) = k) = \frac{\Lambda(A)^k e^{-\Lambda(A)}}{k!}, \text{ with } \Lambda(A) = \int_A \lambda(x)dx \text{ and } k = 0, 1, 2, \ldots, \tag{5.1}$$

   where the $\lambda(x)$ is intensity function and $\Lambda(A)$ the expected number of points in region A.

2. The number of points in disjoint regions are independent, that is $N(A_1), N(A_2), \ldots, N(A_n)$ are independent for $A_1, A_2, \ldots, A_n$ are disjoint subsets of $\mathbb{R}^2$.

The Poisson point process is Homogeneous PPP if $\lambda(x, y) = \lambda$ (constant). And the process is Non-Homogeneous PPP if $\lambda(x, y)$ varies acorss $\mathbb{R}^2$.

The estimation of the $\lambda(x, y)$ is completed through kernel density estimation (KDE) Baddeley et al. (2016),

$$\hat{\lambda}(x, y) = \sum_{i=0}^{n} K(x_i, y_i)w_i, \tag{5.2}$$

where w_i are the weights and $K(x_i, y_i)$ is the kernel function.

For this project, we apply this model to a 610×670 plane, A, in $\mathbb{R}^2$ according the dimension of left (right) court as the dimension of a standard badminton court is 610 cm by 1341 cm shown in Firgure 5.1. And we divide the left (right) court into 128×128 areas, $A_1, A_2, \ldots, A_n$ for $n = 1, 2, \ldots, 128$, with size $h_x \times h_y = 5.76$ cm $\times 5.72$ cm. Since the expected number of shots, intensity function $\lambda(x, y)$, varies across the badminton court, we assume the distributions of shots follows non-homogeneous PPP.

Under these conditions, we use estimator (5.2) with uniform weights

$$w_i = \frac{1}{h_x h_y} \tag{5.3}$$

and Gaussian kernel function

$$K(x, y) = \frac{1}{2\pi} e^{-\frac{1}{2}(x^2+y^2)}, \tag{5.4}$$

which is a bivariate normal density function centered at $(0, 0)$ and identity covariance.

Substituting these into the formula (5.2), the 2D KDE for each area $A_1, A_2, \ldots, A_n$ for $n = 1, 2, \ldots, 128$ is

$$\hat{\lambda}(x, y) = \frac{1}{n h_x h_y 2\pi} \sum_{i=0}^{n} \exp(-\frac{1}{2}((x_i - x)^2 + (y_i - y)^2)). \tag{5.5}$$

In this project, we applied the Poisson point process and kernel density estimation on exploratory analysis of serves by genders and handedness of players and the estimation of transition probability from current shots to next shots given previous shots patterns. The KDE is applied with `density` function in `spatstat` package in R, with output graphs as Firgure 6.1.

5.2 Logistic Regression

In order to identify the significant factors influencing the distribution of serves, the patterns of shots other than serves, and the progressions of scores by players, logistic regression with / without mixed effects models are applied with different methods of estimation.

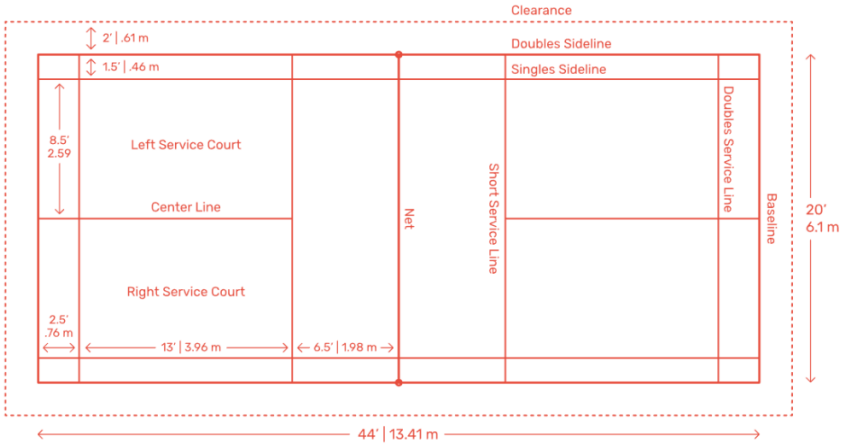

Figure 5.1: The dimension of a standard badminton court.

The general logistic regression with/without mixed effects model in this project is defined as below

$$Y_i \overset{\text{ind}}{\sim} \text{Bernoulli}(p_i), \tag{5.6}$$

where Y_i is the i^{th} observation and p_i is the probability of success of the i^{th} observation.

The log link function is

$$\log\left(\frac{p_i}{1-p_i}\right) = x_i^T\beta + z_i^T b, \tag{5.7}$$

where $x_i^T \in \mathbb{R}^m$ is vector containing the intercept and values of explanatory variables in the i^{th} observation , the $\beta \in \mathbb{R}^m$ is the vector representing the fixed effects of the variables in x_i^T, the $z_i^T \in \mathbb{R}^l$ is the vector of observed constants in the i^{th} observation, and the $b \in \mathbb{R}^l$ is the vector of random effects assumed to follow a normal distribution

$$b_i \overset{ind}{\sim} N(0, \sigma^2) \tag{5.8}$$

.

The likelihood function of this logistic regression with mixed effects model is

$$\begin{aligned} L(p) &= \prod_{i=1}^{n} p_i^{Y_i}(1-p_i)^{1-Y_i} \\ &= \prod_{i=1}^{n} \left[\frac{e^{x_i^T\beta + z_i^T b}}{1+e^{X_i\beta + z_i^T b}}\right]^{Y_i} \left[\frac{1}{1+e^{X_i\beta + z_i^T b}}\right]^{1-Y_i} \end{aligned} \tag{5.9}$$

5.2.1 Serves

To analyze serves, we simplify the serve by categorizing it as "long" or "short" serve according to the criterion of whether passing the line of $y = 1104$ shown in Figure 5.2, which captures the main characteristics of the distribution of the serves shown in Figure 6.1 from the exploratory analysis. The factors, including "gender" (male or female), "handedness" (left or right), "score", "type of last serve", "serve side" (left or right), and "players", are potential significant factors influencing the next serves based on our data.

5.2.1.1 Variable Selection

The first logistic regression model, without random effects, used for variable selection follows model (5.6) where Y_i is the i^{th} observation of serve with $Y_i = 1$ for short serve and $Y_i = 0$ for long serve and p_i is the probability of serving short.

The systematic component has $x_i^T \in \mathbb{R}^7$ is vector containing the intercept, "gender", "handedness", "score", "last serve", "serve side", and "players" observed in the i^{th} observation of serve, the $\beta \in \mathbb{R}^7$ is the vector representing the coefficients of the factor in x_i^T. And this logistic regression model is estimated with `glm` from `stats` package in R.

A likelihood ratio test, conducted using the `anova` from `stats` package in R, is applied to this model without random effects to identify the statistically significant factors influencing the probability of serving short. In the result, "gender" (male or female), "handedness" (left or right), and "players" are the variables identified as statistically significant.

5.2.1.2 Estimation

The estimation and uncertainties of the probabilities of serving short by players are qualified with a logistic regression with mixed effect model(5.6) to encounter the variation of players. And the model is defined with Y_i is the i^{th} observation of serve with $Y_i = 1$ for short serve and $Y_i = 0$ for long serve and p_i is the probability of serving short.

And the systematic component has $x_i^T \in \mathbb{R}^4$ is vector containing intercept and the values of "gender" (male or female), "handedness" (left or right), and "serve side" (left or right) observed in the i^{th} observation , the $\beta \in \mathbb{R}^4$ is the vector representing the fixed effect of the factor in x_i^T, the $z_i^T \in \mathbb{R}^{27}$ is the vector of player indicators, and the $b \in \mathbb{R}^{27}$ is the vector of random effects.

The default weakly informative priors of β defined by `stan_glmer` function are

$$\beta_0 \sim N(0, 2.5^2),\ \beta_1 \sim N(0, 6.68^2),\ \beta_2 \sim N(0, 5^2),\ \text{and}\ \beta_3 \sim N(0, 5.89^2), \tag{5.10}$$

and the distribution of b, in `stan_glmer` function, is defined as

$$b \sim MVN(0, \Sigma), \tag{5.11}$$

where Σ is a diagonal matrix with σ^2 on the diagonal, following

$$\Sigma = DRD, \tag{5.12}$$

where $D = \sigma I$ is matrix of the standard deviations of the random effects and $R = I$ in this case. And the default weakly informative priors of σ defined by `stan_glmer` function follow

$$\sigma \sim \text{Half-Cauchy}(scale = 1). \tag{5.13}$$

Following the likelihood function(5.9) and priors defined above, the effect and uncertainties of serving "short" for each player ara estimated through Hamiltonian Monte Carlo, Betancourt (2017), and Bayesian inference with `stan_glmer` function in `rstanarm` package in R, with 4 independent chains having random selected initial value number of warm-up as 1000 and inferential iterations as 10000.

5.2.2 Patterns of Other Shots

To identify significant factors influencing the distribution of shot other than serves for all the players together, we apply logistic regression models similar to previous part. Since the badminton court is divided in to 9 area shown in the Figure 4.1, we applied 9 logistic regression model comparing the probabilities of shots landing in each of the area and the probabilities of shots landing other 8 areas.

Consider the case of comparing the probabilities of shots landing in area 1 and the probabilities of shots landing other 8 areas, the first logistic regression model, without random effects, used for variable selection follows model (5.6) where Y_i is the i^{th} observation of serve with $Y_i = 1$ for shots landing in area 1 and $Y_i = 0$ for shots landing in other areas and p_i is the probability of shots landing in area 1.

The systematic component has $x_i^T \in \mathbb{R}^9$ is vector containing the intercept, "match", "handedness", "cumulative rallies in a match", "winning on score", "hitting location" (x, y), and "opponent location" (x, y) observed in the i^{th} observation of shot, the $\beta \in \mathbb{R}^9$ is the vector representing the coefficients of the factor in x_i^T. And this logistic regression model is estimated with `glm` from `stats` package in R.

A likelihood ratio test, conducted using the `anova` from `stats` package in R, is applied to this model without random effects to identify the statistically significant factors influencing the probability of serving short. In the result, "hit location" and "opponent location" are the variables identified as statistically significant.

Then, to estimate the probability of shots landing in area 1 and quality the uncertainty of this estimation, we applied a logistic regression with mixed effects model(5.6). The Y_i is same as previous definition. The fixed effect has $x_i^T \in \mathbb{R}^{17}$ is vector containing the intercept, "hitting area" (1 to 9), and "opponent location area" (1 to 9) observed in the i^{th} observation of shot, the $\beta \in \mathbb{R}^{17}$ is the vector representing the coefficients of the factor in x_i^T. The $z_i^T \in \mathbf{R}^{27}$ is the vector of random effect of "players", and the $b \in \mathbb{R}^{27}$ is the vector of random effects. And this logistic regression model is estimated with `glmer` from `lme4` package in R.

In addition to applying this logistic regression with random effects model to all players, we also applied it to the male player with most data, Viktor Axelsen, by replacing the random effect part with $z_i^T \in \mathbf{R}^{11}$ is the indicator vector of "opponents", and the $b \in \mathbf{R}^{11}$ is the vector of random effects.

Figure 5.2: The criterion (red line) characterizing the distribution of all serves. Serves on the left side of the criterion ($y = 1140$) are labeled as "short" and Serves on the right side are labeled as "long".

CHAPTER 6. Results of Next Shot Prediction

Based on results from the methods mentioned in previous section, we are able to analyze the influences of factors, including player and spatial information, on the patterns of serves and other shots for all players. Besides, the significant factors contributing the winning of each point of a specific player can be investigated to generate applicable strategies for training and matches.

6.1 Analysis of All Players

6.1.1 Serves

The Figure 6.1 are results generated from KDE mentioned in section 6.1, showing results of female players (left column) and male players (right column) served on the same side with same handedness on the same row. From these plots, we can observe that the probabilities of female players making long serves is dramatically higher compared to male players regardless of the side of court the serves were made and the handedness of the players.

A summary of the results from the logistic regression with mixed effect model mentioned in section 6.2.1.2 in presented in Table 6.1. From this table, we may assume the Hamiltonian Monte Carlo algorithm converges properly as the $\hat{R}$ for all coefficients are close to 1. The Figure 6.4 and Figure 6.5 show the estimated probabilities of serving short for each players with 95% credible intervals colored by handedness and gender. An interesting finding in Table 6.1 is that the "serve side" and "handedness" were identified as non-significant based on the 95% credible interval, though they are tested as significant in likelihood ratio tests. From the Figure 6.4, we can observe that female players tend to have lower probabilities of serving short compared to male players where as handedness does not play a significant role deciding the type of serves, which is consistent to our exploratory analysis based on KDE.

Therefore, we may expect the encounter short serves when facing male opponents and long serves when facing female opponents, allowing us to prepare more sufficiently on returning the serves and following shots.

6.1.2 Shots Other than Serves

In addition to serves, analysis of distribution of shot were conducted using KDE and the logistic regression models mentioned in section 6.2.2. From Figure 6.2, we can see that the landing sides (left or right) of most shots are consistent with the hit side. And Figure 6.3 shows that the landing sides (front or back, left or right) tends to be on the opposite side with the opponent locations. These results form exploratory analysis shows that "hit area" and "opponent area" are potentially significant to distribution of landing location.

Based on the result of likelihood ratio test in Table 6.2 of the model analyzing factors influencing probability of shots (other than serves) landing in area 1, "cumulative rally", "hit area", and "opponent location" are identified as statistically significant with significance level of 0.05. The results of variable selection for the other 8 models in Table 6.3 analyzing probabilities of shots landing in the other 8 areas. From these results, it is obviously that only "hit area" and "opponent location" are identified as statistically significant constantly among all the models, which lead to consistent results from exploratory analysis based on KDE.

Thus, we apply the factors identified as significant, "hit area" and "opponent location", into the final model predicting the probability of shots (other than serves) landing in each area. The Figure 6.6 present the predicted probabilities of next shot landing in each area given "opponent location" is "1" for each "hit area" based on the final models generated previously. From Figure 6.6, we can see that regardless of the "hit area", the probability of next shot landing in area 8 is always dramatically higher given the opponent is in area 1. When the opponent is in area 2, the places where next shots are mostly likely to land in area 5 and 8, and area 7 is likely to hit to when players are in mid-court areas and the middle of forecourt areas, from Figure 6.7.And given the opponent player is in backcourt area 3, 4, and 9, the next shot would likely to land in

forecourt and mid court areas, area 1, 2, 7, and 8, and area 5 is likely to be hit to when players are in mid-court and backcourt areas from Figure 6.7 and Figure 6.9. When the opponent player stands in mid-court areas, area 5, 6, and 8, the probabilities of next shots landing in forecourt and mid-court areas is obviously higher compared to backcourt areas from Figure 6.8 and Figure 6.9.

6.2 Analysis of the Match Data of Viktor Axelsen

From Table 4.2, Viktor Axelsen is one of the players with most data in our dataset. Since Viktor Axelsen also remains to be the top 5 male badminton player in the ranking from 2018 to 2021, analyzing the strategies against him is also one of our primary objective of this project. To achieve this goal, we analyze both the distribution of future shots given previous opponent and spatial information with in that rally and the factors that specifically contributed to the point-level success in a match.

Using the logistic regression models mentioned in section 6.2.2 and likelihood ratio tests in Table 6.4, "hit area" is identified as statistically significant in all of the cases and "opponent location" is identified as significant in 7 of the 9 cases, whereas "handedness", "cumulative rally", and "if ahead on points" are only identified as significant under certain cases.

Then, we can include the "hit area" and "opponent location" into the final model predicting the probability of shots (other than serves) landing in each area. Given the opponents stand in area 1, Viktor Axelsen tends to make next shot landing in area 8, 6, and 1, from Figure 6.10, forcing opponent to use back-steps for next shot while also limiting the chance for oppoent to attack. If the opponent is in area 2, the probability for next shot landing in area 7 and 8 is dramatically higher than other probability based on Figure 6.11. Under same situation, Viktor Axelsen also tends to make next shot landing in area 3 when he is in the forecourt, forcing the opponent to move the farthest distance as possible while also saving himself more time to get prepared for next shot. When the opponent is in the backcourt, area 3, 4, and 9, the probability of next shot lands in forecourt and mid-court area is obviously higher, either forcing opponent to lift the shuttle up when he is also in forecourt area or attacking when he is already in the

backcourt from Figure 6.11 and Figure 6.13. Given the opponent is in mid-court areas, area 5 and 6, the probabilities of next shot landing in backcourt areas, area 3, 4, and 9, are generally lower when Viktor Axelsen is in the mid-court or backcourt areas from Figure 6.12, which can effectively limit the chance for opponent to attack. Regardless of the current position of the opponents, the prabobilities of next shot landing in area 5 and 6 increase as Viktor Axelsen moving the the backcourt, especially in area 4 and 5, meaning he tends to attack when he is in the backcourt areas.

Compared to the general results for all players in section 6.1.2, the pattern of non-serve shots of Viktor Axelsen is mainly distinguished by situations where opponents are in forecourt areas, area 1 and 2. When opponent is in area 1, Viktor Axelsen tends to make attacks to area 6 when he is in backcourt areas, instead making attacks to area 5 compared to other players, shown in Figure 6.10 and from Figure 6.6. Similarly, Viktor Axelsen nearly make no shots to area 5 when he is in backcourt areas and opponents are in area 2, where other players prefer to return shots to area 5.

6.3 Potential Training Plans and Match Strategies against Viktor Axelsen

Based on the analysis of match data of Viktor Axelsen from previous section, we generate some detailed suggestions for potential training plans and match strategies when preparing to play against Viktor Axelsen in future matches.

6.3.1 Anticipate Specific Landing Areas

Viktor Axelsen frequently makes shots toward middle court areas, area 7 and 8, when he is in the backcourt areas and the opponents are in forecourt areas, area 1 and 2, which is attacking with quick smash before the opponents return to proper positions of defense. Thus, players should practice quick and precise foot movements returning to mid-court areas from forecourt areas, emphasizing on lateral agility and transition from front court positions.

When opponents lift the shuttles toward backcourt areas from area 1 and 2, Viktor Axelsen mostly return the shots to area 6 only (left side of mid-court), likely quick attacks, regardless of standing side (left and right) of opponents. Therefore, players should be prepared to defense in area 6 when they lift the shuttles toward backcourt areas from forecourt areas.

6.3.2 Prepare for Forecourt and Mid-court Engagements

Viktor Axelsen often place shots in forecourt and mid-court areas, disirupting opponents' paith and positioning, when opponents are in backcourt areas. Thus, opponents should be prepared for forward movements toward forecourt areas if they are currently located in backcourt areas. Training focusing on quick forward movements would also be helpful.

6.3.3 Improve Backward Movement and Overhead Techniques

When Viktor Axelsen hits from forecourt positions, he often sends the shuttle deep, forcing opponents backward. Players should be prepared for backward movements when Viktor Axelsen is passively returning shots from forecourt areas, and reacting with overhead shots, such as clears, smashes, and drops.

6.3.4 Avoid Vulnerable Mid-court Positions

Axelsen strategically reduces opportunities for opponents to attack by minimizing shots to backcourt when opponents are already in mid-court. Players should anticipate repositioning from mid-court to forecourt, preparing for high quality net shot to avoid becoming targets of Axelsen's controlled attacks.

6.3.5 Physical Conditioning for Long Rallies

Axelsen employs strategies that move opponents around the court extensively. Players should invest significantly in physical conditioning, stamina, and speed endurance training, preparing physically and mentally for prolonged exchanges and frequent directional changes.

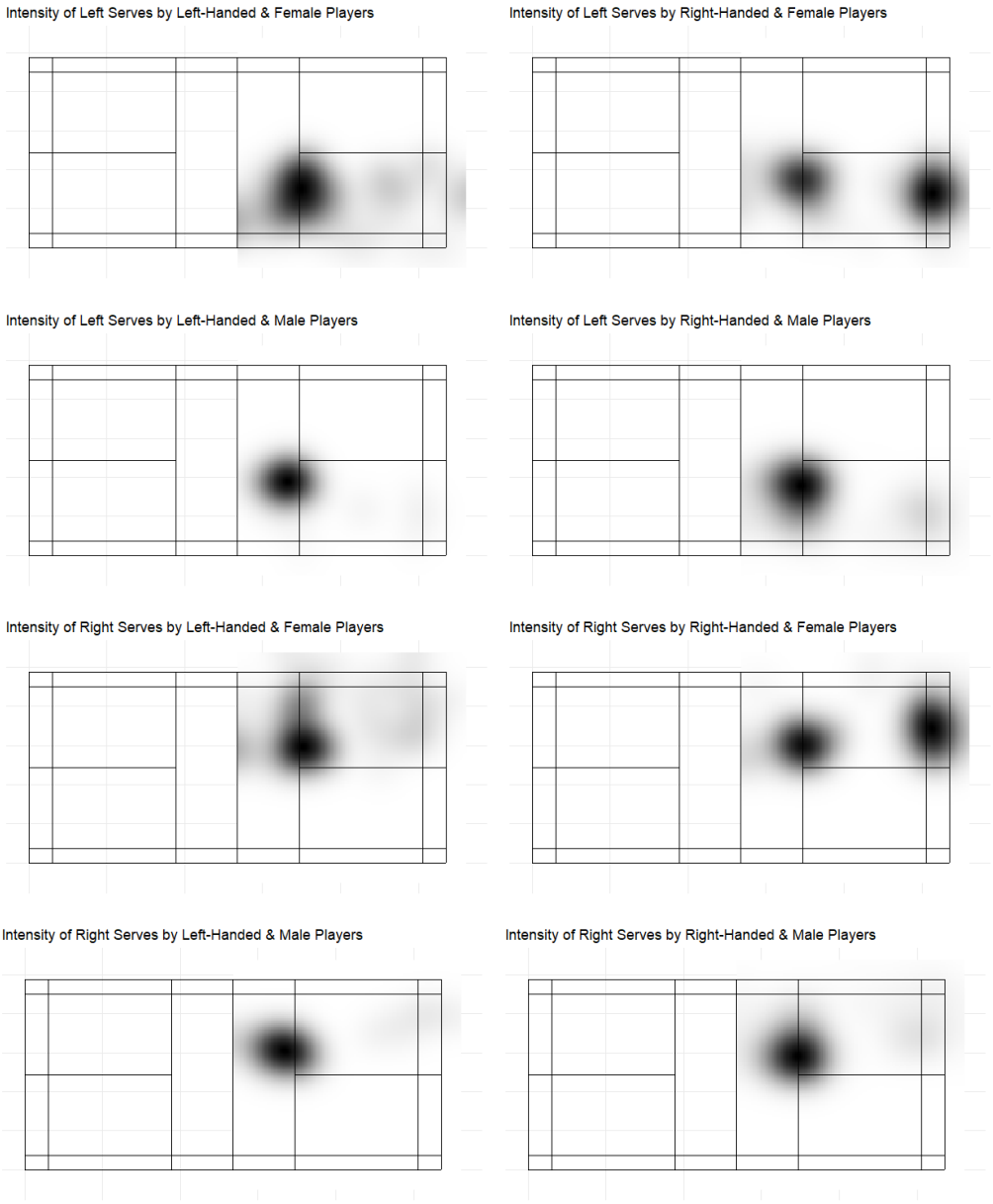

Figure 6.1: The kernel density estimation of the distributions of serves shows that female players (left column) tends to make more long serves compared to male players (right column) regardless of handedness (rows).

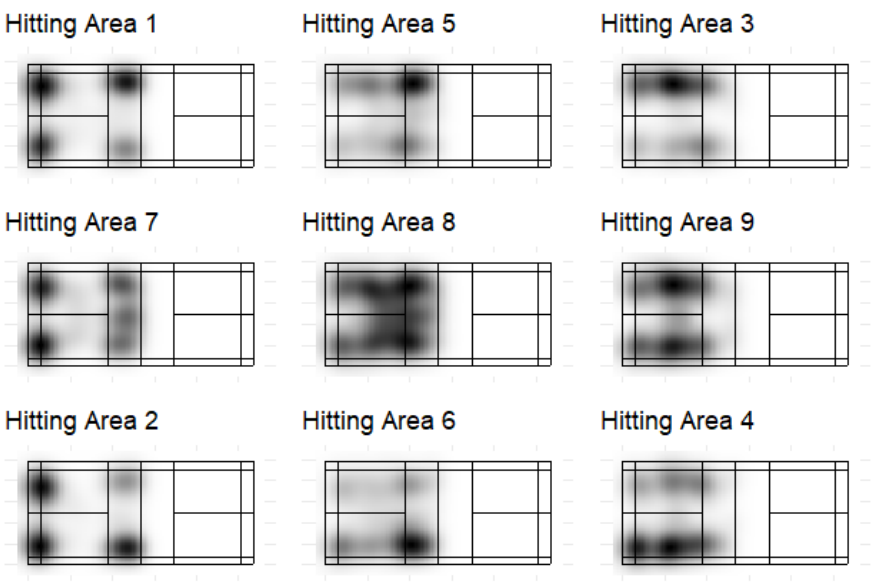

Figure 6.2: The kernel density estimation of the distributions of landing location categorized by hit areas, which shows landing location of shots tends to be on the same side (front or back, left or right) with hit location.

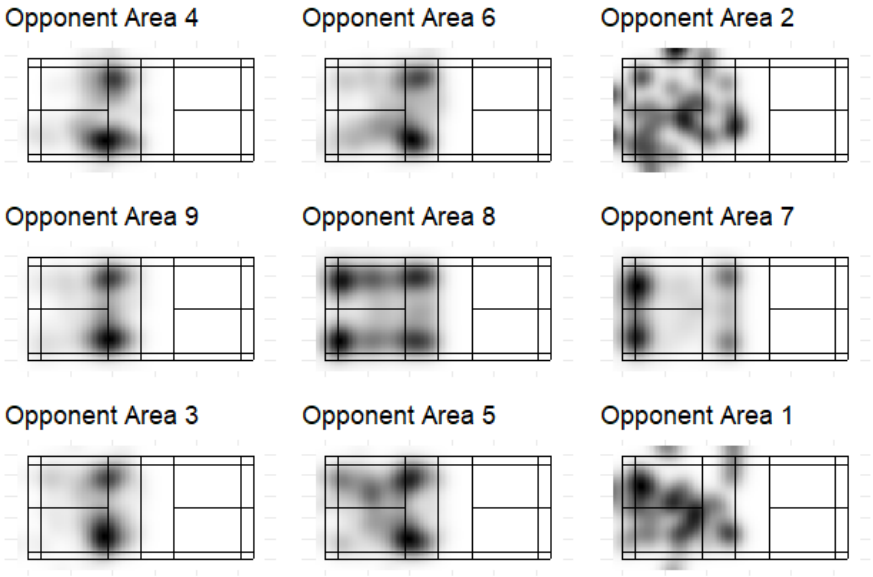

Figure 6.3: The kernel density estimation of the distributions of landing location categorized by opponent areas, which shows landing location of shots tends to be on the opposite side (front or back, left or right) with hit location.

	mean	sd	2.5%	97.5%	R_hat
(Intercept)	0.76	0.83	-0.99	2.48	1.00
handednessright	-0.94	0.87	-2.74	0.91	1.00
serve_sideright	-0.03	0.12	-0.27	0.21	1.00
gendermale	2.74	0.59	1.55	4.00	1.00
Random Effects (Intercept by Players)					
An_Se_Young	0.69	0.51	-0.35	1.72	1.00
Anders_ANTONSEN	-1.62	0.42	-2.49	-0.81	1.00
Anthony_Sinisuka_GINTING	0.15	0.49	-0.83	1.14	1.00
Busanan_ONGBAMRUNGPHAN	-1.69	0.82	-3.55	-0.22	1.00
Carolina_MARIN	0.90	0.84	-0.83	2.66	1.00
CHEN_Long	1.36	0.76	0.03	3.11	1.00
CHOU_Tien_Chen	-0.28	0.45	-1.19	0.59	1.00
Evgeniya_KOSETSKAYA	0.24	0.57	-0.91	1.40	1.00
Hans-Kristian_Solberg_VITTINGHUS	-0.49	0.53	-1.55	0.58	1.00
Jonatan_CHRISTIE	-0.29	0.47	-1.24	0.66	1.00
Kento_MOMOTA	-0.92	0.90	-2.82	0.91	1.00
KIDAMBI_Srikanth	1.47	0.98	-0.19	3.89	1.00
LEE_Cheuk_Yiu	0.90	0.79	-0.50	2.72	1.00
LEE_Zii_Jia	0.68	0.81	-0.80	2.53	1.00
LIEW_Daren	-1.26	0.57	-2.40	-0.13	1.00
Mia_BLICHFELDT	2.39	0.94	0.78	4.68	1.00
Michelle_LI	-2.65	0.89	-4.83	-1.13	1.00
Neslihan_YIGIT	0.88	0.63	-0.36	2.19	1.00
NG_Ka_Long_Angus	2.02	0.72	0.75	3.69	1.00
Pornpawee_CHOCHUWONG	-0.24	0.55	-1.35	0.86	1.00
PUSARLA_V._Sindhu	-0.47	0.61	-1.73	0.76	1.00
Rasmus_GEMKE	-1.14	0.67	-2.45	0.21	1.00
Ratchanok_INTANON	-0.21	0.55	-1.34	0.91	1.00
Sameer_VERMA	0.23	0.87	-1.34	2.16	1.00
SHI_Yuqi	0.84	0.79	-0.58	2.71	1.00
Supanida_KATETHONG	0.21	0.86	-1.54	1.99	1.00
Viktor_AXELSEN	-1.28	0.41	-2.14	-0.48	1.00

Table 6.1: Summary of Bayesian Model of Serves

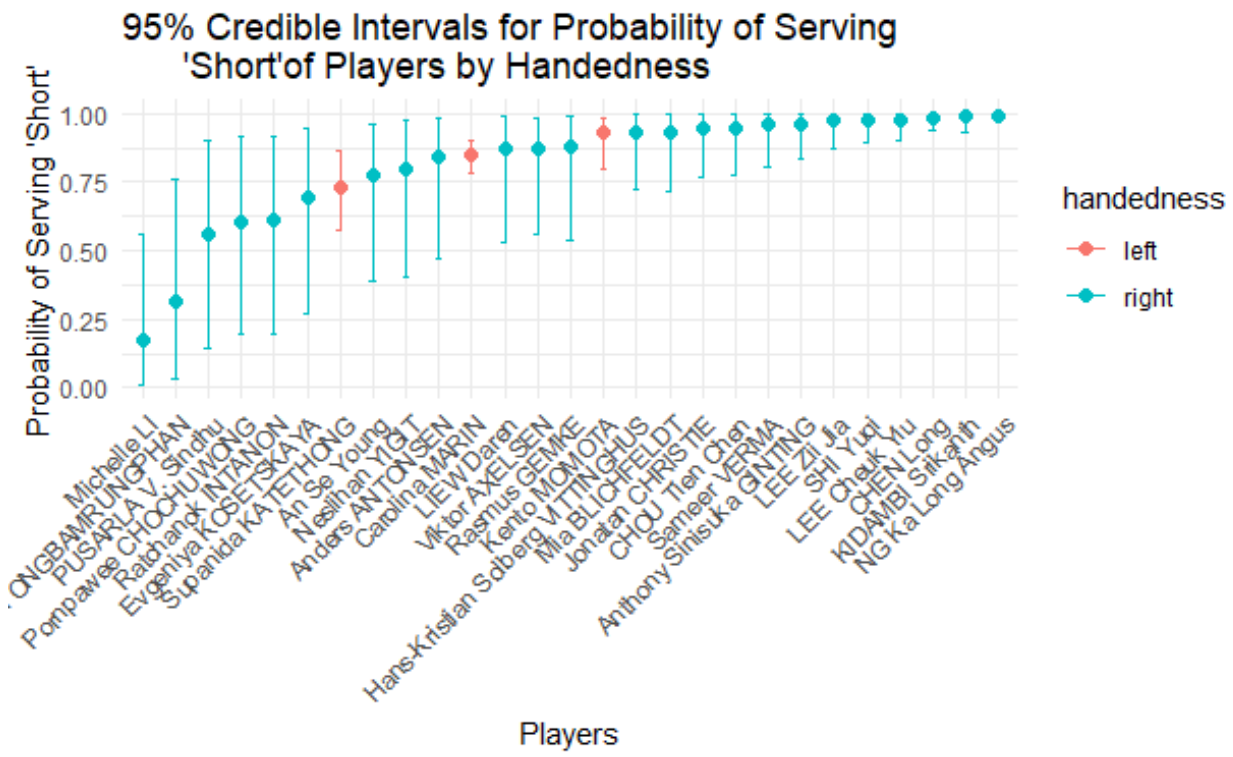

Figure 6.4: The estimated probabilities of serving "short" shows that the handedness is not significant on predicting the type of the serves.

Area 1	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
handedness	1	0.65	27257	21023.77	0.4187
cum_rally	1	3.96	27256	21019.82	0.0467
ahead	1	0.65	27255	21019.16	0.4186
hit_area	14	583.19	27241	20435.97	0.0000
opponent_location_area	12	130.32	27229	20305.65	0.0000

Table 6.2: Test of significance factors influencing the probability of shots (other than serves) landing in area 1. From this table, cumulative rally, hit area, and opponent location are significant factors influencing the landing area of next shot.

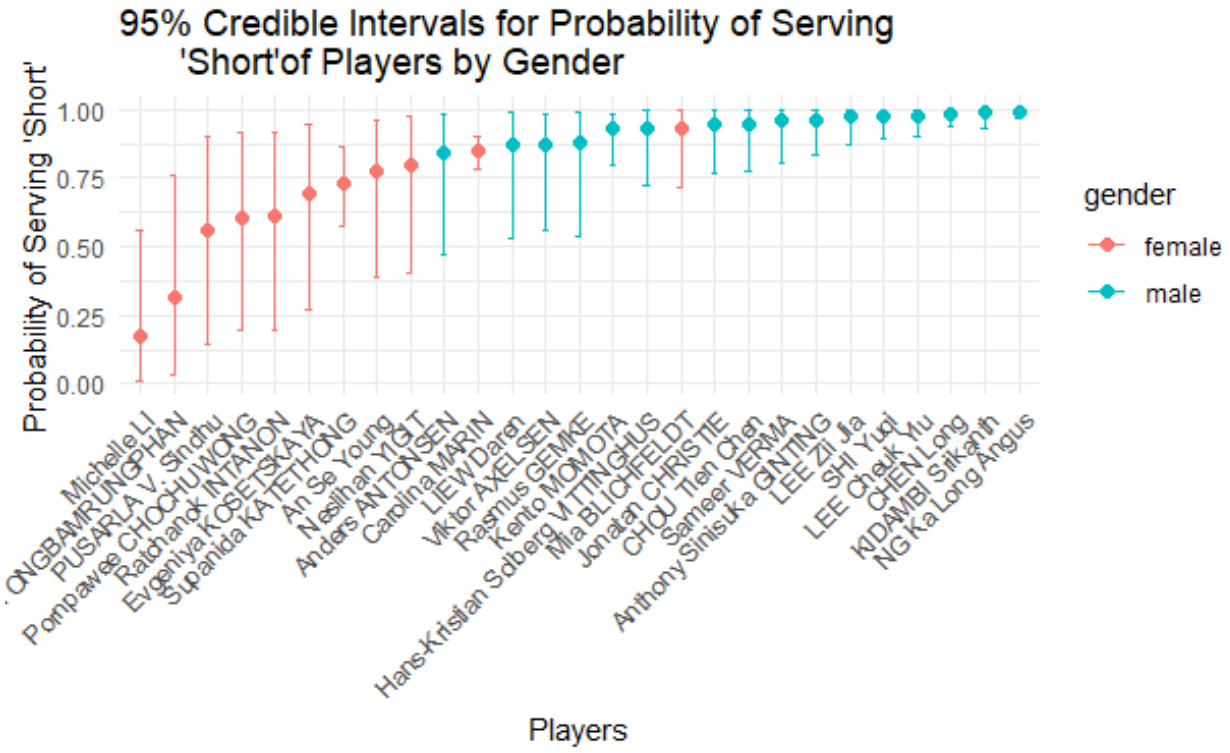

Figure 6.5: The probabilities of serving "short" shows that the gender is obviously significant on the predicting the type of serves.

Area 2	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
handedness	1	4.80	27257	20674.18	0.0285
cum_rally	1	0.03	27256	20674.15	0.8603
ahead	1	1.28	27255	20672.86	0.2571
hit_area	14	583.28	27241	20089.58	0.0000
opponent_location_area	12	58.28	27229	20031.31	0.0000
Area 3	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
handedness	1	19.31	27257	16081.10	0.0000
cum_rally	1	1.51	27256	16079.59	0.2191
ahead	1	2.99	27255	16076.60	0.0836
hit_area	14	821.89	27241	15254.71	0.0000
opponent_location_area	12	128.02	27229	15126.69	0.0000
Area 4	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
handedness	1	12.48	27257	13653.82	0.0004
cum_rally	1	16.15	27256	13637.67	0.0001
ahead	1	12.29	27255	13625.37	0.0005
hit_area	14	722.78	27241	12902.60	0.0000
opponent_location_area	12	89.10	27229	12813.49	0.0000
Area 5	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
...					
Area 7	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
handedness	1	0.96	27257	19397.08	0.3276
cum_rally	1	7.20	27256	19389.88	0.0073
ahead	1	0.19	27255	19389.69	0.6646
hit_area	14	1522.93	27241	17866.76	0.0000
opponent_location_area	12	54.69	27229	17812.07	0.0000
Area 8	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
handedness	1	4.97	27257	21329.62	0.0259
cum_rally	1	9.24	27256	21320.38	0.0024
ahead	1	11.43	27255	21308.95	0.0007
hit_area	14	797.62	27241	20511.33	0.0000
opponent_location_area	12	113.23	27229	20398.10	0.0000
Area 9	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
handedness	1	0.00	27257	13640.23	0.9627
cum_rally	1	0.20	27256	13640.03	0.6527
ahead	1	7.56	27255	13632.47	0.0060
hit_area	14	743.94	27241	12888.53	0.0000
opponent_location_area	12	252.09	27229	12636.44	0.0000

Table 6.3: Test of significance factors influencing the probability of shots (other than serves) landing in other 8 areas. From these tables, we can see that hit area and opponent location are always significant factors influencing the landing area of next shot in all models.

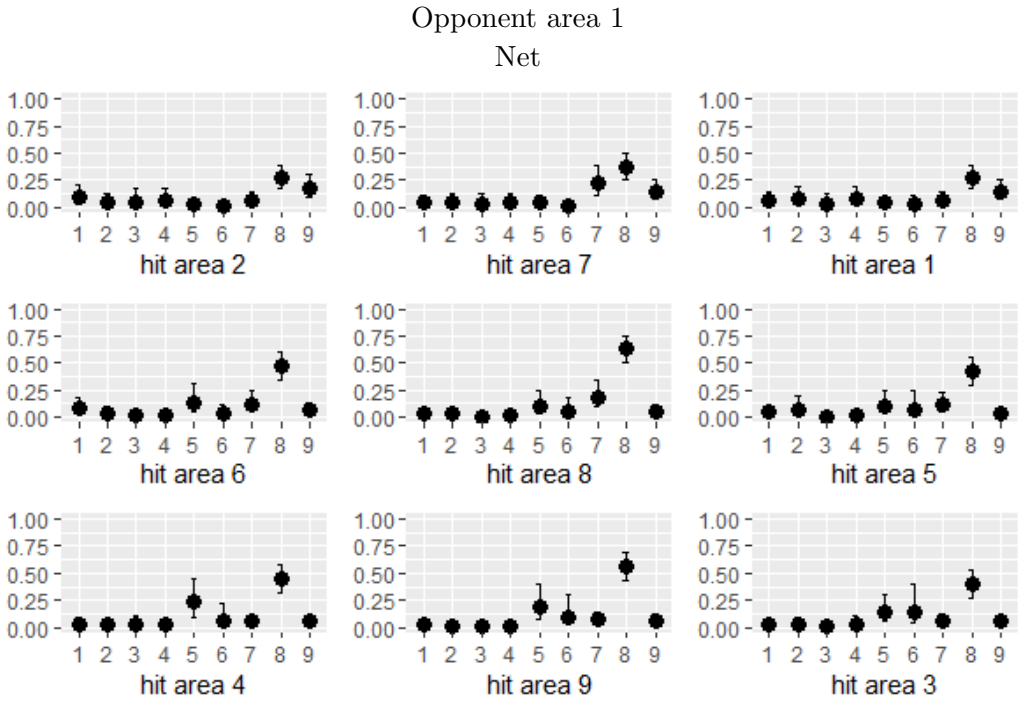

Figure 6.6: The predicted probabilities of next shot landing in each area given "opponent location" is "1" and "hit area". For each of the plot, the x axis represents the "landing area" and y axis represents the probabilities.

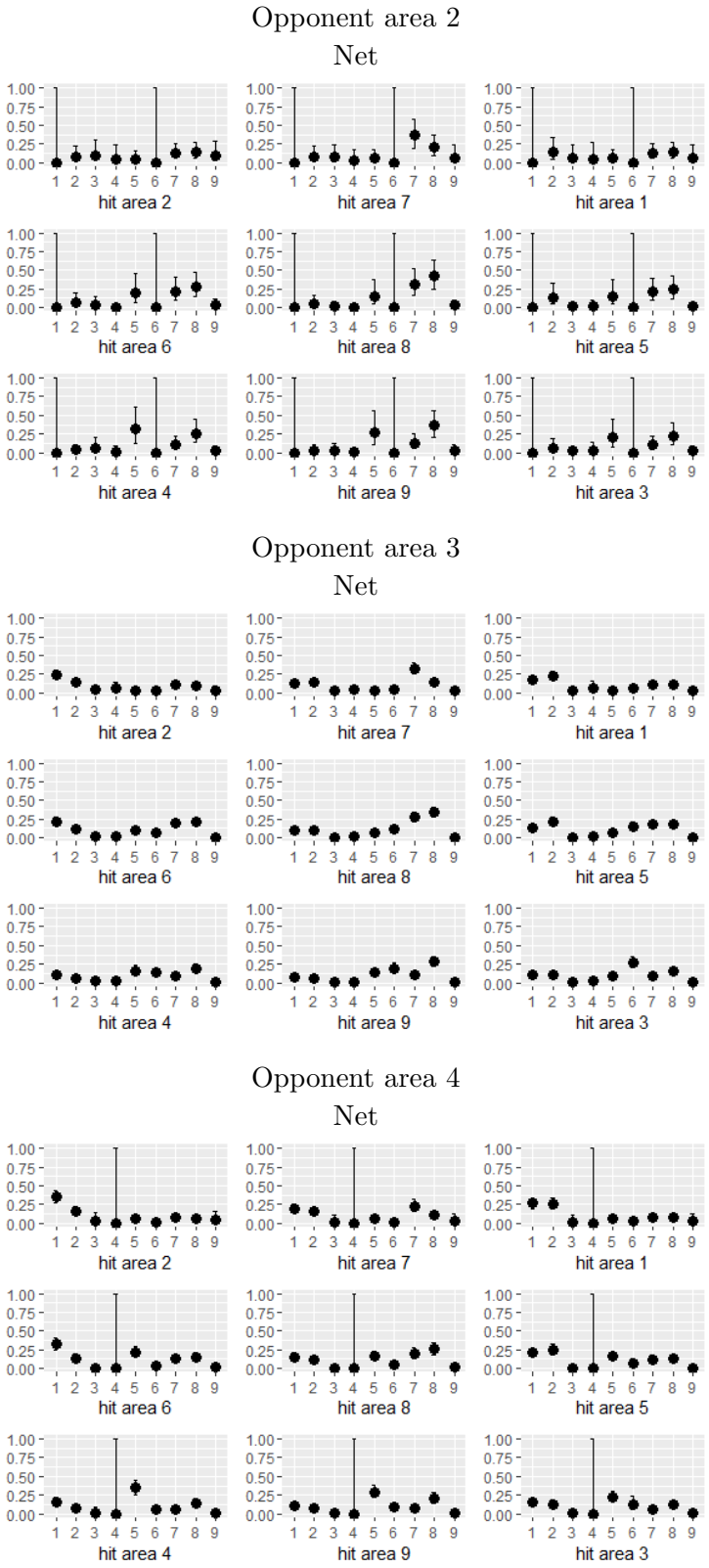

Figure 6.7: The predicted probabilities of next shot landing in each area given "opponent location" are "2", "3", and "4", and "hit area". For each of the plot, the x axis represents the "landing area" and y axis represents the probabilities.

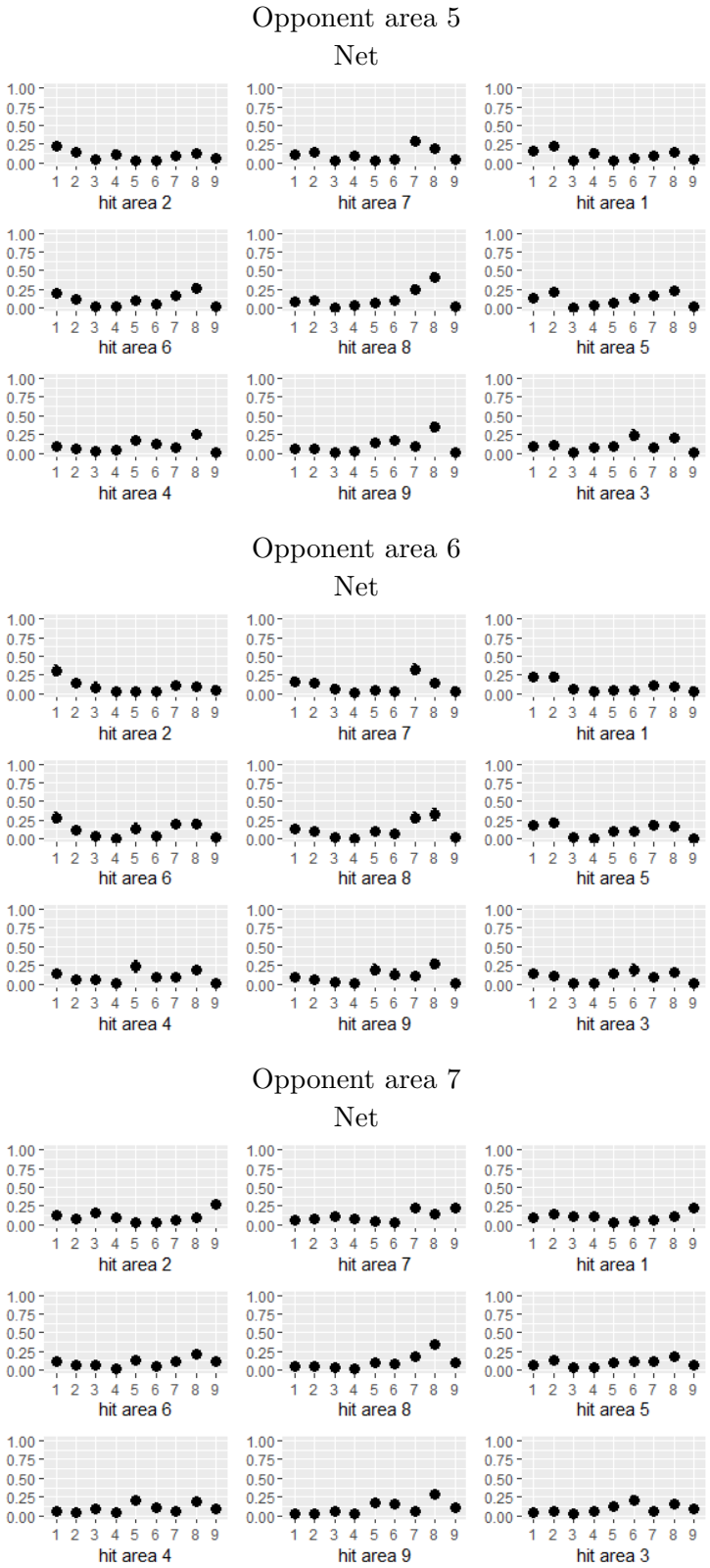

Figure 6.8: The predicted probabilities of next shot landing in each area given "opponent location" are "5", "6", and "7", and "hit area". For each of the plot, the x axis represents the "landing area" and y axis represents the probabilities.

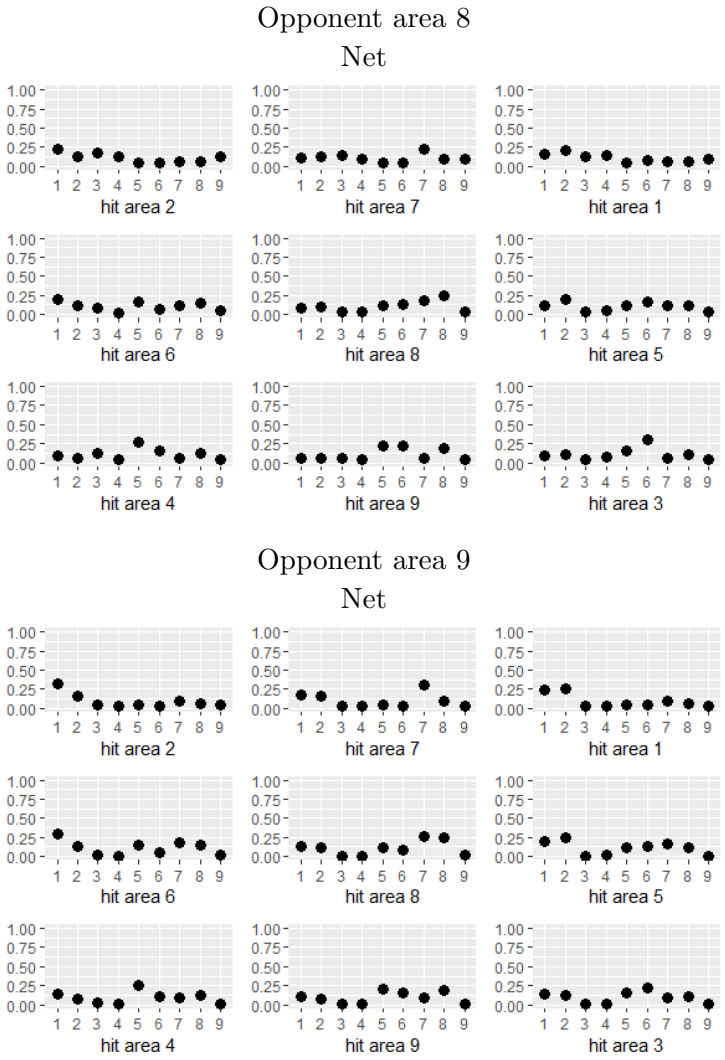

Figure 6.9: The predicted probabilities of next shot landing in each area given "opponent location" are "8" and "9" and "hit area". For each of the plot, the x axis represents the "landing area" and y axis represents the probabilities.

Area 1	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
handedness	1	3.20	6617	5029.54	0.0735
cum_rally	1	2.75	6616	5026.78	0.0971
ahead	1	0.41	6615	5026.38	0.5237
hit_area	13	158.10	6602	4868.28	0.0000
opponent_location_area	12	35.58	6590	4832.70	0.0004
Area 2	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
handedness	1	6.75	6617	4881.28	0.0094
cum_rally	1	0.18	6616	4881.11	0.6747
ahead	1	5.39	6615	4875.72	0.0203
hit_area	13	157.02	6602	4718.70	0.0000
opponent_location_area	12	12.13	6590	4706.57	0.4354
Area 3	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
handedness	1	0.63	6617	3912.44	0.4276
cum_rally	1	0.00	6616	3912.44	0.9622
ahead	1	7.05	6615	3905.38	0.0079
hit_area	13	212.22	6602	3693.17	0.0000
opponent_location_area	12	52.94	6590	3640.23	0.0000
...					
Area 8	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
handedness	1	5.11	6617	4906.65	0.0238
cum_rally	1	10.11	6616	4896.54	0.0015
ahead	1	0.86	6615	4895.67	0.3526
hit_area	13	267.73	6602	4627.94	0.0000
opponent_location_area	12	19.44	6590	4608.50	0.0784
Area 9	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
handedness	1	3.22	6617	3409.39	0.0730
cum_rally	1	1.60	6616	3407.79	0.2057
ahead	1	7.73	6615	3400.07	0.0054
hit_area	13	255.29	6602	3144.78	0.0000
opponent_location_area	12	76.32	6590	3068.46	0.0000

Table 6.4: Test of significance factors influencing the probability of shots (other than serves) landing in area 1 for Viktor Axelsen. From these tables, we can see that hit area and opponent location are significant factors influencing the landing area of next shot in most of the models.

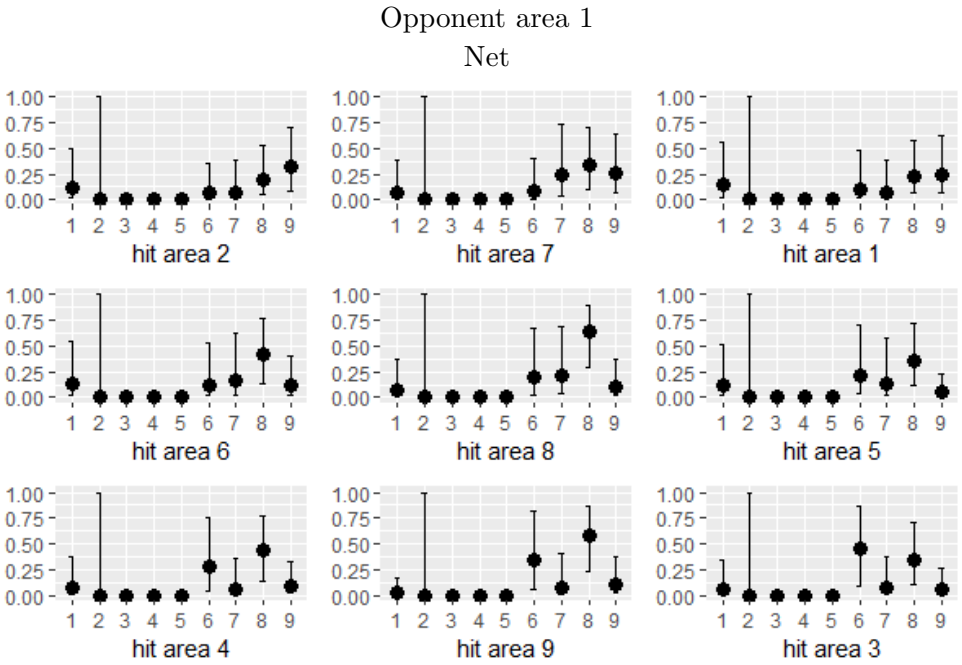

Figure 6.10: The predicted probabilities of next shot landing in each area given "opponent location" is "1" and "hit area" for Viktor Axelsen. For each of the plot, the x axis represents the "landing area" and y axis represents the probabilities.

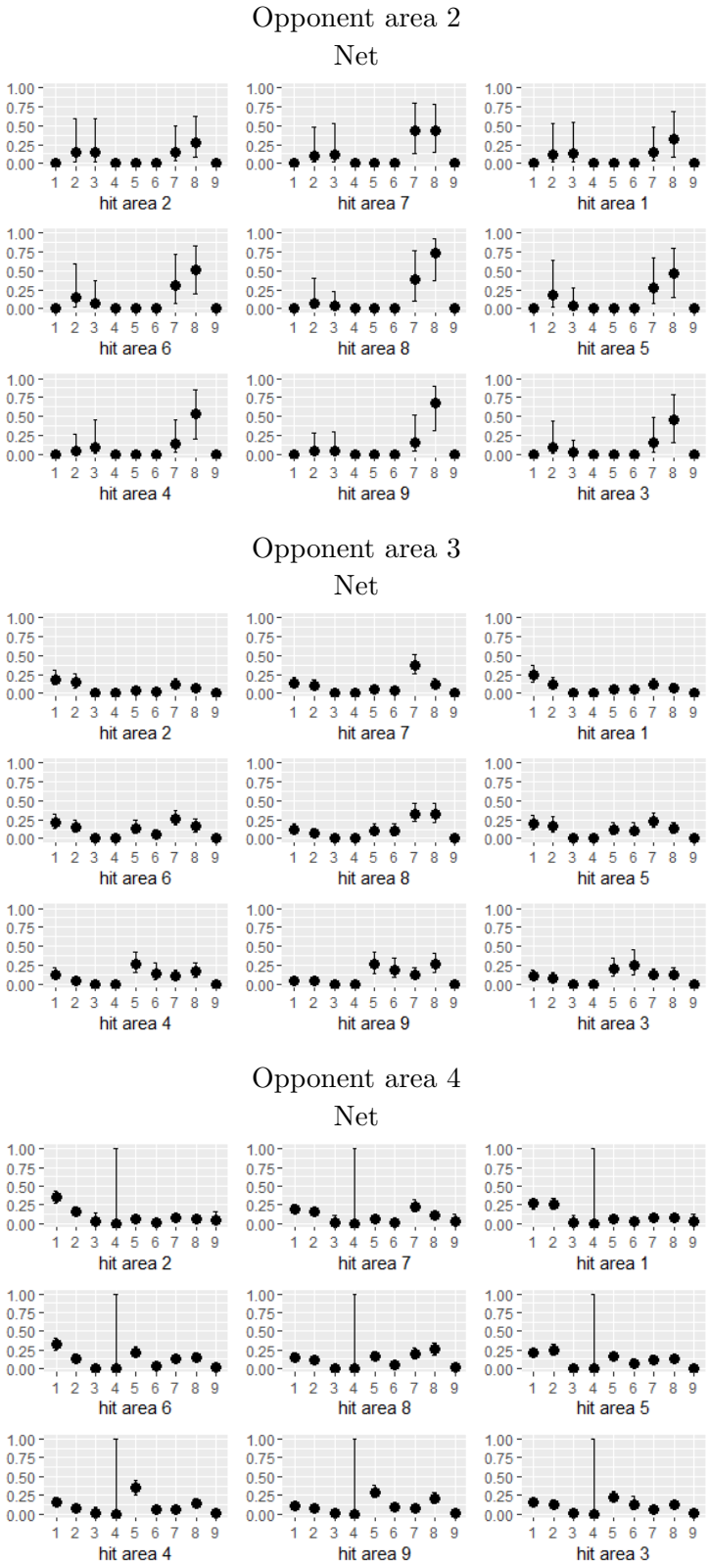

Figure 6.11: The predicted probabilities of next shot landing in each area given "opponent location" are "2", "3", and "4", and "hit area" for Viktor Axelsen. For each of the plot, the x axis represents the "landing area" and y axis represents the probabilities.

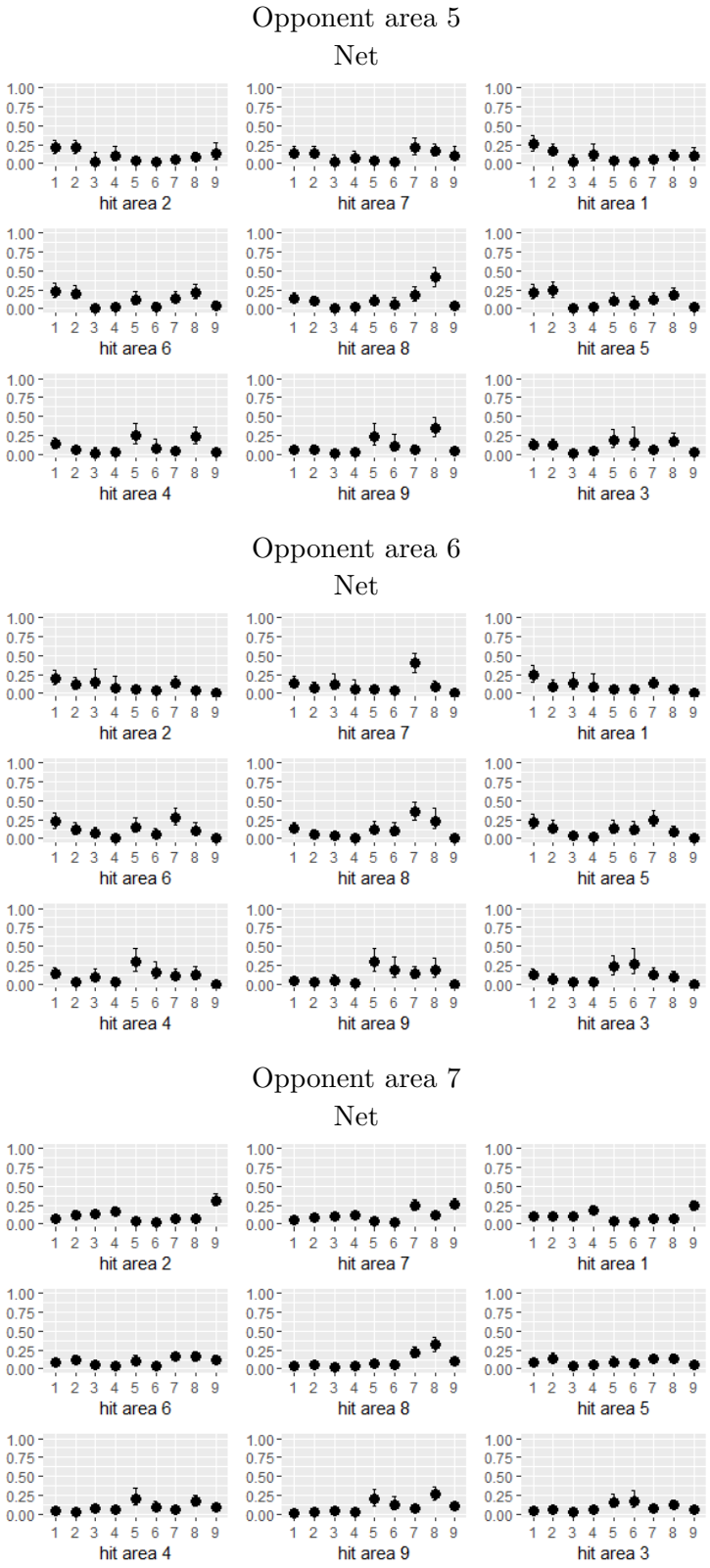

Figure 6.12: The predicted probabilities of next shot landing in each area given "opponent location" are "5", "6", and "7", and "hit area" for Viktor Axelsen. For each of the plot, the x axis represents the "landing area" and y axis represents the probabilities.

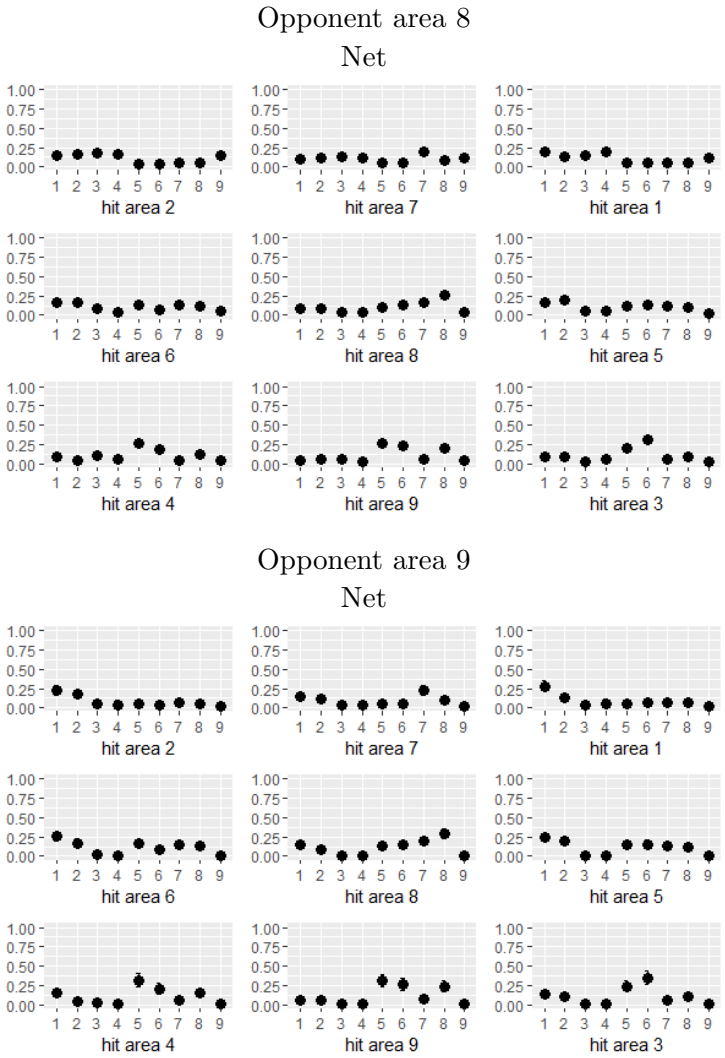

Figure 6.13: The predicted probabilities of next shot landing in each area given "opponent location" are "8" and "9" and "hit area" for Viktor Axelsen. For each of the plot, the x axis represents the "landing area" and y axis represents the probabilities.

CHAPTER 7. Discussion

In this project, we conducted statistical analysis of the patterns of shots and key factors of winning points based on methods, such as Kernel Density Estimate and logistic regression with mixed effect models, trying to summarize some general characteristics and generating the player-specific strategies. Additionally, we developed a reproducible workflow using the `targets` package in R, creating an efficient data pipeline that takes in standardized datasets of badminton matches and generates necessary output for further statistical analysis. While applying these statistical methods, we observed few potential issues on the data and the model assumptions.

The "landing location" (x, y) in the original data represents where the shots were returned by the opponent players most of the time, instead of the actually landing location. As shown in Figure 6.1, a great proportion of the serves were estimated to be "out" (not reaching the front service line), caused by the "landing location" problem, which are potentially "in". Since we only characterize serves as "short" and "long", this problem does not influence the estimation of the probabilities of serving "short". Another potential issue is that the "landing locations" of 891 among the 3087 serves are missing due to the change in camera view. But we believe this problem of missing data did not cause significant influences on further analysis as the data is missing completely at random. Besides, the locations of shots and players in the original data is marked with the pixel locations in the recordings, which led to inaccuracy as the information on the height of these location is ignored. And this issue have some potential influences on our analysis as the number of shots recorded as out (passing the back boundary line) would be higher than the actual number. From Figure 6.3, there are obvious disorders in distribution of landing location for opponent area 1 and 2, which is caused by errors of labeling of opponent areas in original data. For coding convenience, we solved the homography matrix H in equation (4.1) based on assumption of $h_{33} = 1$, which satisfies equation (4.1) but might not be the solution that minimize

the transformation error (noise). And if we have more than four matching pairs of source and destination points, H must be solved using constrained least squares problem,

$$\min_{h} ||Ah||^2, \text{ such that } ||h||^2 = 1, \tag{7.1}$$

for A is the matrix composed by equations (4.2) of all matching pairs of points. And the H (h) can be solved by finding the eigenvector corresponding to the smallest eigenvalue of matrix $A^T A$ after applied Lagrangian method on least squares problem (7.1). And the homography transformation of the shot locations tends to treat all shots as they are on the same horizontal plane, which ignored the fact that each shot is hit on different height. Therefore, to reduce the influences from these issues, it would be better to record the height information of each shots in the original data and the potential landing location could be estimated with certain fluid dynamic models modeling the trajectory of shots in physics.

In the part of analyzing the probabilities of shots landing in different areas, we applied logistic regression model estimating probability of shot landing in each area compared to all other areas, which lead to a potential problem that the total probability does not sum up to 1, as each probability is estimated separately. Selecting proper multinomial regression models could be a better choice for this problem. The logistic regression model is the main statistical method we applied in this project, where we assume shots to be independent from each other in most of the situations. However, attacks and advantages could be constructed through sequential shots by player intentionally. Thus, more complicated stochastic process models and time series models could be applied to this data for further analysis on the dependence among shot. For the estimation through Bayesian simulation, we only applied default priors of `stan_glmer` function, which give reasonable results, as we do not have further prior knowledge on the distribution of coefficients. But, further examination on prior sensitivity should be applied.

CHAPTER 8. Reference

8.1 References

Baddeley, A., Rubak, E., and Turner, R. (2016). *Spatial point patterns: methodology and applications with R*, volume 1. CRC press Boca Raton.

Betancourt, M. (2017). A conceptual introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434*.

Chen, T., Fan, Q., Liu, K., and Le, L. (2021). Identifying key factors in momentum in basketball games. *Journal of Applied Statistics*, 48.

Dubrofsky, E. (2009). Homography estimation. *Diplomová práce. Vancouver: Univerzita Britské Kolumbie*, 5.

Gabrio, A. (2021). Bayesian hierarchical models for the prediction of volleyball results. *Journal of Applied Statistics*, 48.

Galeanoa, J., Gomez, M.-A., Rivasc, F., and M. Buldu d, J. (2022). Using Markov chains to identify players performance in badminton. *Chaos, Solitons Fractals*, 165.

Klaassen, F. J. G. M. and Magnus, J. R. (2001). Are points in tennis independent and identically distributed? evidence from a dynamic binary panel data model. *Journal of the American Statistical Association*, 96.

Magnus, J. R. and Klaassen, F. J. G. M. (1999). The final set in a tennis match: Four years at Wimbledon. *Journal of Applied Statistics*, 26.

Percy, D. F. (2009). A mathematical analysis of badminton scoring systems. *Journal of the Operational Research Society*, 60.

Wang, W.-Y., Huang, Y.-C., Ik, T.-U., and Peng, W.-C. (2023). Shuttleset: A Human-Annotated Stroke-Level Singles Dataset for Badminton Tactical Analysis. *Knowledge Discovery and Data Mining*.