Applying Multi-Resolution Stochastic Modeling to Individual Tennis Points

by

CALVIN MICHAEL FLOYD

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Applied and Computational Mathematics
School of Mathematical Sciences, College of Science

Rochester Institute of Technology
Rochester, NY

May 30, 2017

Committee Approval:

Dr. Matthew Hoffman Date

School of Mathematical Sciences

Director of Graduate Programs / Thesis Co-Advisor

Dr. Ernest Fokoué Date

School of Mathematical Sciences

Thesis Co-Advisor

Dr. Nathan Cahill Date

School of Mathematical Sciences

Committee Member

ABSTRACT

Individual tennis points evolve over time and space, as each of the two opposing players are constantly reacting and positioning themselves in response to strikes of the ball. However, these reactions are diminished into simple tally statistics such as the amount of winners or unforced errors a player has. In this thesis, a new way is proposed to evaluate how an individual tennis point is evolving, by measuring how much a player can expect each shot to contribute to a won point, given who struck the shot and where both players are located. This measurement, named "Expected Shot Win Rate" (ESWR), derives from stochastically modeling each shot of individual tennis points. The modeling will take place on multiple resolutions, differentiating between the continuous player movement and discrete events such as strikes occurring and duration of shots ending. Multi-resolution stochastic modeling allows for the incorporation of information-rich spatiotemporal player-tracking data, while allowing for computational tractability on large amounts of data. In addition to estimating ESWR, this methodology will be able to identify the strengths and weaknesses of specific players, which will have the ability to guide a player's in-match strategy.

CONTENTS

1 INTRODUCTION

1.1 INSPIRATION

Before moving into the discussion of this thesis, it is important to understand where the inspiration for its research originated. In the sport of basketball there are many statistics which are used to measure its players and its teams. These include points, rebounds, assists and field goal percentage. Traditionally, there are not many ways to measure how well players interact with each other in regards to space and time. A single game of basketball can be broken down into individual possessions for each team, and each of these possessions can be broken down further into increments of time. Using these discretizations, multi-resolution stochastic models have been applied to individual basketball possessions. In the paper *A Multiresolution Stochastic Process Model for Predicting Basketball Possession Outcomes* by Daniel Cervone, Alex D'Amour, Luke Bornn and Kirk Goldsberry, Cervone et al. incorporated NBA player-tracking data to create an Expected Possession Value (EPV) estimator, defined as "the expected number of points the offense will score on a particular possession conditional on that possession's evolution up to time t" [5]. Beyond calculating the expected points of a given possession, the EPV estimator and its methodology were used to evaluate individual players' decision-making and effectiveness. [5] is the main inspiration for this work, and it will be referred to often.

Tennis is similar to basketball in the sense that the most common measures of a tennis match are simple tally statistics such as how many aces, unforced errors or winners a player has, or what percentage of first serves were won. Not many of these common statistics measure how a player is doing in regards to positioning or game strategy. This thesis will attempt to tackle that problem. There are many aspects of the EPV approach that can be applied to the game of tennis. This includes the facts that tennis points can also be broken down into discrete states and, like basketball, there is continuous player movement occurring at all times. It is the intent of this thesis to analyze how the chances of winning for a specific player evolve throughout the course of a point, incorporating this continuous player movement.

1.2 THE APPROACH OF CERVONE ET AL.

For the last three seasons, the National Basketball Association (NBA) has used a player-tracking system during every one of their games. This player-tracking system gives the exact (x, y) coordinates of each player on the court and the (x, y, z) coordinates of the ball at every 25^{th} of a second. Along with the coordinates, the data includes the actions of each player at each moment, including whether a player passes, shoots or turns the ball over. The United States Tennis Association (USTA) also collects similar player-tracking data

during its US Open tennis tournaments. Along with the coordinates of each player[1] and the ball, the USTA also records variables such as amount of spin on shots, speed of shots, game score, whether a ball coordinate refers to the ball being hit, bouncing or crossing the net, among others. Stochastic models have been applied to basketball to incorporate the evolution of possessions up to time t, in order to predict what will happen next, and the same approach can be applied to tennis. Markov chain models work well for a potential model of individual tennis points as it is necessary to incorporate the information of what happened at the time of the most previous shot, in order to predict what event will happen next and to calculate how much each shot helps each player to a winning point. In [5], the EPV estimator was formulated as

$$\nu_t = \sum_{c \in C} \mathbb{E}[X | C_{\delta_t} = c] \mathbb{P}(C_{\delta_t} = c | \mathcal{F}_t^{(Z)}). \tag{1}$$

Equation (1) expresses EPV as "the expectation given by a homogeneous Markov chain on $\mathcal{C}$ with a random starting point C_{δ_t}, where only the starting point depends on the full-resolution information $\mathcal{F}_t^{(Z)}$" [5]. The most important variables of ν_t include X, which is the number of points scored for the possession, and C_{δ_t} which is the outcome state of transitions such as passes, turnovers, or shots. The intention of this thesis is to produce an estimator similar to EPV, but mold it in a way so it conforms to the sport of tennis.

One of the key ideas this thesis will be borrowing from [5] is the way Cervone et al. coarsened their states, which made sense to the flow of a basketball possession. Cervone et al. separated basketball possessions into three distinct states: The possession state consisting of a player handling the ball, the transition state consisting of a pass, shot or turnover, and the end state consisting of the outcome of either 0, 2 or 3 points. Visually, the basketball possessions and its data were coarsened into:

[1] This thesis will be analyzing strictly singles matches, and no doubles matches. Therefore, there will be two players on the court at all times.

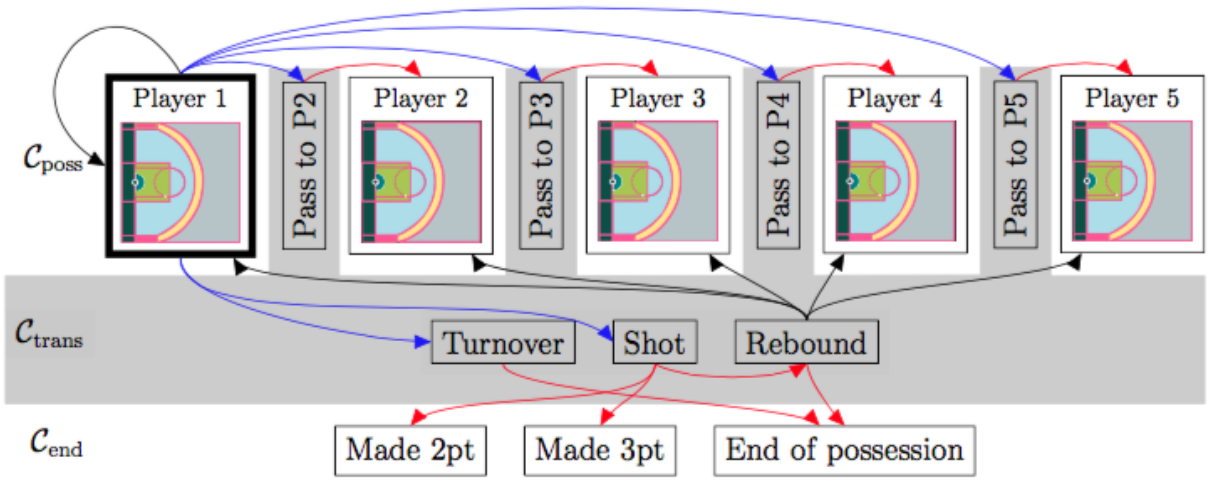

FIGURE 1: Visualization from [5] showing how Cervone et al. coarsened their basketball possessions.

Using the states described in Figure 1, Cervone et al. were able to create a general Markov transition probability matrix to model basketball possessions. This general Markov transition probability matrix was then incorporated into the formula estimating the EPV.

1.3 GOALS OF WORK

The main objective of this thesis is to construct a model which incorporates the locations of player i and player j on the court at each time t, in order to evaluate how much every shot contributes to each player winning or losing the point. There are many things to consider with this model: What shot is being hit by the player, where the players are at each hit, where the ball is likely to go next, what are each player's tendencies, etc. This methodology will potentially be able to identify strengths and weaknesses of specific players in regards to their striking ability and their return ability. The results from this thesis could help tennis players strategize for matches, based on personal strengths and opponents' weaknesses. The model could also potentially be able to identify where a point "swung" and gave the advantage to the winning player.

Expected Shot Win Rate (ESWR), this thesis's version of the EPV estimator, and its methodology ideally will be able to describe how a tennis point is evolving. Cervone et al. created a "stock-ticker" visualization of the EPV metric and how it changed over the course of a possession. They analyzed a specific possession between the Miami Heat, on offense, and the Brooklyn Nets, on defense. Figure 2 is their EPV visualization of that possession:

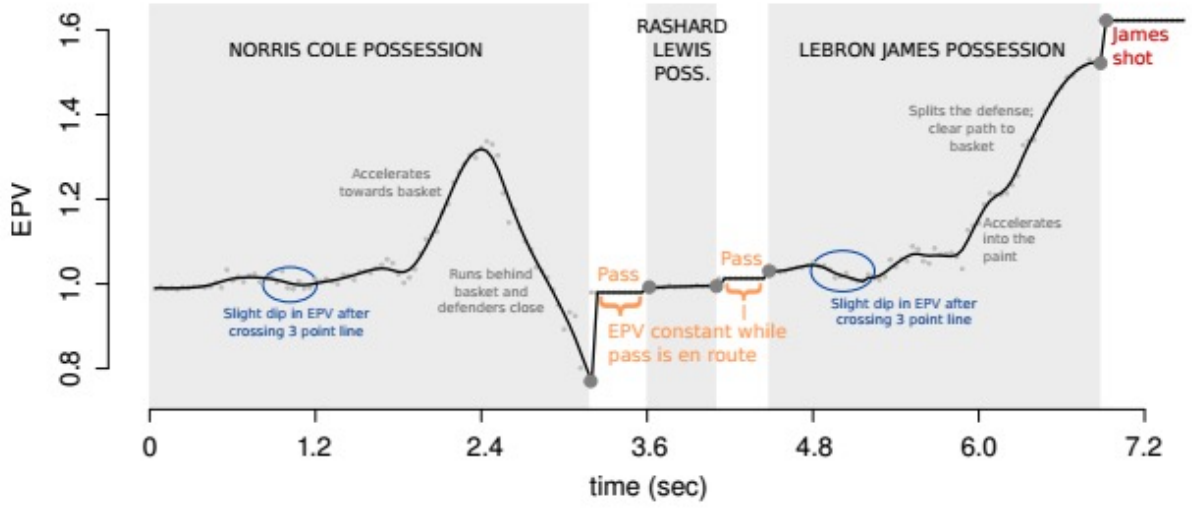

FIGURE 2: The stock-ticker visualization Cervone et al. used to demonstrate how their EPV metric evolved over the course of a basketball possession.

Now, it should be noted that a basketball possession is much more continuous and free-flowing than a tennis point. During a basketball possession, the ball-handler possesses the ball for long periods of time, and he can shoot or pass whenever he wants. Because of this, the stock-ticker visualization in Figure 2 appears very continuous (with some obvious exceptions, like when the ball is passed to another player.) During a tennis point, the tennis ball is "possessed" for extremely small periods of time. The only time a player can dictate where the ball goes to next, and therefore impose his[2] will on the point, is the fraction of a second when he strikes the ball. Compare this to a basketball possession, where the ball-handlers can impose their will on the possession for the multiple seconds they handle the ball. As a result, the "stock-ticker" visualization for the ESWR estimator will appear much choppier than that in Figure 2. This visualization can be found in Section 5.3.

1.4 DATASET AND RELATED WORK

The data used in this thesis will all come from the data collected during the USTA's 2015 US Open tournament data. The technology which collects and outputs this data is mostly used at the major tennis tournaments to assist chair-umpires with calling a shot bounce in-bounds or out-of-bounds. Though the applications of this data is plentiful, there has not been much analysis of this specific type of tennis player-tracking data in regards to player performance or game prediction. This is partly because the data has not

[2]Women's tennis is an extremely popular sport, and arguably as popular, if not more popular, than men's tennis. But for this thesis, only data from men's singles matches from the 2015 US Open was analyzed, so we will assume all players being discussed are male.

been made accessible to the public, but also because there is not as much motivation for new and more informative analytics in tennis, compared to other sports. In baseball, for example, many sports fans are familiar with the so-called "Moneyball" analytics revolution, based on the book *Moneyball*, by Michael Lewis [12]. At a certain level this refers to creating game strategies which use players in the most effective ways, in order to produce as many wins as possible, but the Moneyball analytics revolution mostly refers to professional baseball front offices using analytics to identify undervalued and under-priced players in order make the most out of their money. In professional tennis, there are no front offices, there are no teams which pay players' salaries, and there are no major parties who necessarily care if a certain player wins or not. Thus, the lack of analytical progress in tennis can be at least partially explained by organizations in tennis not necessarily having financial motivation to use their information-rich player-tracking data to make informed decisions or to guide player strategy.

There will be two distinct datasets used in this thesis. One dataset is a five-match sample from the 2015 US Open, with all possible information available including bounce locations, shot speed and shot spin. Thus, any exact (x, y) coordinate, or any reference to speed or spin in this thesis, will be from the five-match sample. The other dataset will have information from all available data from the 2015 US Open. The USTA was gracious enough to allow the ESWR estimator and its related methodology to be run on the entirety of its 2015 US Open data, in order to give the ESWR's results a dependable sample size. However, this dataset will not contain any specific information of the matches, such as coordinates and shot speed. This dataset will mostly be used in the results section, Section 5. For any calculation in this thesis, it will be specified which dataset was used. Also, the names of the players we are analyzing will not be used in this thesis, in order to protect their privacy. Thus, there will be a focus on the mathematics and the potential of the ESWR estimator and its methodology, rather than the performance of specific, well-known tennis players.

The five-game sample alone is still a significant amount of tennis player-tracking data, compared to past works which used this type of data, since the data is difficult to obtain. Prior to this thesis's use of the tennis player-tracking data, the main tennis player-tracking dataset available for public use was for the entire 2012 Australian Open, although there has been work done on the 2013 and 2014 tournaments, as well. In [18], Mora and Knottenbelt, who used the 2012 Australian Open data, discussing the system used to produce the player-tracking data, stated: "*...the complexity and cost of this system make it only available to the main events at major tournaments. Therefore the data obtained with such systems is rich but limited in the number of matches and of difficult access.*" This sentiment is echoed in other papers, like Loukianov [13] who analyzes by-hand a 2009 Australian Open first set between Roger Federer and Rafael Nadal. He attempts to Markov model a tennis point, similar to this thesis, but, ultimately, Loukianov simply comments on the "potential of Hawkeye TM technology," as the data is difficult to obtain. That being said, there has been interesting work done on the aforementioned 2012 Australian Open data. A Bayesian Network framework

was used to "model player behavior using ball and player tracking information" in [21]. The same Australian Open data was used to evaluate fatigue and how it affected tennis players depending on variables such as how much distance they traveled during matches, how long each match lasted and how many days they had been playing for [19]. In [20], Wei et al. were able to not only analyze the 2012 Australian Open data, but also tennis player-tracking data from the 2013 and 2014 Australian Open tournaments. They used their three tournaments worth of data to predict serve styles, based on the server's tendencies, the opponent's tendencies and the match context at the time of the serve.

Tennis is not the only professional sport to start incorporating spatiotemporal player-tracking data during their events. Of course, basketball and the NBA has its own player-tracking data, on which there has been a good amount of meaningful analysis done already. The potential of this spatiotemporal data was introduced by Kirk Goldsberry in 2012 with [8], showing significant differences in players' shooting abilities. The importance of players residing in certain regions on a basketball court was analyzed in [4]. New defensive metrics and evaluations based on spatial analysis were created in [6] and [9]. The strategy of teams to get players "open" to shoot the basketball was analyzed with the help of the spatiotemporal data [14]. Also, by using the player-tracking data, Miller et al. were able to develop a machine learning approach to represent and analyze shot selection [17]. Soccer is another sport which has embraced spatiotemporal player-tracking data, from an analysis perspective. Using soccer player-tracking data, classifications of passes in soccer matches have been made [10], and so has improved shot prediction [15]. There has also been much analysis made on team strategy [1, 16].

There have also been uses of stochastic models in other sports. Cervone et al.'s [5] has already been discussed and its applications of stochastic models in basketball. In [2] the single evolution of a play leading up to a goal by Argentina in the 2006 World Cup was analyzed, and the flow of the play was modeled by a stochastic process. In a different study, football drives were analyzed with absorbing Markov chains in order to predict their end states [7]. Markov chains were applied in baseball in order to make logical decisions in regard to pitching, defense and base-running [3].

2 REGION SYSTEM AND LOCATION CATEGORIZATIONS

Before deriving the ESWR estimator, it is necessary to define a few terms that have been and will be used often in this thesis. *Shot* will be a general term describing the process of making contact with the tennis ball and attempting to land it in-bounds on the opposite side of the court. For every tennis point, one player will be given a *server* label, the player who begins the point with a serve, and the other player will be labeled as the *receiver*. The server and receiver labels will not change between players during a point. For every shot during a given point, one player will be given a *striker* label and the other will be given a *returner* label. The striker will be the player who strikes the tennis ball, and the returner will be the player who is attempting to

return this strike. Unlike the server and receiver labels, the striker and returner labels *will* change between players during the point, given that the point constitutes of more than one shot. To be clear, the *strike* refers to the very beginning of a shot, when a player makes contact with the tennis ball using his racket.

2.1 REGION SYSTEM

In [5], if a player possessed the basketball at time t, they were categorized with their player ID, whether they were defended by an opposing player and which region ID they resided in. Cervone et al. divided up half of the basketball court into seven distinct regions of importance: The two corner three-pointer areas, the area right around the basket, the painted area before the free-throw line, the mid-range area outside of the painted area, the area around the three-point line and the area past the three-point line. Obviously, tennis courts are much different than basketball courts. There are no curved regions like a basketball court has, and there is no central point which all tennis points revolve around, like how a basketball possession revolves around the basket. Also, unlike basketball players, tennis players can be located in the out-of-bounds region of the court. Thus, we need to come up with a much different set of regions for the tennis court. Figure 3 shows the layout of a tennis court, with x and y axes, measured in meters.

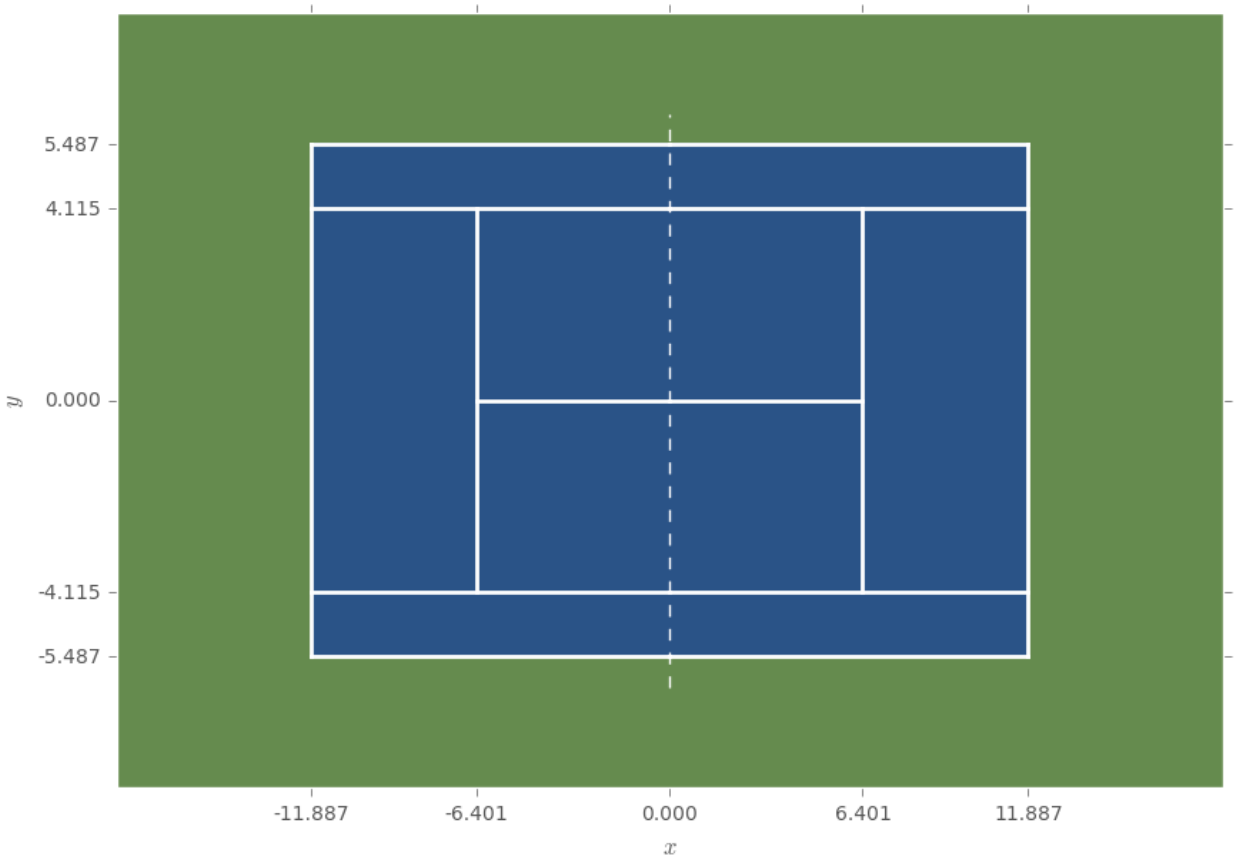

FIGURE 3: Layout of a tennis court. The blue areas indicate the in-bounds areas of the court, with the exception of the "double's allies." The green areas represent the out-of-bounds areas of the court. The dashed white line at $x = 0$ meters indicates the net.

The court's in-bounds x-coordinates from baseline to baseline range from $x = -11.9135$ meters to $x = 11.9135$ meters.[3] Only men's singles matches will be analyzed, thus the "doubles allies" will not be in play, which are the long rectangular portions of the court with their y-coordinates ranging from $y \in [4.115, 5.487]$ meters and $y \in [-5.487, -4.115]$ meters. Therefore, the in-bounds y-coordinates from sideline to sideline range from $y =$ -4.1415 meters to $y = 4.1415$ meters. Any shot bounce landing outside of these in-bounds x and y values will be deemed an out-of-bounds shot. This does not mean a player cannot go out-of-bounds to return a shot, however, so the region system must extend outside of the in-bounds section of the court. For this research we will consider the locations of both players at the time of each shot. Below in Figures 4 and 5 are visualizations of where all 2,404 ground-strokes (non-serves) in the five-match sample data were struck from, and where the returners were located at the time of those strikes. Note that we also did we did not

[3]The center of the baselines are located at $x = \pm$ 11.887 meters, but if a ball touches any part of the line, it is considered in-bounds. Therefore, we must take into account the width of the lines on the court, which are 0.053 meters thick. This will be applied to the in-bounds sidelines on the court, as well.

map the locations of strikes which were returns of serves, as those will be dealt with later on in this section.

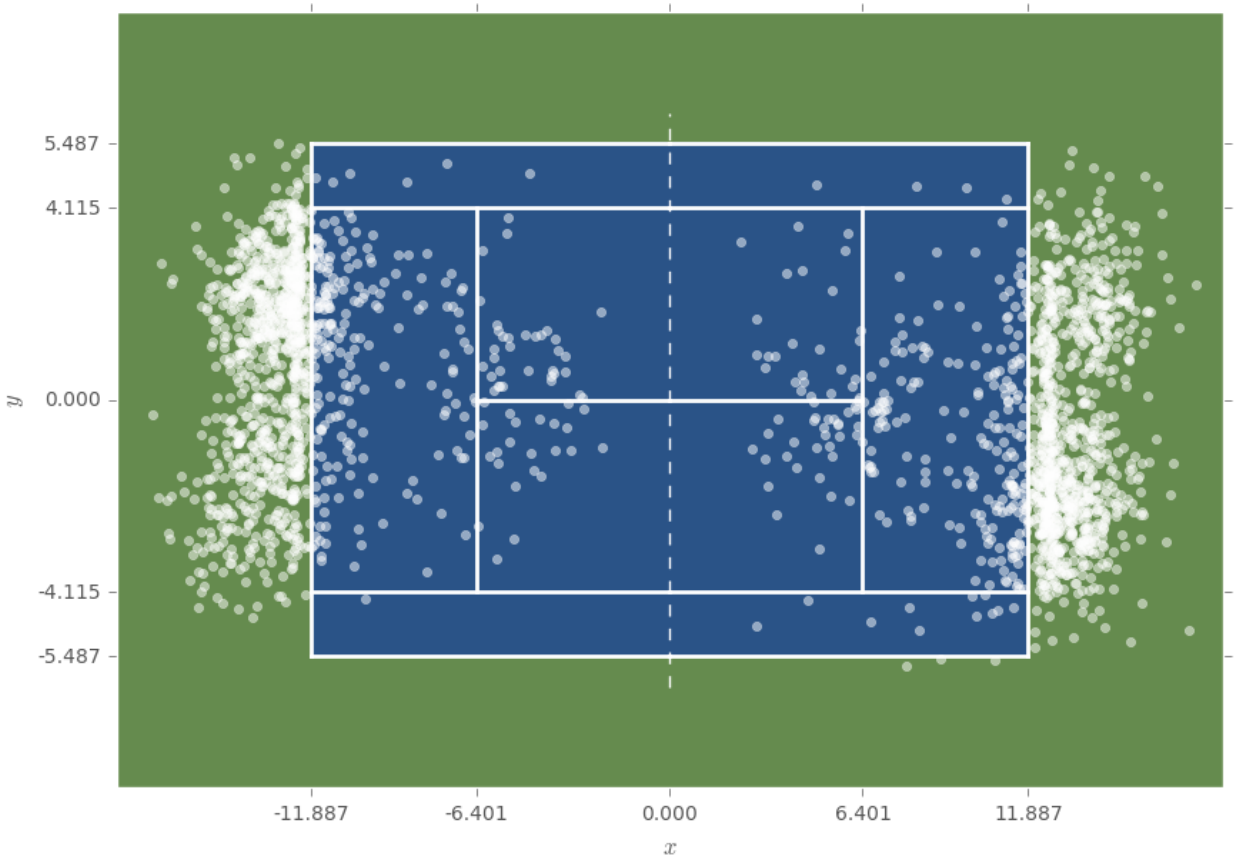

FIGURE 4: Locations of all strikes of ground-strokes in the five-match sample data, represented by small white, transparent circles. Areas with a more solid white color indicate a higher concentration of strike locations.

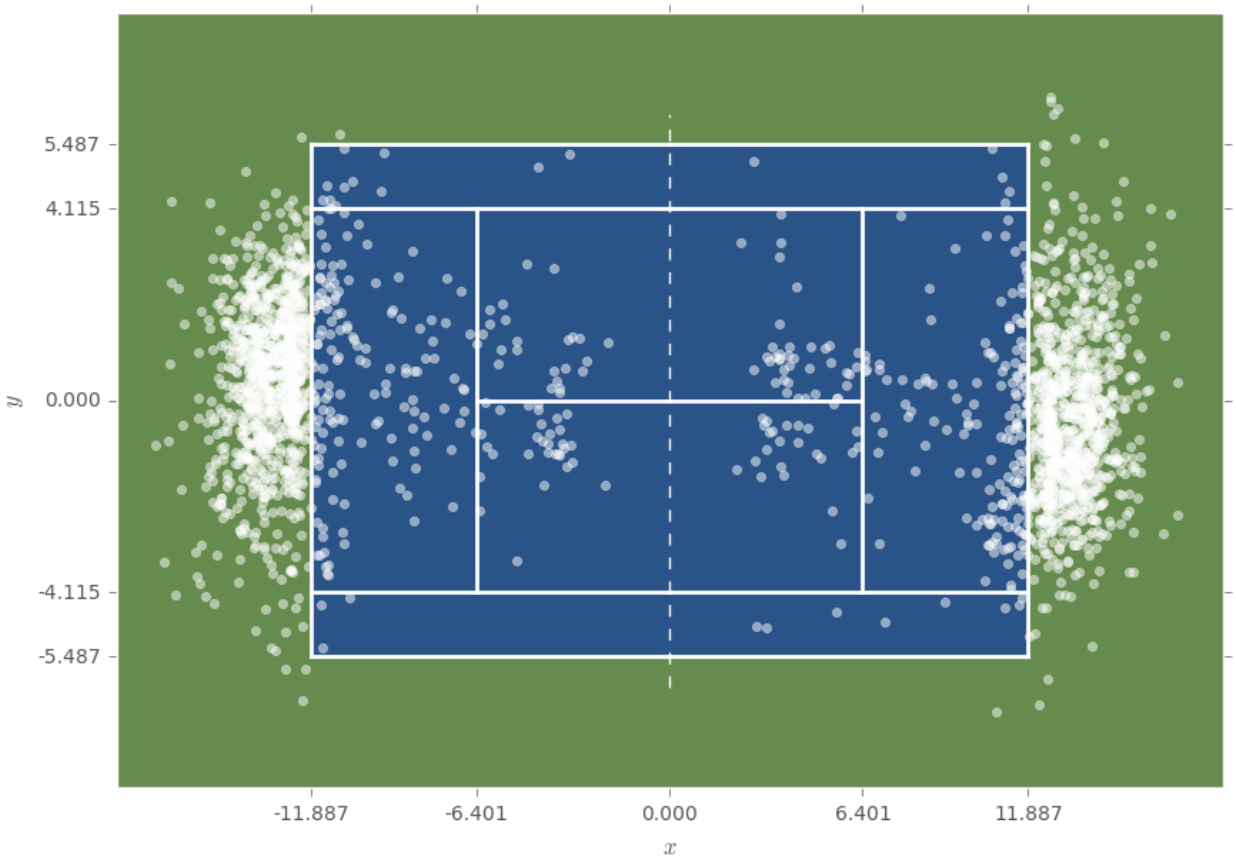

FIGURE 5: Locations of returners at the time of strikes of ground-strokes in the five-match sample data, represented by small white, transparent circles. Areas with a more solid white color indicate a higher concentration of returner locations.

Clearly, there are a high number of ground-strokes struck from the baseline area and, at the same time, a high number of returners who are setting up near the baseline area. This must be accounted for when creating the region system.

In [5], Cervone et al. used many distinctly shaped regions because of the three-point line and the basket. In tennis, there are not the same motivations, as there are no areas on the court which are more valuable to strike the ball from than others.[4]

[4]This is in the sense that no area on the court is made more valuable due to the rules of the game, and that there is no area where it is easier to score points, such as right next to the basket in basketball.

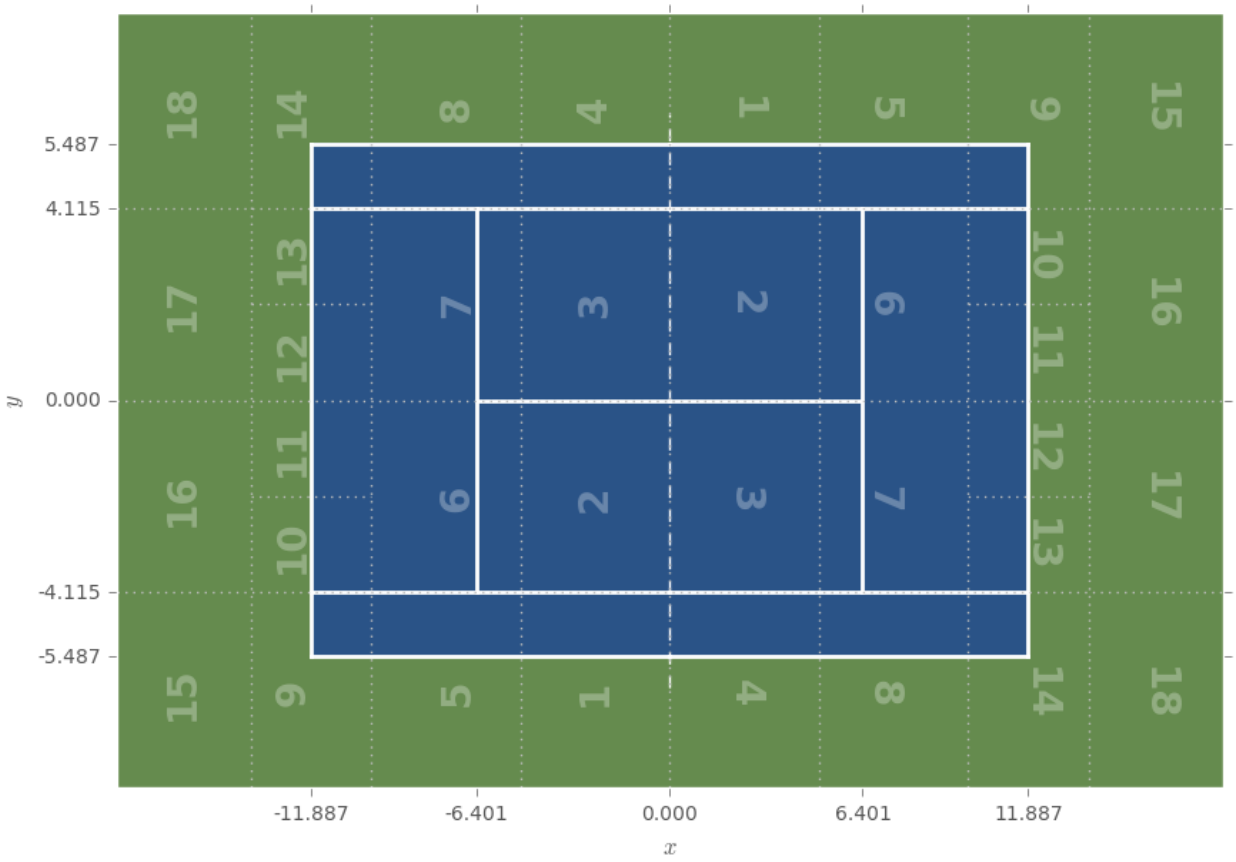

FIGURE 6: The region system which will be used for analysis, with each region given a number 1-18.

In the region system depicted in Figure 6, the area around the baseline at $x = \pm 11.887$ meters is accounted for with regions 10 through 13 encompassing its in-bounds area. As was evident in Figures 4 and 5, many shots are struck from the baseline area, and many players set up to return strikes from baseline area. Thus, it is important to have more regions along the baseline, than it is, say, near the net, where there is not as much player traffic. Regions 9 and 14 encompass the areas of the baseline which are out-of-bounds. For regions 9 through 14, the x-coordinates encompassed are two meters on either side of the middle of the baseline, such that $x \in [9.887, 13.887)$ meters and $x \in (-13.887, -9.887]$ meters. This is reasonable, as regions 9 through 14 are to represent the baseline area, and creating these regions too large would include too much court area not necessarily near the baseline. Figures 7 and 8 are the strike locations of ground-strokes and the locations of the returners five-match sample data superimposed on the region system, to visually validate its construction:

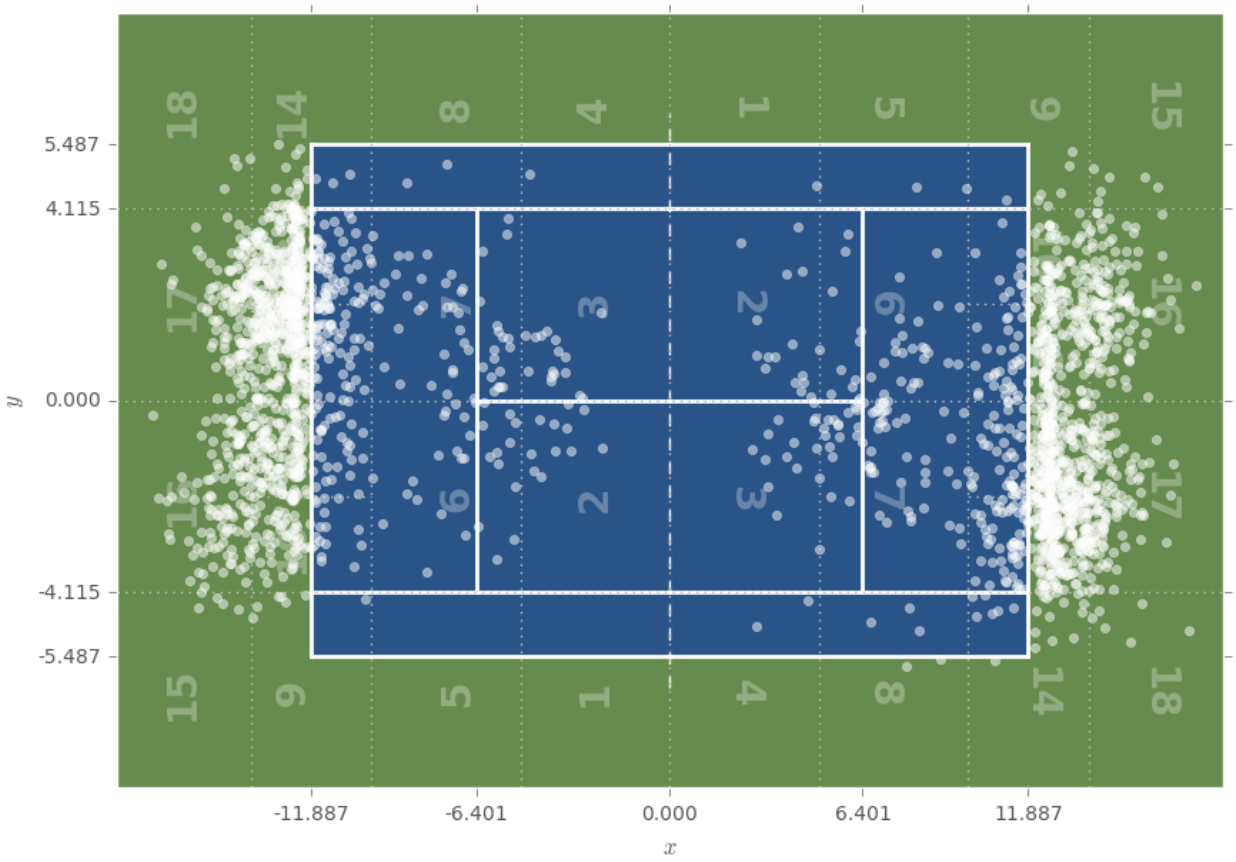

FIGURE 7: The region system and the locations of all strikes of ground-strokes from the five-match sample data, represented by small white, transparent circles.

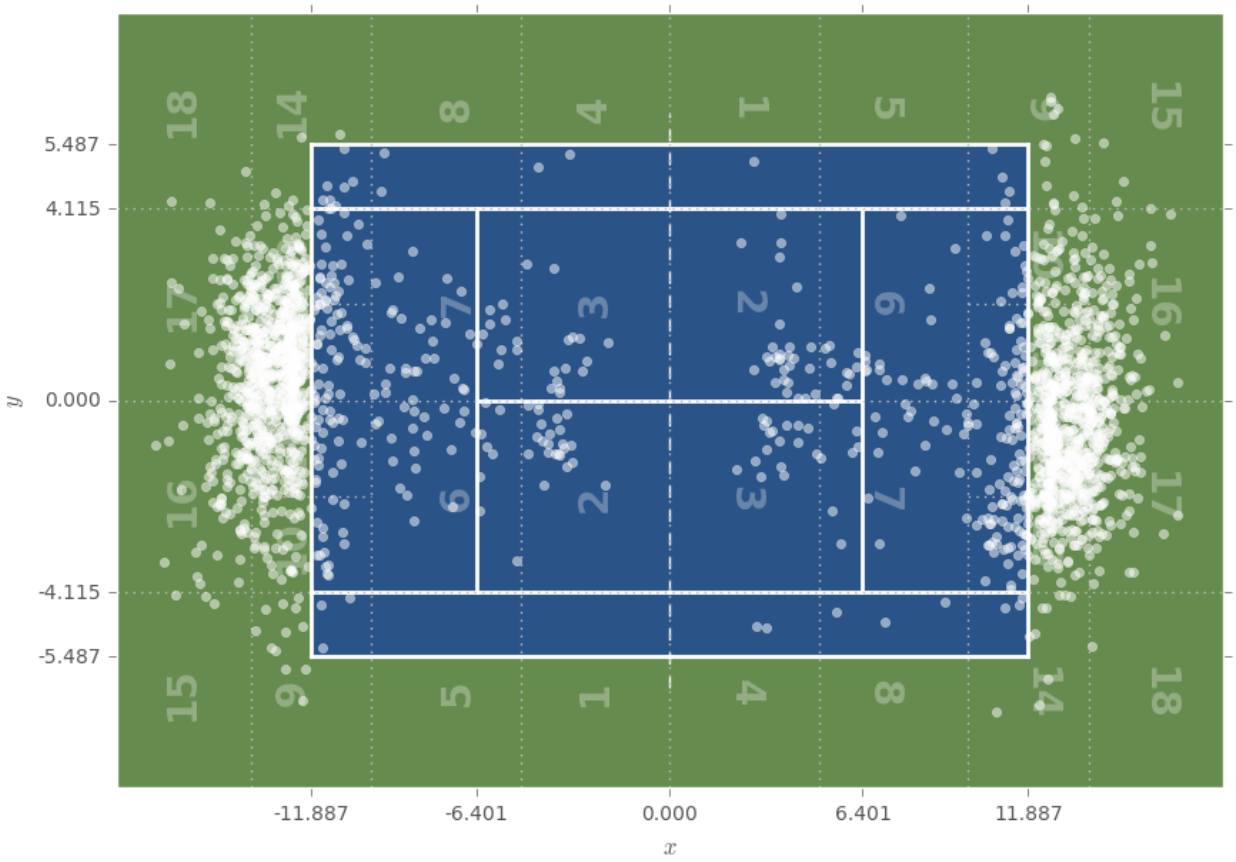

FIGURE 8: The region system and the locations of returners at the time of strikes of ground-strokes from the five-match sample data, represented by small white, transparent circles.

Regions 10 through 13 take care a majority of the shots struck near the in-bounds baseline area and a majority of the returners who set up near the in-bounds baseline area, which is their main purpose. Continuing on, regions 5 through 8 represent strikes and returns from an intermediate area of the court that is not quite at the baseline, but it is not close to the net, either. These regions' x-coordinates range from the end of the baseline regions to halfway to the net. That is, the x-coordinates for regions 5 through 8 range from $x \in [4.9435, 9.887)$ meters and $x \in (-9.887, -4.9435]$ meters.

Regions 1 through 4 encompass the area close to the net, ranging from $x \in (0, 4.9435)$ meters and $x \in (-4.9435, 0)$ meters. The region system also takes care of any shots that are far back from the baseline area, as regions 15 through 18 encompass any x-coordinates where $x \in [13.887, \infty)$ meters or $x \in (-\infty, -13.887]$ meters. It should be noted that the the regions on the outside with the largest y magnitudes range from the middle of the sideline at $y = -4.115$ and $y = 4.115$ meters to $y = -\infty$ and $y = \infty$, respectively. This is to ensure any strikes or returns with large y magnitudes are accounted.

2.2 PLAYER LOCATION COMBINATION AND TREATMENT OF GROUND-STROKES

To demonstrate how exactly it will be used, Figure 9 is the region system with the striker denoted with an "S" and the returner denoted with an "R".

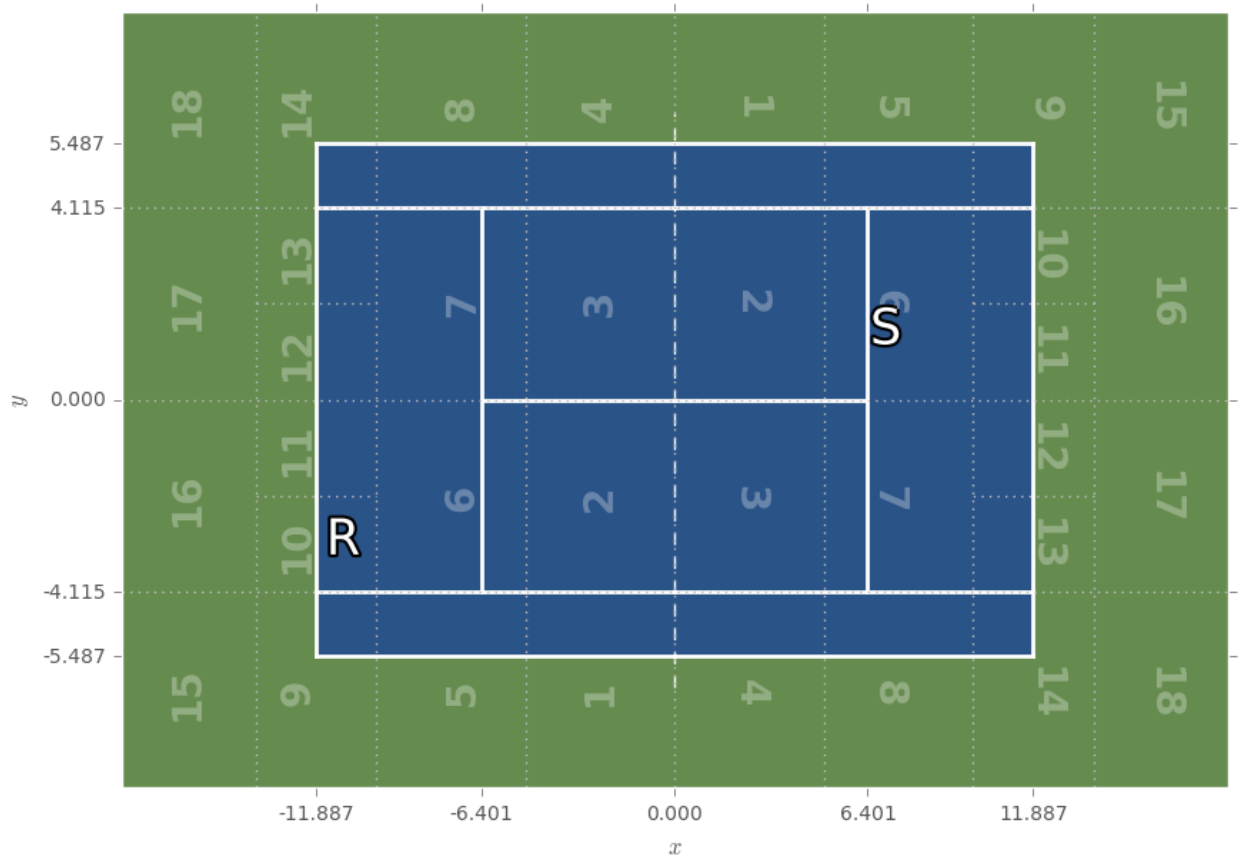

FIGURE 9: With the striker denoted with an "S" and the returner denoted with an "R", this player location combination (PLC) would be set as 6-10.

The pseudocode for translating a player's coordinates to the region he is residing in can be found in Appendix A. The driving belief behind this research is that the most important factor of where a striker chooses to hit the ball next is where the striker and the returner are located at the time of the strike. This region system will assist in grouping together similar shots, based on where the striker and returner are located. First, we will find which region the striker is striking the ball from, and then we will find which region the returner is residing in at the time of the strike. In Figure 9, if the striker struck the ball at that very instant, the shot would be categorized as 6-10 since the striker is in region 6 and the returner is in region 10.

We will name this region combination of where the striker and returner are located as the *player location combination* (PLC). These PLC's will be used to help categorize locations of players, in order to assess their striking and return abilities. Since there are 18 distinct regions on either side of the court, there are 18^2 (or 324) possible PLC's, which will all be included to help categorize the locations of each shot. All possible *location categorization*'s for a ground-stroke (GS) reside in the set

$$\mathcal{P}^{\text{GS}} = \{a - b : a = \{1, 2, \ldots, 18\}, b = \{1, 2, \ldots, 18\}\}.$$

Let Ω represent the space of all possible tennis points in full detail, with $\omega \in \Omega$ describing the full path of a particular point. At time $t \geq 0$ during a given point path ω, the most previous shot will be given a *strike state*, $\beta_t^S(\omega)$, and a *return state*, $\beta_t^R(\omega)$. Most times, $\beta_t^S(\omega)$ and $\beta_t^R(\omega)$ will be the same, as they will simply be the PLC at time t of point path ω, but there are special instances where this will not be the case, as explained in the Section 2.3.

2.3 LOCATION CATEGORIZATION OF SERVE-RELATED SHOTS

One of the goals of this thesis is to assess both a player's strike ability and return ability from each unique location categorization, in order to find weaknesses and strengths in a player's skill set. Those familiar with the game of tennis know that ground-strokes and serves are very different shots. Every point begins with a serve, and, often times, it sets the tone for the rest of the point. On a serve, the server and the returner line up in approximately the same locations every time: The server must line up behind the baseline, on the opposite half of the court to the serve box he is attempting to land his serve in, and the returner usually lines up along the baseline behind the serve box the server is aiming for. Thus, for a majority of first and second serves, the PLC will be either 11-10 or 12-13. Figures 10 and 11 show this described lack of variety in player locations during serves, plotting the locations of all 1,272 first or second serves in the five-match sample dataset, and also the locations of the returners at the time of those serves.

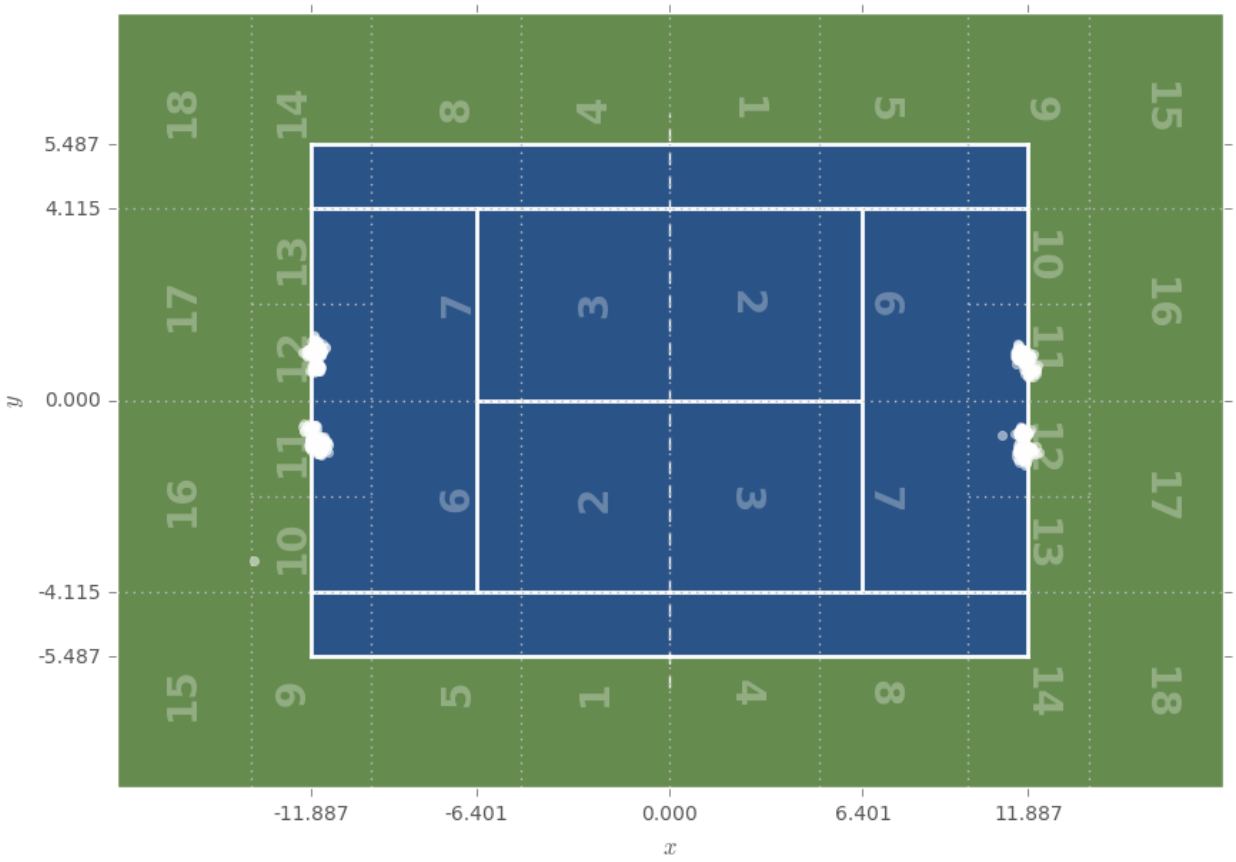

FIGURE 10: Locations of strikes of first and second serves, showing high concentrations in regions 11 and 12, near the baseline.

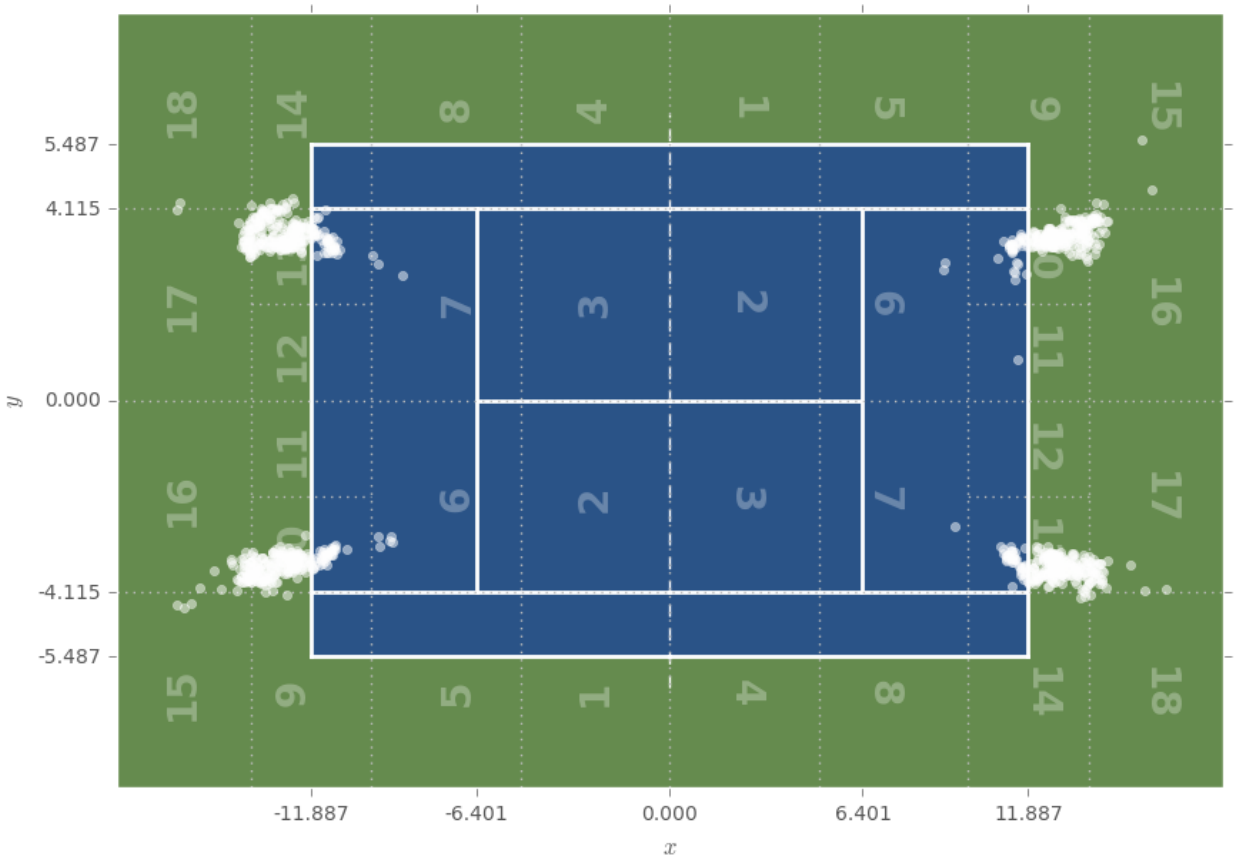

FIGURE 11: Locations of receivers at the time of the first and second serves, showing high concentrations in regions 10 and 13, right behind the baseline.

To lump serves and serves' returns in with 11-10 and 12-13 ground-strokes is unfair, as serves and serves' returns are much different than the average ground-stroke. Because of this and since serves are so important, and are often times difficult to return for a variety of reasons[5], we will treat the location categorizations of both serves' and the returns of serves' differently.

Note that all the cases described assume a shot being struck at time $t \geq 0$ of a specific point path ω. When a first serve is struck, $\beta_t^S(\omega)$ will be set to 1S (First Serve) and $\beta_t^R(\omega)$ will be set to 1SR (First Serve Return). For a second serve, $\beta_t^S(\omega)$ will be set to 2S (Second Serve) and $\beta_t^R(\omega)$ will be set to 2SR (Second Serve Return). When the return shot of a first serve is struck, the $\beta_t^S(\omega)$ will be set to 1SR, and $\beta_t^R(\omega)$ will be set to the PLC at that time. Similarly, when the return of a second serve is struck, $\beta_t^S(\omega)$ will be set to 2SR, and $\beta_t^R(\omega)$ will be set to the PLC at that time. Thus, for a strike or return which is serve-related

[5]These include the fact the server is able to take his time and aim his shot, and that serve speeds are much faster than ground-stroke speeds.

(S-R), the possible location categorization's reside in the set

$$\mathcal{P}^{\text{S-R}} = \{\text{1S, 2S, 1SR, 2SR}\}.$$

All possible location categorization's will reside in the union, $\mathcal{P}$, of $\mathcal{P}^{\text{S-R}}$ and $\mathcal{P}^{\text{GS}}$:

$$\mathcal{P} = \mathcal{P}^{\text{S-R}} \cup \mathcal{P}^{\text{GS}}.$$

3 SHOT SIGNIFICANCES

Through the perspective of a tennis player, every point either ends in a won point, given a point value of 1, or a lost point, given a point value of 0. There are shots in tennis which lead directly to a won point - such as pure winners - and shots which lead directly to a lost point - such as shots that land out-of-bounds or go directly into the net. Given the total shots from the data sample, the shots that lead directly to a point value of 0 or 1, which we call *direct shots* only make up approximately 29% of the total strikes in the entirety of the 2015 US Open data. In order to properly find strengths and weaknesses of a given player, we need to take into account more than just these direct shots and look at *indirect shots*. These indirect shots do not directly lead to the outcome of a point, but they will contribute in some way to a won or lost point. Since indirect shots will now be considered, we cannot limit ourselves to only looking at the binary values of 0 and 1, but also at the values in-between 0 and 1. In this section, we will attempt to quantify just how much these indirect shots contribute to a won point.

3.1 STRIKE SIGNIFICANCES

All strikes during a point will fall into one of the eight different categories described in Table 1. The pseudocode for sorting shots into strike significance categories can be found in Appendix B.

Strike Significance (λ_{SS})	**Description**
Pure Winner (λ_{PW})	In-bounds strike which the returner cannot return. Direct shot which leads to a won point for the striker.
Out-of-Bounds (λ_{OB})	A strike which lands out-of-bounds. Direct shot which leads to a lost point for the striker.
Net (λ_{N})	A strike which goes directly into the net, unable to make it over. Direct shot which leads to a lost point for the striker.
Set-Up of Pure Winner (λ_{SUPW})	The strike prior to a λ_{PW} by the winning player. An indirect shot.
Set-Up of Opponent's Pure Winner (λ_{SUOPW})	The strike by the losing player prior to the winning player striking a λ_{PW}. An indirect shot.
Forced Out-of-Bounds (λ_{FOB})	The strike by the winning player prior to the losing player striking a λ_{OB}. An indirect shot.
Forced Net (λ_{FN})	The strike by the winning player prior to the losing player striking a λ_{N}. An indirect shot.
Non-Impactful (λ_{NI})	A strike that does not fall into any of the other seven categories. An indirect shot.

TABLE 1: Table describing the different Strike Significances (λ_{SS}).

We want to make sure that every strike that could potentially contribute to the outcome of a point is accounted for, which is why there are seven significance categories (not including the neutral λ_{NI} category), so we can feel confident we are analyzing most of the strikes that could potentially contribute to the outcome of the point. Now, it would be naive to claim this is a perfect approach, as some strikes may be categorized as contributing in some way to a won or lost point, when in reality they probably should be categorized as λ_{NI}, but this approach to finding the most important strikes of a given point is more than reasonable. Sorting every strike available in the data will help predict how much future strikes will potentially contribute to a winning point.

Let all of the strike significances reside in the set Λ_{SS}:

$$\Lambda_{SS} = \{\lambda_{PW}, \lambda_{SUPW}, \lambda_{SUOPW}, \lambda_{OB}, \lambda_{N}, \lambda_{FOB}, \lambda_{FN}, \lambda_{NI}\}.$$

Some of these categories will contribute more to a winning point than others, and that needs to be quantified for the work in this thesis. Below in Table 2 are weights (w_{SS}) for each strike significance (λ_{SS}) corresponding to how much each strike significance theoretically contributes to a winning point.

$\lambda_{\mathbf{SS}}$	$w_{\mathbf{SS}}$
λ_{PW}	1.00
λ_{SUPW}	0.75
λ_{FOB}	0.75
λ_{FN}	0.75
λ_{NI}	0.50
λ_{SUOPW}	0.25
λ_{OB}	0.00
λ_{N}	0.00

TABLE 2: Table describing the weights (w_{SS}) for each Strike Significance (λ_{SS}).

The neutral λ_{NI} weight was set to 0.50 so that anything greater than 0.50 indicates a greater-than-average contribution to a won point, and anything below indicates a less-than-average contribution. The weights w_{PW}, w_{OB} and w_{N} weights are self-explanatory since they are direct shots. Also, it should be noted there is one specific strike state which will never be penalized with w_{OB} or w_{N} weights: When the server is striking a first serve, he cannot directly lose the point on that specific strike. If he strikes the ball into the net or outside of the serve box he is aiming for, then the first serve is considered a "fault" and he continues onto the second serve. Thus, when $\beta_t^S(\omega) = 1\text{S}$ and the strike is categorized as λ_{OB} or λ_{N}, the corresponding weight will be given a neutral 0.50 weight instead of a 0.00 weight.

The weight w_{SUPW} is 0.75 since it is a set-up to a λ_{PW}, which has a weight of 1.00, so the mean of 0.50 and 1.00 was chosen. The weight w_{SUOPW} is 0.25 since it is not fully contributing to a lost point for the striker, but it is giving the returner the ability to strike a λ_{PW} the next shot, so the mean of 0.00 and 0.50 was chosen. The w_{FOB} and w_{FN} weights were somewhat tougher to decide upon, but they both clearly deserved weights above 0.50. Ultimately, 0.75 was decided upon since if it is truly the belief that these strikes help cause the opponent to hit an λ_{OB} or λ_{N} strike, then the λ_{FOB} and λ_{FN} strikes deserve as much weight as a λ_{SUPW} strike.

Now, these categorizations of strikes may be somewhat counterintuitive. For example, if player i strikes a λ_{FOB}, and player j then strikes an λ_{OB}, does player j's strike truly deserve a 0.00 weight if player i hit a great strike which then caused player j to then hit an errant strike? There will be a lot of gray area and room for interpretation when it comes to sorting strikes into significance categories. But unless we were to watch each individual point and by-hand figure out which shots contribute more to a winning point than others, it is difficult to start sorting strikes into different categories. Ultimately, it is the intention for this ESWR methodology and its corresponding Python code to be able to run on extremely large amounts of player-tracking data. Thus, going in and assessing each individual point by-hand is not practical.

3.2 RETURN SIGNIFICANCES

The ability to strike and to place the ball wherever he wants is a very important skill for a tennis player to have. But it is fair to say a player's ability to return strikes is just as important. Half of the game of tennis is a player putting himself into position to be able to return the ball and not letting the other player dictate how the point will be played. All strikes with the ability to be returned during a point will be categorized into one of the seven Return Significances described in Table 3. The pseudocode for sorting shots into return significance categories can be found in Appendix C.

Return Significance ($\lambda_{\mathbf{RS}}$)	**Description**
Returned Pure Winner (λ_{RPW})	Returned a λ_{PW}. Direct shot which leads to a won point for the returner.
Returned Losing Strike (λ_{RLS})	Returned either an λ_{OB} or λ_{N} strike. Direct shot which leads to a lost point for the returner.
No Return (λ_{NR})	No return made on a λ_{PW} strike. Direct shot which leads to a won point for the striker.
Returned Set-Up of Pure Winner (λ_{RSUPW})	Returned a λ_{SUPW}, an indirect shot.
Returned Set-Up of Opponent's Pure Winner ($\lambda_{\mathrm{RSUOPW}}$)	Returned a λ_{SUOPW}, an indirect shot.
Returned Forced Losing Strike (λ_{RFLS})	Returned either a λ_{FOB} or λ_{FN} strike, an indirect shot.
Returned (λ_{R})	Returned a strike that does not fall into any of the other six categories, an indirect shot.

TABLE 3: Table describing the different Return Significances (λ_{RS}).

It should be noted that "strikes with the ability to be returned" includes all strikes that do not land out-of-bounds or go directly into net. Shots that do not land in-bounds prevent the returner from having the opportunity to return them. All shots like this as are labeled as "Insignificant" and they were not included in any analysis of a player's return ability. Also, since they will have the same weights, returns that have led to a λ_{OB} or λ_{N} strikes have been grouped together in the λ_{RLS} Return Significance, as have returns that have led to λ_{FOB} or λ_{FN} strikes in the λ_{RFLS} Return Significance. A λ_{NR} return is when the returner is unable to get his racket on the ball, resulting in a pure winner for the striker.

Let all of the Return Significances reside in the set Λ_{RS}. That is,

$$\Lambda_{\mathrm{RS}} = \{\lambda_{\mathrm{RPW}}, \lambda_{\mathrm{RSUPW}}, \lambda_{\mathrm{RSUOPW}}, \lambda_{\mathrm{RLS}}, \lambda_{\mathrm{RFLS}}, \lambda_{\mathrm{NR}}, \lambda_{\mathrm{R}}\}.$$

Below in Table 4 are weights (w_{RS}) for each Return Significance (λ_{RS}) corresponding to how much each return significance theoretically contributes to a winning point.

$\lambda_{\mathbf{RS}}$	$w_{\mathbf{RS}}$
λ_{RPW}	1.00
λ_{RSUPW}	0.75
λ_{RFLS}	0.75
λ_{R}	0.50
λ_{RSUOPW}	0.25
λ_{RLS}	0.00
λ_{NR}	0.00

TABLE 4: Table describing the weights (w_{RS}) for each Return Significance (λ_{RS}).

Since, the λ_{NR} is a return where a player is unable to get his racket on the ball, directly leading to a won point for the striker, $w_{\text{NR}} = 0.00$. It was the intent of this research to keep the weights of the Strike Significances and the Return Significances symmetric, therefore, the remaining Return Significance weights, w_{RS}, correspond to the same Strike Significance weights, which are explained in the previous section, Section 3.1.

4 MULTI-RESOLUTION MODELING

Recall that $\omega \in \Omega$ describes the full path of a particular point. For any point path ω, we denote by $Z(\omega)$ the optical tracking time series generated by this point so that $Z_t(\omega) \in \mathcal{Z}$, with $t > 0$, is a "snapshot" of the player-tracking data exactly t seconds from the start of the point, beginning with the server setting up for his serve. $\mathcal{Z}$ is a high-dimensional space that includes the (x, y) coordinates for both players on the court, the (x, y, z) coordinates for the tennis ball, summary information such as which players are on the court, which player is the striker and which is the returner, how each shot is classified, and event annotations that are observable in real time, such as the ball being struck or the ball traveling between players. The possible point values of a particular point path ω for player i are either 0 or 1, denoted by $X^i(\omega) \in \{0, 1\}$. Letting $T(\omega)$ denote the time at which a point following the path ω ends, the point's outcome for player i then is a deterministic function of the full resolution data at this time, $X^i(\omega) = h^i(Z_{T(\omega)}(\omega))$.

We will denote the *shot win rate* as $\Gamma^i(\omega) \in [0, 1]$, which will evaluate how much a given shot contributes to a won point for player i, for point path ω. As mentioned earlier, this thesis will focus on calculating each player's *expected shot win rate*, for every shot during a given point. This means we will have to look at shot win rate from two different perspectives: The striker's perspective, and the returner's perspective. Thus, two sub-functions will condition $\Gamma^i(\omega)$:

$$\Gamma^i(\omega) = \begin{cases} \gamma_S^i(\omega), \text{ if player } i \text{ was the striker of the shot in question} \\ \gamma_R^i(\omega), \text{ if player } i \text{ was the returner of the shot in question} \end{cases} \tag{2}$$

Taking the intuitive view of ω as a sample space of all individual tennis point paths, we define $Z_t(\omega)$ for each $t > 0$ as a random variable in $\mathcal{Z}$. Let $\mathcal{F}_t^{(Z)}$ represent the collection of all information from the player-tracking data up to time t of a particular point path ω such that $\mathcal{F}_t^{(Z)} = \{Z_s(\omega) : 0 \leq s \leq t\}$. We will define the expected shot win rate for player i, depending on all available information up to time t (which includes if the player is the striker or returner on a given shot, and where both players are located) as follows:

Definition 4.1. The *expected shot win rate*, or ESWR, for player i at time $t \geq 0$ during a given point is $\nu_t^i(\omega) = \mathbb{E}[\Gamma^i(\omega)|\mathcal{F}_t^{(Z)}]$.

Now, $\Gamma^i(\omega)$ will be a function of $X^i(\omega)$, translating the binary $X^i(\omega)$ to $\Gamma^i(\omega)$'s continuous interval of [0,1], based on $\mathcal{F}_t^{(Z)}$. More explicitly, let us now define $\Gamma^i(\omega)$ as a function of both $X^i(\omega)$ and $\mathcal{F}_t^{(Z)}$:

$$\Gamma^i(\omega) = \begin{cases} \gamma_S^i(\omega) = f_S(X^i(\omega), \mathcal{F}_t^{(Z)}), \text{ if player } i \text{ was the striker of the shot in question} \\ \gamma_R^i(\omega) = f_R(X^i(\omega), \mathcal{F}_t^{(Z)}), \text{ if player } i \text{ was the returner of the shot in question} \end{cases} \tag{3}$$

The expectation $\mathbb{E}[\Gamma^i(\omega)|\mathcal{F}_t^{(Z)}]$ is an integral over the distribution of future paths the current point can take. Recall our definition of $X^i(\omega)$ such that $X^i(\omega) = h^i(Z_{T(\omega)}(\omega))$. Thus, evaluating ESWR amounts to integrating over the joint distribution of $(T(\omega), Z_{T(\omega)})$:

$$\begin{aligned} \nu_t^i(\omega) &= \mathbb{E}[\Gamma^i(\omega)|\mathcal{F}_t^{(Z)}] \\ &= \int_\Omega f_{S/R}(X^i(\omega), \mathcal{F}_t^{(Z)}) \mathbb{P}(d\omega|\mathcal{F}_t^{(Z)}) \\ &= \int_t^\infty \int_{\mathcal{Z}} f_{S/R}(h^i(z), \mathcal{F}_t^{(Z)}) \mathbb{P}(Z_s(\omega) = z|T(\omega) = s, \mathcal{F}_t^{(Z)}) \mathbb{P}(T(\omega) = s|\mathcal{F}_t^{(Z)})\, dz\, ds \end{aligned} \tag{4}$$

Now that we have defined ESWR in (4) as a theoretical quantity, we must now develop a methodology to calculate it. In order to get the most out of the ESWR estimator, we must require it to be stochastically consistent. Using a stochastically consistent ESWR estimator guarantees that changes in the resulting ESWR over the course of a point derive from players' on-court actions rather than artifacts or inefficiencies of the data analysis. We must also require the ESWR estimator to be sensitive to the fine-grained details of the data without incurring undue variance or computational complexity. Applying a Markov chain-based estimation approach would require discretizing the data by mapping the observed spatial configuration $Z_t(\omega)$

into a simplified summary $C_t(\omega)$ potentially violates this criteria by trading potentially useful information in the player-tracking data for computational tractability.

In order to develop a methodology that meets both of the criteria above, we must understand that the information-computation trade-off results from choosing a single level of resolution at which to model the tennis point and compute all expectations. This thesis's strategy, which borrows from [5], for estimating ESWR combines models for the tennis point at two distinct levels of resolution. Namely, a fully continuous model of player movement and actions, and a Markov chain model for a highly coarsened view of the point. This multi-resolution approach leverages the computational simplicity of a discrete Markov chain model while conditioning on exact spatial locations and high-resolution data features.

4.1 COARSENING OF THE SPATIOTEMPORAL DATA

The Markov chain portion of our method requires a coarsened view of the data. For all time $0 < t \leq T(\omega)$ during a point path ω, let $C(\cdot)$ be a coarsening that maps $\mathcal{Z}$ to a finite set $\mathcal{C}$, and call $C_t(\omega) = C(Z_t(\omega))$ the "state" of the point. To make the Markovian assumption plausible, we populate the coarsened state space $\mathcal{C}$ with summaries of the full resolution data so that transitions between these states represent meaningful events in a tennis point, described below:

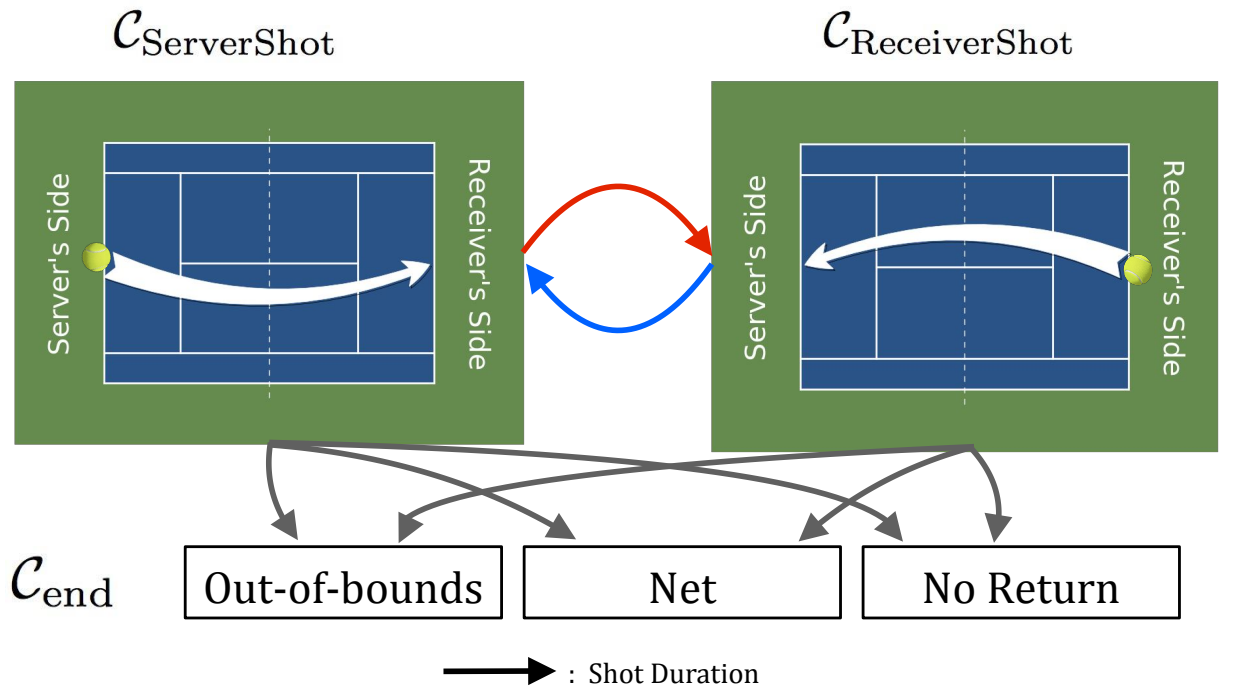

FIGURE 12: Visual representation of how the three coarsened states $\mathcal{C}_{\text{ServerShot}}$, $\mathcal{C}_{\text{ReceiverShot}}$ and $\mathcal{C}_{\text{end}}$ will interact with each other. The arrows between each state represent a "shot duration."

First, we should define what is meant by "duration" of a shot. The duration of a shot begins with the striker striking the tennis ball with his racket, and ends with either the returner striking the tennis ball with his racket or with the point ending in a state of $\mathcal{C}_{\text{end}}$. As is made clear by the coarsened state space $\mathcal{C} = \mathcal{C}_{\text{ServerShot}} \cup \mathcal{C}_{\text{ReceiverShot}} \cup \mathcal{C}_{\text{end}}$, we will be breaking down individual tennis points shot-by-shot. Also, let $\mathcal{C}_{\text{Shot}} = \mathcal{C}_{\text{ServerShot}} \cup \mathcal{C}_{\text{ReceiverShot}}$, which will be utilized later on in this section.

Now, we will define the individual states. Let us start with $\mathcal{C}_{\text{end}}$, and define what tennis events would fall in this state. Since $\mathcal{C}_{\text{end}}$ categorizes the end of a point, it is the case that $C_{T(\omega)}(\omega) \in \mathcal{C}_{\text{end}}$ and for all $t < T(\omega)$, $C_t(\omega) \notin \mathcal{C}_{\text{end}}$. A tennis point can end in three different ways: The bounce of a shot lands out-of-bounds, the shot goes directly into the net and lands on the striker's side of the court, or the ball is not returned by the returner and the ball bounces twice on the returner's side of the court[6]. Thus, $\mathcal{C}_{\text{end}} = \{\text{out-of-bounds, net, no return}\}$. Through the perspective of the striker and the returner, these end states will result in different point values. Let $X_S(c)$ and $X_R(c)$ denote the point value of the given end state $c \in \mathcal{C}_{\text{end}}$ through the perspective of the striker and returner, respectively. From the striking perspective, $X_S(\text{out-of-bounds}) = 0$, $X_S(\text{net}) = 0$ and $X_S(\text{no return}) = 1$. From the returning perspective, $X_R(\text{out-of-bounds}) = 1$, $X_R(\text{net}) = 1$ and $X_R(\text{no return}) = 0$.

Next, whenever a strike by the server is the most recent shot at time t, we assume that $C_t(\omega)$ is equal to the strike state at that time, or $C_t(\omega) = \beta_t^S(\omega)$. The possible values of $C_t(\omega)$, if a strike by the server is the most recent shot at time t, thus live in $\mathcal{C}_{\text{ServerShot}} = \{\mathcal{P}\}$. Similarly, whenever a strike by the receiver is the most recent shot at time t, we again assume that $C_t(\omega)$ is equal to the strike state at that time, or $C_t(\omega) = \beta_t^S(\omega)$. The possible values of $C_t(\omega)$, if a strike by the receiver is the most previous shot at time t, live in $\mathcal{C}_{\text{ReceiverShot}} = \{\mathcal{P}\}$.

4.2 MODEL ASSUMPTIONS

We must make a few more assumptions about the processes Z and C, which allow them to be combined into a functional ESWR estimator.

(A1) C is marginally semi-Markov.

We need to assume C is marginally semi-Markov, since our underlying process of entering different states does not have a geometrically distributed duration (i.e. the time between strikes of the ball is not the same every time). This semi-Markov assumption guarantees that the embedded sequence of disjoint states $C^{(0)}, C^{(1)}, \ldots, C^{(K)}$ is a Markov chain.

We now need to specify the relationship between coarsened and full-resolution conditioning. This requires

[6]Points which end in the striker striking a shot that bounces directly on his side of the court without going into the net will not be analyzed, as these points do not occur nearly often enough to warrant consideration.

us to define two additional time points which mark changes in the future evolution of the point:

$$\tau_t = \begin{cases} \max\{s : s < t, C_s(\omega) \notin \mathcal{C}_{\text{ServerShot}}\} + \epsilon, \text{ if } C_t(\omega) \in \mathcal{C}_{\text{ServerShot}} \\ \max\{s : s < t, C_s(\omega) \notin \mathcal{C}_{\text{ReceiverShot}}\} + \epsilon, \quad \text{if } C_t(\omega) \in \mathcal{C}_{\text{ReceiverShot}} \end{cases} \tag{5}$$

$$\delta_t = \begin{cases} \min\{s : s > \tau_t, C_s(\omega) \notin \mathcal{C}_{\text{ServerShot}}\}, \text{ if } C_t(\omega) \in \mathcal{C}_{\text{ServerShot}} \\ \min\{s : s > \tau_t, C_s(\omega) \notin \mathcal{C}_{\text{ReceiverShot}}\}, \quad \text{if } C_t(\omega) \in \mathcal{C}_{\text{ReceiverShot}} \end{cases} \tag{6}$$

where ϵ is the temporal resolution of the player-tracking data. In this case, ϵ will be equal to 1/25 second. Thus, τ_t is the first time increment after the beginning the shot duration occurring at time t, and δ_t is the endpoint of this shot duration, leaving the shot duration state into either the opposing player's shot duration state or into $\mathcal{C}_{\text{end}}$. We will assume that when a new shot duration state is passed into, this decouples the future of the point after time τ_t with its history up to time t. Let $\rho^S_{\tau_t}(\omega)$ denote the player ID of the striker of a shot occurring at τ_t, and let $\rho^R_{\tau_t}(\omega)$ denote the player ID of the returner of a shot occurring at τ_t during point path ω.

(A2) For all $s > \tau_t$ and $c \in \mathcal{C}$, $\mathbb{P}(C_s(\omega) = c \,|C_{\tau_t}(\omega), \rho^S_{\tau_t}(\omega), \mathcal{F}^{(Z)}_t) = \mathbb{P}(C_s(\omega) = c \,|C_{\tau_t}(\omega), \rho^S_{\tau_t}(\omega))$.

Our second assumption (A2) intuitively says that for predicting coarsened states beyond some point τ_t, all information in the point history up to time t is summarized by the distribution of $C_{\tau_t}(\omega)$, and by who the striker is at τ_t, $\rho^S_{\tau_t}(\omega)$. The dynamics of tennis make this second assumption reasonable: Every time a player strikes the ball, this represents a structural transition in the tennis point to which both players (mostly the returner) react. Their actions prior to this transition are not likely to influence their actions after this transition. Given $C_{\tau_t}(\omega)$ - which includes the locations of each player at the beginning of the current shot duration - and given who the striker is of this shot duration, $\rho^S_{\tau_t}(\omega)$, data prior to the corresponding shot duration will not help predict future evolutions of the points. Since we are going shot-by-shot with the ESWR estimator and its methodology, the information at the beginning of the shot duration, $C_{\tau_t}(\omega)$ and $\rho^S_{\tau_t}(\omega)$, will be used to predict the outcome of the shot duration, $C_{\delta_t}(\omega)$.

Using these two assumptions we are able to simplify expression (4), which combines aspects from the full-resolution and coarsened views of the process.

Theorem 1. *Under assumptions (A1) and (A2), the full resolution ESWR ν^i_t, for player i at time t of a point being played against player j, can be rewritten:*

$$\nu_t^i = \mathbb{E}[\Gamma^i|\ \mathcal{F}_t^{(Z)}] = \sum_{c_v \in \mathcal{C}} \mathbb{E}[\Gamma^i|\ C_{\delta_t} = c_v]\ \mathbb{P}(C_{\delta_t} = c_v|\ \mathcal{F}_t^{(Z)})$$

$$= \begin{cases} \sum_{c_v \in \mathcal{C}} \mathbb{E}[\gamma_S^i|\ C_{\delta_t} = c_v]\ \mathbb{P}(C_{\delta_t} = c_v|\ C_{\tau_t} = c_u,\ \rho_{\tau_t}^S = i),\ \textit{if } i = \rho_{\tau_t}^S \\ \sum_{c_v \in \mathcal{C}} \mathbb{E}[\gamma_R^i|\ C_{\delta_t} = c_v]\ \mathbb{P}(C_{\delta_t} = c_v|\ C_{\tau_t} = c_u,\ \rho_{\tau_t}^S = j),\ \textit{if } i = \rho_{\tau_t}^R \end{cases} \tag{7}$$

The proof of Theorem 1 follows directly from assumptions (A1) and (A2) and is therefore omitted. Note that the dependence on ω was also omitted in (7) in order to save space. Heuristically, (7) expresses $\nu_t^i(\omega)$ as the expectation given by a homogeneous Markov Chain on $\mathcal{C}$ with a random starting point $C_{\delta_t}(\omega)$, where only the starting point depends on the full-resolution information $\mathcal{F}_t^{(Z)}$, which now can be summarized with $C_{\tau_t}(\omega)$ and $\rho_{\tau_t}^S(\omega)$.

4.3 TRANSITION MODEL SPECIFICATION

The representation of the ESWR estimator in (7) shows that estimating ESWR does not require a full-blown model for the entire tennis point at high resolution. Instead, the priority is to accurately predict how much the next state, denoted by $C_{\delta_t}(\omega)$, will contribute to a won point for both the striker and the returner. Let us define the *Markov transition probability matrix* [11] for when player i is the striker as $\mathbf{P}^i$, with

$$P_{uv}^i = \mathbb{P}(C^{(n+1)} = c_v | C^{(n)} = c_u, \rho^S = i) \tag{8}$$

for any $c_u,\ c_v \in \mathcal{C}$, and where ρ^S indicates the striker's player ID. Without any other probabilistic structure assumed for $\mathbb{P}(C^{(n+1)} = c_v|C^{(n)} = c_u, \rho^S = i)$ other than Markov, for all u,v, the maximum likelihood estimator of P_{uv}^i is the observed transition frequency,

$$P_{uv}^i = \frac{N_{uv}^i}{\sum_{v'} N_{uv'}^i}$$

where N_{uv}^i counts the number of transitions $c_u \to c_v$ when player i is the striker. The pseudocode for forming the transition probability matrix can be found in Appendices D and E. Intuitively, this states that the striker ultimately decides what the next state will be, c_v, based on the current state, c_u, whether that be the end of the point or the next strike state. Also, note that the results from this estimator will be undesirable if the number of visits to a particular state c_u is small. For example, striking from a 1-4 strike state is very uncommon (both players are at the net, in the out-of-bounds area of the court, right next to each other), and the estimated transition probabilities from that state may be degenerate. Uncommon strike states will pop up every now and then, but since they are uncommon, and due to the limited amount of data and time,

we will not worry ourselves with these kinds of shots. In Section 8, possible solutions to this problem are proposed.

4.4 STRIKE WIN RATE AND RETURN WIN RATE

Since we are incorporating transition probability matrices to predict how much the outcome state, or the *next* state, will contribute to a won point, we must consider all possibilities. Say player i is striking the ball from a given strike state. If that strike by player i does not result in the end of the point, which will give player i a defined point value of either 1 or 0, and the point continues, then we must attempt to quantify how much the next potential return state will contribute to a won point for player i. Conversely, during a match between players i and j, if player i is attempting to return player j's strike, and that strike does not end the point, then we must look ahead and attempt to quantify how much the next potential strike state will contribute to a won point for player i. To do this we will create two metrics: *Strike win rate* and *return win rate*.

First, let $\mathcal{S}$ contain every strike contained in the specified dataset, which includes who struck the shot, its strike state and its corresponding strike significance. Thus, the possible values contained in $\mathcal{S}$ reside in the set

$$\mathcal{S} = \{\text{player ID}\} \times \{\mathcal{P}\} \times \{\Lambda_{\text{SS}}\}.$$

Let $S^i \subseteq \mathcal{S}$ denote the subset of all strikes in $\mathcal{S}$ struck by player i. Furthermore, let $S^i_{\beta^S} \subseteq S^i$ denote the subset of all strikes in $\mathcal{S}$ struck by player i from the strike state $\beta^S \in \mathcal{P}$. Lastly, let $S^i_{\beta^S,\lambda_{\text{SS}}} \subseteq S^i_{\beta^S}$ denote the subset of all strikes in $\mathcal{S}$ struck by player i from the strike state β^S, with the strike significance $\lambda_{\text{SS}} \in \Lambda_{\text{SS}}$. Similarly, let $\mathcal{R}$ contain every return in the specified dataset, which includes who is returning the shot, its return state and its corresponding return significance. Thus, the possible values contained in $\mathcal{R}$ reside in the set

$$\mathcal{R} = \{\text{player ID}\} \times \{\mathcal{P}\} \times \{\Lambda_{\text{RS}}\}.$$

Let $R^i \subseteq \mathcal{R}$ denote the subset of all returns in $\mathcal{R}$ returned by player i. Furthermore, let $R^i_{\beta^R} \subseteq R^i$ denote the subset of all returns in $\mathcal{R}$ returned by player i from the return state $\beta^R \in \mathcal{P}$. Lastly, let $R^i_{\beta^R,\lambda_{\text{RS}}} \subseteq R^i_{\beta^R}$ denote the subset of all returns in $\mathcal{R}$ returned by player i from the return state β^R, with the return significance $\lambda_{\text{RS}} \in \Lambda_{\text{RS}}$.

Now, *strike win rate* will be defined as how much a given strike from a specific strike state contributes to a won point. The strike win rate for player k from strike state β^S will be defined as follows:

$$\text{SWR}(k, \beta^S) = \sum_{\lambda_{\text{SS}} \in \Lambda_{\text{SS}}} w_{\lambda_{\text{SS}}} \frac{|S^k_{\beta^S,\lambda_{\text{SS}}}|}{|S^k_{\beta^S}|}, \tag{9}$$

where $|x|$ is the cardinality of set x, counting the number of elements it contains. A strike win rate of 0.50 indicates a neutral strike which does not necessarily contribute to a winning point or a losing point. A strike win rate above 0.50 indicates an above-average win contribution for a strike, and below 0.50 indicates a below-average win contribution.

Next, *return win rate* will be defined as how much a given return contributes to a won point. The return win rate for player k from return state β^R will be defined as follows:

$$\text{RWR}(k, \beta^R) = \sum_{\lambda_{\text{RS}} \in \Lambda_{\text{RS}}} w_{\lambda_{\text{RS}}} \frac{|R^k_{\beta^R, \lambda_{\text{RS}}}|}{|R^k_{\beta^R}|}. \tag{10}$$

A return win rate of 0.50 indicates a neutral return state which the player does a good job of keeping the ball in play from, but does not necessarily tend to return winning strikes or losing strikes. A return win rate above 0.50 indicates an above-average win contribution for a return, and below 0.50 indicates a below-average win contribution.

4.5 DEFINING THE CALCULATION OF ESWR

Recall our rewritten definition of the ESWR estimator (7):

$$\begin{aligned}
\nu^i_t &= \sum_{c_v \in \mathcal{C}} \mathbb{E}[\Gamma^i |\ C_{\delta_t} = c_v]\ \mathbb{P}(C_{\delta_t} = c_v |\ \mathcal{F}^{(Z)}_t) \\
&= \begin{cases} \sum_{c_v \in \mathcal{C}} \mathbb{E}[\gamma^i_S |\ C_{\delta_t} = c_v]\ \mathbb{P}(C_{\delta_t} = c_v |\ C_{\tau_t} = c_u,\ \rho^S_{\tau_t} = i), \text{ if } i = \rho^S_{\tau_t} \\ \sum_{c_v \in \mathcal{C}} \mathbb{E}[\gamma^i_R |\ C_{\delta_t} = c_v]\ \mathbb{P}(C_{\delta_t} = c_v |\ C_{\tau_t} = c_u,\ \rho^S_{\tau_t} = j), \text{ if } i = \rho^R_{\tau_t} \end{cases}
\end{aligned}$$

First, let us look at the $\mathbb{P}(C_{\delta_t}(\omega) = c_v |\ \mathcal{F}^{(Z)}_t)$ portion of our ESWR estimator. To calculate this value, we will rely on the transition probability matrices we have created for each player. For a match being played between player i and player j, this probability will be calculated using the transition probability matrices $\mathbf{P}^i$ and $\mathbf{P}^j$. If we are attempting to calculate $\nu^i_t(\omega)$, and $i = \rho^S_{\tau_t}(\omega)$, then the probability $\mathbb{P}(C_{\delta_t}(\omega) = c_v |\ \mathcal{F}^{(Z)}_t)$ will be equivalent to P^i_{uv}:

$$\mathbb{P}(C_{\delta_t}(\omega) = c_v |\ \mathcal{F}^{(Z)}_t) = \mathbb{P}(C_{\delta_t}(\omega) = c_v |\ C_{\tau_t}(\omega) = c_u, \rho^S = i) = P^i_{uv}. \tag{11}$$

Conversely, if we are attempting to calculate $\nu^i_t(\omega)$, and $j = \rho^S_{\tau_t}(\omega)$, then the probability $\mathbb{P}(C_{\delta_t}(\omega) = c_v |\ \mathcal{F}^{(Z)}_t)$ will be equivalent to P^j_{uv}:

$$\mathbb{P}(C_{\delta_t}(\omega) = c_v |\ \mathcal{F}^{(Z)}_t) = \mathbb{P}(C_{\delta_t}(\omega) = c_v |\ C_{\tau_t}(\omega) = c_u, \rho^S = j) = P^j_{uv}. \tag{12}$$

From a tennis perspective this makes sense, since the striker ultimately controls where the ball travels next, dictating the next player location combination or if the point ends as a result of this strike.

Now, we must look at the $\mathbb{E}[\gamma_S^i(\omega)|\ C_{\delta_t}(\omega) = c_v]$ and $\mathbb{E}[\gamma_R^i(\omega)|\ C_{\delta_t}(\omega) = c_v]$ portions of our ESWR equation. As was discussed in Section 4.4, we must take into account not only the possibility that the point will end on a given strike, but also the possibility that the point continues. Thus, we will need to break down $\mathbb{E}[\gamma_S^i(\omega)|\ C_{\delta_t}(\omega) = c_v]$ and $\mathbb{E}[\gamma_R^i(\omega)|\ C_{\delta_t}(\omega) = c_v]$ into two parts: One part which takes into account the potential end of the point, and the other which takes into account the potential next state and the continuance of the point. From player i's striking perspective, we will incorporate the return win rate metric, RWR(i,β^R), and the point outcome from end states, $X_S(c \in \mathcal{C}_{\text{end}})$. From player i's returning perspective we will incorporate the strike win rate metric, SWR(i,β^S), and the point outcome from end states, $X_R(c \in \mathcal{C}_{\text{end}})$. Thus, for a match being played between players i and j, the ESWR for player i and point path ω will be calculated as:

$$\begin{aligned}\nu_t^i(\omega) &= \sum_{c_v \in \mathcal{C}} \mathbb{E}[\Gamma^i(\omega)|\ C_{\delta_t}(\omega) = c_v]\ \mathbb{P}(C_{\delta_t}(\omega) = c_v|\ C_{\tau_t}(\omega) = c_u, \mathcal{F}_t^{(Z)}) \\ &= \begin{cases} \sum\limits_{c_w \in \mathcal{C}_{\text{Shot}}} \text{RWR}(i, c_w) P_{uw}^i + \sum\limits_{c_x \in \mathcal{C}_{\text{end}}} X_S(c_x) P_{ux}^i, \text{ if } i = \rho_{\tau_t}^S(\omega) \\ \sum\limits_{c_w \in \mathcal{C}_{\text{Shot}}} \text{SWR}(i, c_w) P_{uw}^j + \sum\limits_{c_x \in \mathcal{C}_{\text{end}}} X_R(c_x) P_{ux}^j, \text{ if } i = \rho_{\tau_t}^R(\omega) \end{cases} \end{aligned} \tag{13}$$

5 RESULTS

In order to see if the proposed ESWR estimator is actually producing results players and coaches can rely on, comparisons will be made between players who won more matches during the 2015 US Open and players who did not win as many. The strike win rate and return win rate metrics are backbones of the ESWR estimator, thus, we need to make sure these two metrics are properly playing their part when calculating win contribution. Since these metrics measure how much strikes and returns contribute to a winning point, theoretically, the players who win more games should separate themselves from the players who win less games. As was mentioned earlier, the USTA, on top of the five-match data sample, allowed the ESWR estimator and its methodology to be run on the entirety of their 2015 US Open data. Thus, every shot of every match on a major court during the 2015 US Open was analyzed using the ESWR methodology, which includes 71,368 shots from 73 different players. Sections 5.1 and 5.2 will attempt to validate the ESWR methodology and show the SWR and RWR metrics are producing reliable results. The ESWR estimator will be applied to an individual tennis point in Section 5.3, and the potential of the ESWR and its methodology will be shown in Section 5.4, with regards to guiding player strategy and insight.

5.1 ANALYSIS OF STRIKE WIN RATE

In theory, players who have won more matches should have more strike states which exceed 0.50, or strike states that give above-average win contributions. Below is a visualization of the percent of total strike states, β^S's, with SWR(k, β^S)'s above 0.50 for each individual player k in the entirety of the 2015 US Open data. These players will then be grouped together based on how many matches they played in. Since the US Open is a single-elimination tournament, more matches for a player is equivalent to saying more wins and more won points. Also, it should be noted only strike states where each player had at least 1 strike from were considered.

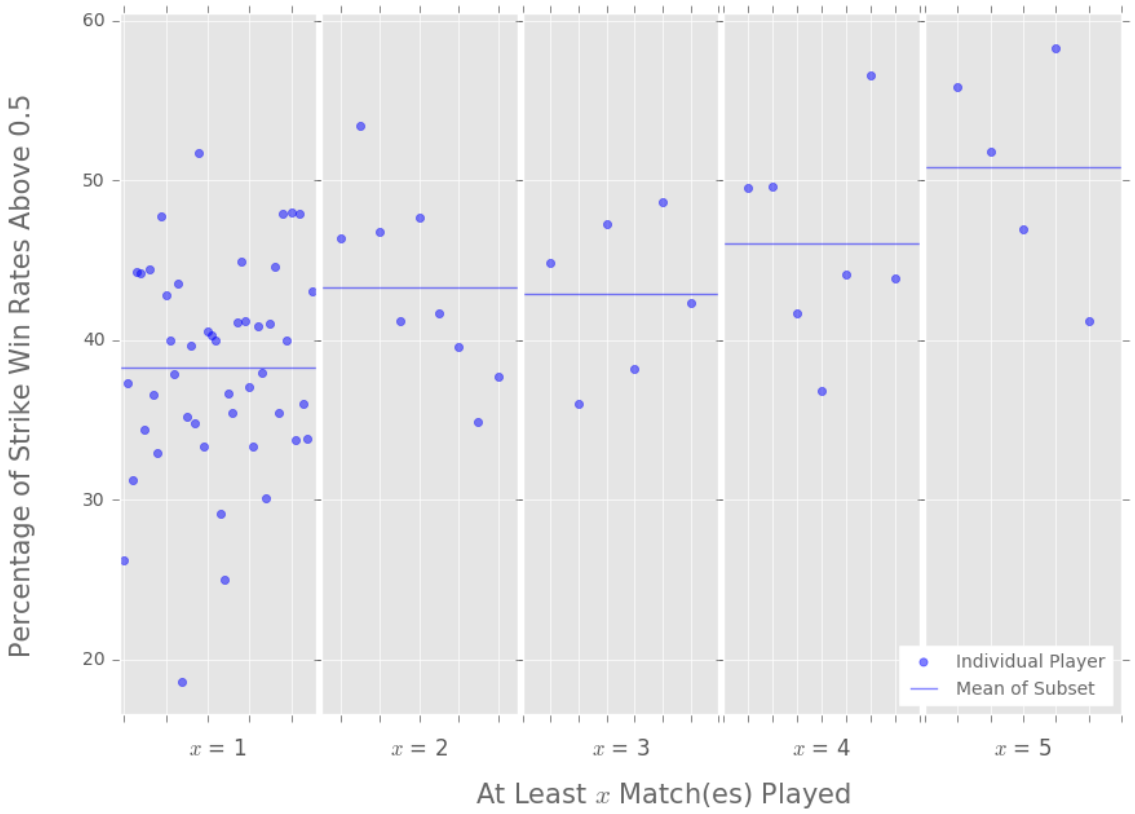

FIGURE 13: Visualization of the percentage of strike states with strike win rates above 0.50, by individual players, and by groups of players. The dots represent individual players, and the lines represent the mean of each subset.

As is shown in Figure 13, as players play and win more matches, they naturally have a higher percentage of strike states where they have above 0.50 strike win rates. The subset of players who played at least 5 matches[7], on average, had more than half of their strike states have positive win contributions. Compare

[7]This subset of players includes players who played in 5, 6 and 7 matches. The reason for grouping these players together is that they have a comparable amount of wins, which exceed the other subsets of players. Also, if we created new subsets for

this to the subset of players who only played 1 match, and lost, who had on average less than 40% of their strike states having positive win contributions. Clearly, there are players who only played and lost one match and still had high proportions of their strike states above average. But, as a general trend, players who win more have higher proportions of their strike win rates above 0.50.

5.2 ANALYSIS OF RETURN WIN RATE

Again, in theory, players who have won more matches should have more return states with return win rates which exceed 0.50, or return states that give above-average win contributions. Figure 14 is a visualization of the percent of total return states, β^R's, with RWR(k, β^R)'s above 0.50 for each individual player k. These players are then grouped together based on how many matches they played in, similar to the previous section. Only return states where each player had at least 1 return were considered.

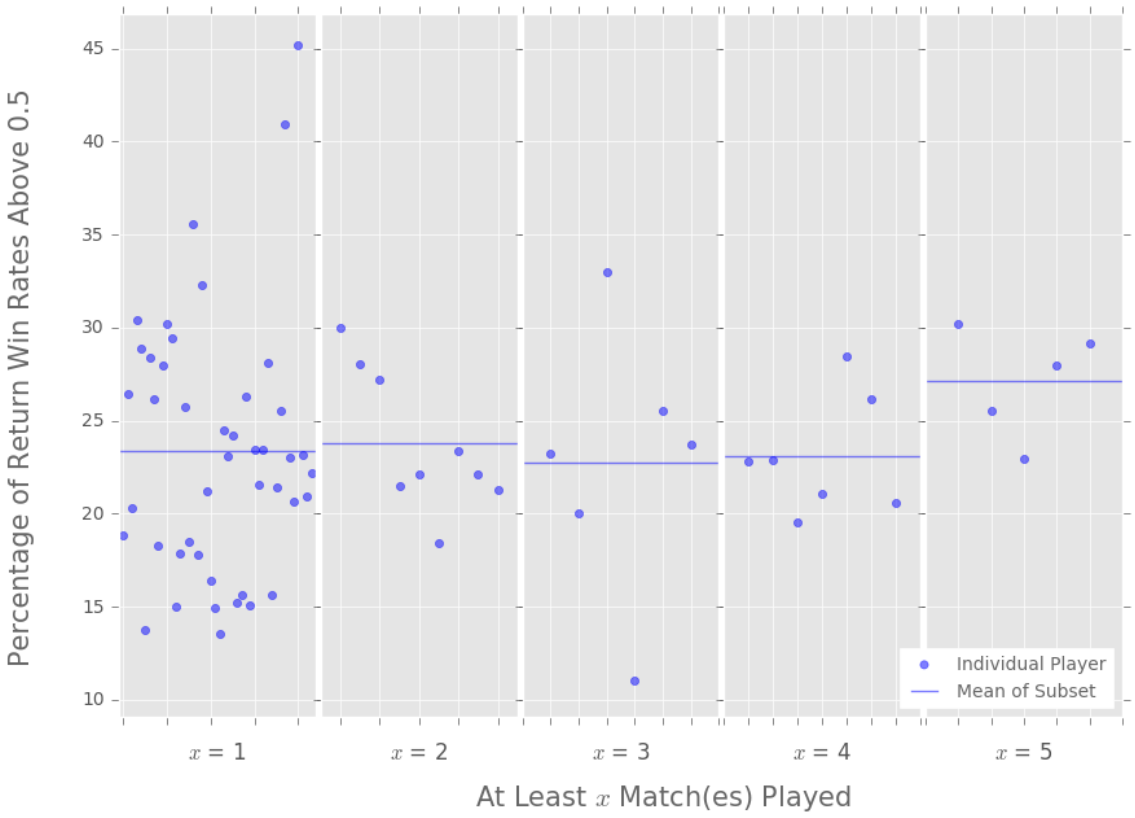

FIGURE 14: Visualization of the percentage of return states with return win rates above 0.50, by individual players, and by groups of players. The dots represent individual players, and the lines represent the mean of each subset.

players who played in exactly 5, 6 and 7 matches, the number of players in each one would be very small compared to the other subsets.

By observing Figure 14, we can see that, generally, the mean percent of return win rates above 0.50 is about the same for each subset of players, with the mean marginally increasing from players who played 1 match to players who played at least 5 matches. This is not totally surprising, considering returning is much more random than striking. When striking, the player is able to control where the ball travels next. But, when returning, the player has no control of where the ball travels to next, making the ability to produce quality returns much more random. A returner can *influence* where the ball might go to next by setting up in certain locations, but, ultimately, the striker has control of the point. Generally, however, the players who win more matches are able to produce quality returns from more return states than those who win less.

5.3 ESWR ESTIMATOR APPLIED TO AN INDIVIDUAL POINT

Figures 15 and 16 are the ESWR versions of [5]'s stock-ticker visualization. Since they take turns being on the offensive and the defensive, two visualizations, from the perspective of each player, were created to properly describe the evolution of the given tennis point. The labels next to each point in the graph describe the corresponding strike state or return state, dependent on whether the player was the striker or returner of the most previous shot. The marker at 10 seconds marks the end of the point, resulting from player 2 striking a pure winner. Thus, player 1's marker at the end of the point has an ESWR of 0 (Figure 15), and player 2's marker at the end has an ESWR of 1 (Figure 16). Note that the five-match sample data was used for this analysis, in order to properly analyze player decisions in Section 5.4.

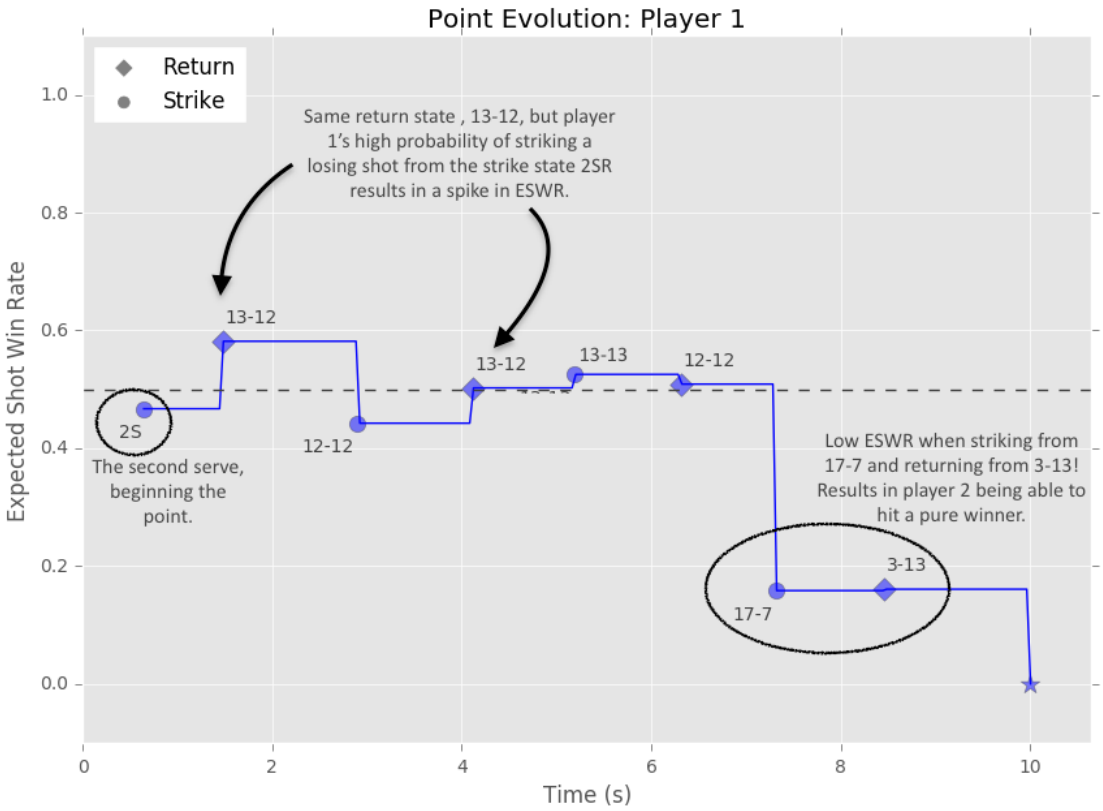

FIGURE 15: The stock-ticker visualization of the ESWR estimator for a specific point through player 1's perspective.

When player 2 attempts to strike a second-serve return, player 1 achieves his highest ESWR of the point at 0.582. His ESWR is never as high again (Figure 15). After the second-serve return, a series of shots occur next, with both players hovering around the baseline area, and the corresponding ESWR's hovering around 0.50. Player 1 ends up having to strike from 17-7, dooming him. His ESWR from this strike state drops to 0.158, and this results in player 1 striking a weak shot, which player 2 is able to volley (not let the ball bounce) and strike a pure winner from the 3-13 strike state. Thus, having to strike from the 17-7 significantly contributes to player 1 losing the point, at least through the lens of the ESWR estimator.

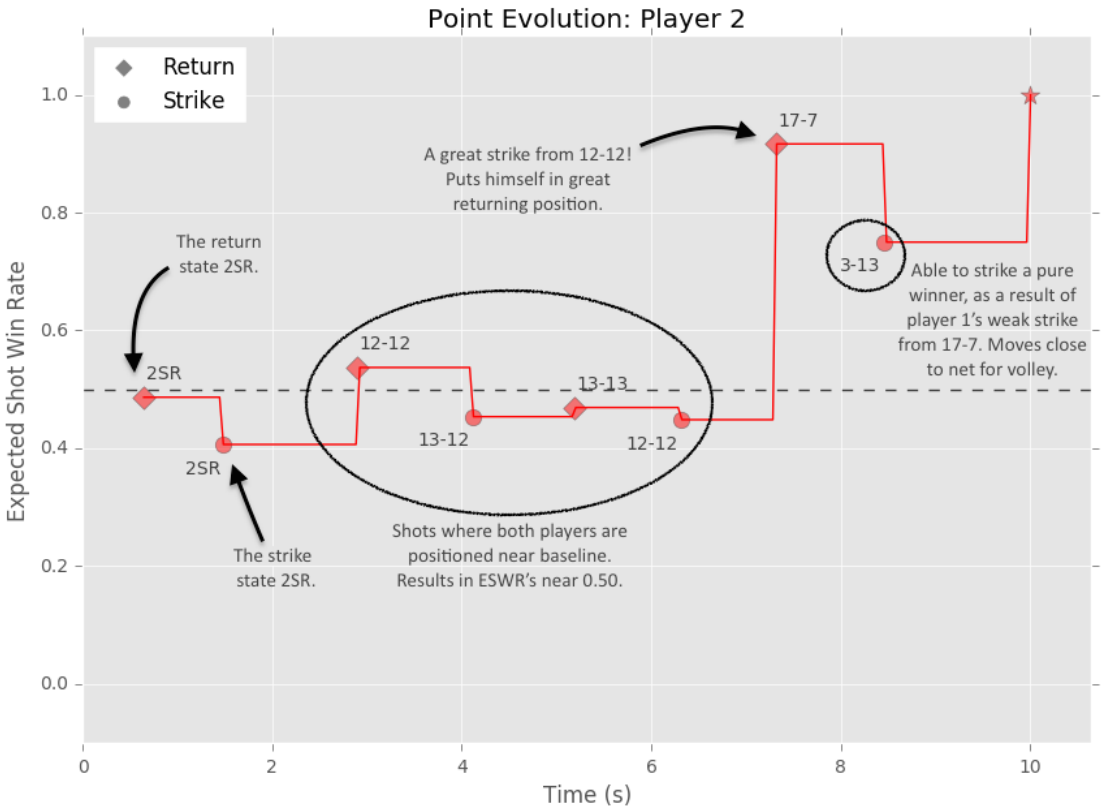

FIGURE 16: The stock-ticker visualization of the ESWR estimator for a specific point through player 2's perspective.

For the first six shots of the point, player 2 had ESWR's hovering around 0.50, indicating neutral strikes and returns (Figure 16). On the third-to-last strike of the point, player 2 strikes an unexpected great strike from the strike state 12-12. This strike forced player 1 to strike from region 17, and, at the same time, player 2 moved up towards the net to region 7. This 17-7 return state gave player 2 an ESWR of 0.917, his highest of the point. As was mentioned when analyzing Figure 15, 17-7 is not a good strike state for player 1, and this results him striking a weak shot. Player 2 continues his advance towards the net, and ends up volleying and striking a pure winner from the strike state 3-13. Thus, player 2 being able to hit a great set-up shot from the strike state 12-12 ultimately swung and won the point for him.

Now, let us go into the estimator and figure out what caused the spikes in ESWR for player 2 and the dips in ESWR for player 1. Below in Figure 17 the probability of the next state is evaluated at each strike of the point analyzed in this section. Given the strike state of the past strike at τ_t, $\beta^S_{\tau_t}(\omega)$, and who struck it, $\rho^S_{\tau_t}(\omega)$, three probabilities were calculated: The probability of a won point as result of the strike, $\mathbb{P}(W|\beta^S_{\tau_t})$, the probability of a lost point, $\mathbb{P}(L|\beta^S_{\tau_t})$ and the probability the point continues, $\mathbb{P}(!(W \vee L)|\beta^S_{\tau_t})$.

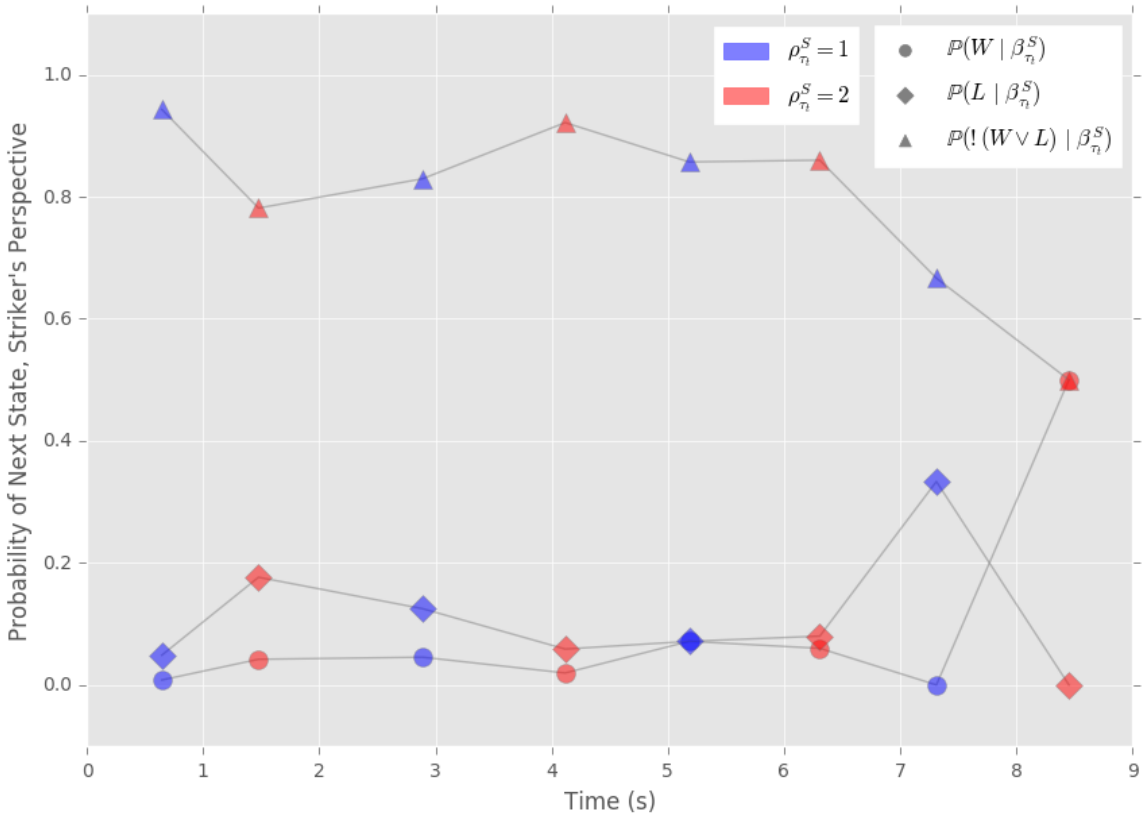

FIGURE 17: The probability of the next state evaluated at each strike of the point being analyzed between player 1 and player 2. Dependence on ω is omitted in order to save space.

We can see that when player 2 strikes from strike state 2SR at around 1.5 seconds there is a relatively high probability of a lost point as a result of the strike, near 0.20. This results in a high ESWR for the returner, player 1, and a low ESWR for the striker, player 2. For the next four strikes, the probabilities of the point ending as a result of the strike are considerably low. Since 12-12 is a common strike state, where both players are located near the center of the baseline, it is unexpected when player 2 hits a great set-up shot from this strike state. This leads to a 17-7 strike state for player 1, where he is highly likely to hit a losing strike, with a probability around 0.30. He manages to keep the ball in play, however, but this merely gives player 2 a strike state of 3-13, where he has a 0.50 probability of hitting a pure winner. Player 2 does just that, as he volleys and places a pure winner in front of player 1, who is back in region 13 at the time of the strike. Also, it should be noted that there were only two shots in the point where the striker had a higher ESWR value than the returner. This is because the ESWR estimator takes into account the probability of the striker striking the ball out-of-bounds or into the net, and usually this probability is higher than that of striking a pure winner. Thus, this probability is given a weight of 0 for the striker and a weight of 1 for the returner, which results in the returner having a higher ESWR value most times.

The approach taken in this thesis to evaluating the evolution of individual tennis points was to go from shot-to-shot and look at each shot individually, taking into account both direct shots and indirect shots. This specific point was included because it supports this approach, as the outcome of the point was swung on a single shot near the middle of the point. In the point evaluated in Figures 15 and 16, the five groundstrokes following the first serve were all shots where both players were near the baseline, resulting in a lack of variation in the ESWR since these are common types of shots. But, when player 2 is able to force player 1 to strike from the 17-7 strike state, player 2 is able to take control of the point. Thus, this point's outcome was ultimately decided by the third-to-last strike of the point, when player 2 forced some variation to where both players were located.

5.4 GUIDING PLAYER STRATEGY AND INSIGHT

In order to demonstrate how ESWR and its methodology can guide player strategy, let us look at one of the players' decisions from the point analyzed in Section 5.3. Specifically, after player 2 struck from the 12-12 strike state at around 6.5 seconds, did he pick the best region to set up and return player 1's next strike from region 17? What is the worst region he could have set up in?

Assume that player 1 will still strike the ball from region 17. Let us examine player 2's return win rate when the striker is striking from region 17, and let us examine player 1's strike win rate when he is striking from region 17. Now, player 2 traveled a distance of 3.04 meters from when he struck from the 12 region to when he set up to return player 1's strike from region 17. This means that in the same distance or less, player 2 could have instead traveled to regions 6, 11, 13, 14, 17 or, of course, stayed put in region 12. Let us look specifically at the strike win rates for player 1 and return win rates for player 2 from the corresponding strike and return states where the striker resides in region 17 and the returner sets up in one of the regions where player 2 could have gone to. Note that we are not analyzing ESWR in Table 5, but strike win rates and return win rates, as they give a good representation of how a player strikes and returns. Also, the two metrics take into consideration pure winners, out-of-bounds strikes and strikes which go directly into the net, as does the ESWR estimator.

β^S/β^R	SWR(1,β^S)	RWR(2,β^R)
17-6	0.5000	0.5500
17-7	0.2500	0.6500
17-11	0.5000	0.5750
17-12	0.4706	0.5139
17-13	0.5179	0.5500
17-14	N/A	N/A
17-17	0.7500	0.3929

TABLE 5: Describes the strike win rates of player 1, SWR(1,β^S), and return win rates of player 2, RWR(2,β^R). The considered return states and strike states consisted of the striker residing in region 17, and the returner residing in several different regions. "N/A" indicates player 1 had no prior history striking from the strike state of 17-14, and player 2 had no prior history returning from the return state 17-14.

Since the weights used in SWR and RWR are symmetric, the two metrics are able to be compared. According to the strike win rates and return win rates in Figure 5, player 2 picked the best area to set up to return player 1's strike, based on differences between RWR and SWR. Player 2 could have traveled to virtually any of the other regions, and still held the advantage over player 1 striking from region 17. However, if player 2 had chosen to backpedal to region 17, this would have given player 1 the advantage. From the strike state 17-17, player 1 has a strike win rate of 0.7500, and and from the return state 17-17, player 2 has a return win rate of 0.3929. Player 1 was in a very difficult position striking from region 17, but if player 2 made a mistake in regards to his positioning, player 1 may have been able to salvage this point.

Another perspective in which the ESWR methodology can assist player strategy is where to land a strike, depending on the returner's weaknesses. Again, we will look at two players, player 1 and player 2 (not necessarily the same player 1 and player 2 analyzed previously.) Say player 1 is striking the ball from region 11, and player 2 is setting up to return the strike from region 12. Based on this location categorization of 11-12, where should player 1 aim to strike the ball next? What bounce locations has player 2 had trouble returning before, from the return state of 11-12? Below is visualization describing this situation:

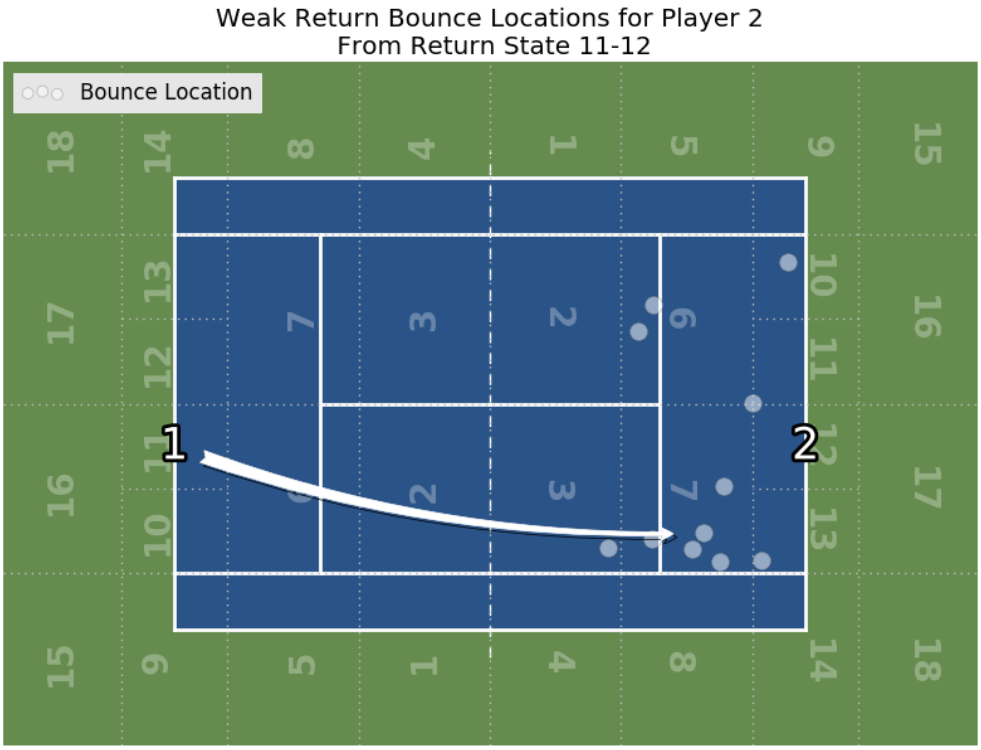

FIGURE 18: Player 1 as the striker in region 11, and player 2 as the returner in region 12. The white circles indicate bounce locations of strikes player 2 has had trouble returning. The arrow indicates where player 1 should be aiming to land his next shot.

Now, when we say "bounce locations player 2 has had trouble returning," we refer to the bounces of returns categorized with either λ_{RLS}, λ_{RN}, or λ_{RSUOPW} return significances. In Figure 18 are the bounce locations of returns given one of those three return significances, when player 2 had a return state β^R of 11-12. The arrow is pointing to the median (x, y) location of all the bounce locations, which is located near the right sideline in region 7. Thus, while there have been shots on the opposite sideline which player 2 has had trouble returning in the past, player 1 should aim to land his strike near the right sideline since that is where a majority of the bounces player 2 has had trouble returning landed.

From an insight perspective, it would interesting to know how a player strikes or returns in comparison to other players. More specifically, how does a given player perform in comparison to how the average player performs? Does he perform above or below average from certain strike states or return states, and by how much?

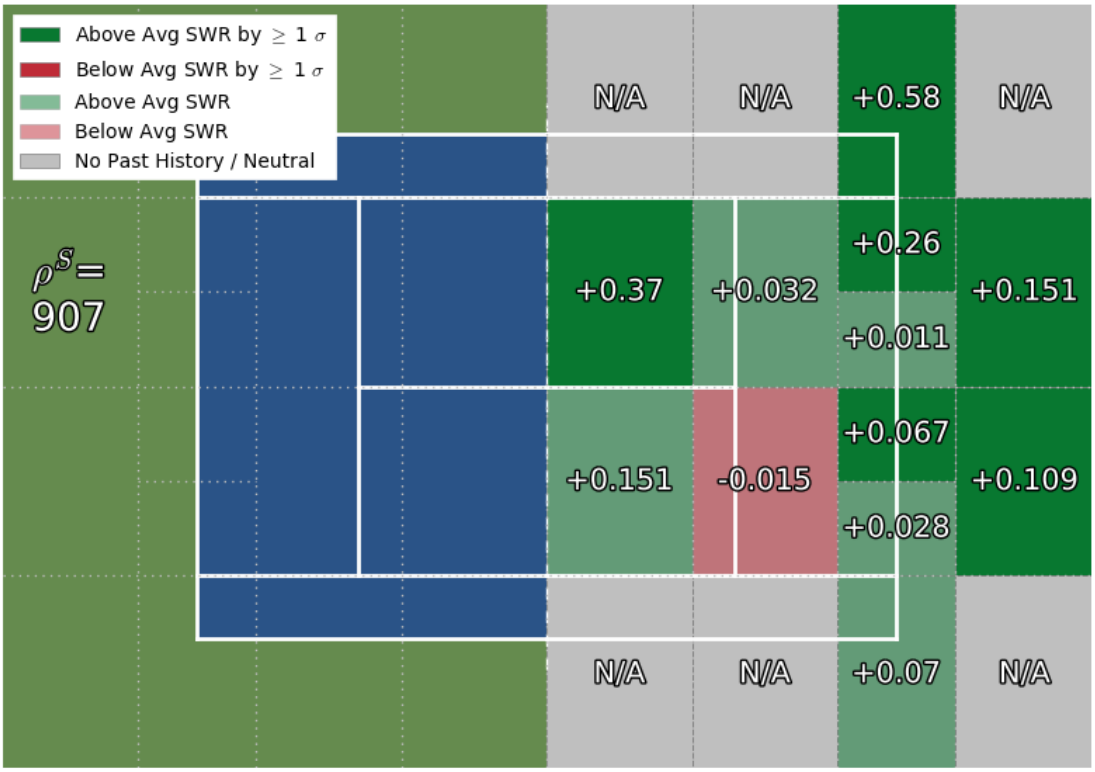

FIGURE 19: Player 907 striking from region 17. Each region on the opposite side of the court color-coded to describe how many standard deviations above/below player 907's strike win rate is to the average, if the returner was to set up in that region. The numbers inside the regions describe the net difference between player 907's strike win rate and the average for the corresponding strike state. The gray regions with "N/A" are regions where no opponent has set up to return against player 907 when player 907 strikes from region 17.

Figure 19 attempts to visualize just this. Say player 907 is the striker, $\rho^S = 907$, and he set up to strike in region 17. Which regions that the returner chose to set up in would give player 907 an above or below average strike win rate, from the corresponding strike state? Say the returner chose to set up in region 7. This would give player 907 a below average strike win rate by 0.015, where this number was produced by calculating SWR(907,17-7), and then subtracting the mean of SWR(k,17-7) for all other players k. For all other regions where he has had past returners set up against him when he strikes from region 17, player 907 is above average, and above average by at least one standard deviation when returners set up in regions 2, 9, 10, 12, 16 and 17. Thus, it is clear player 907 enjoys to strike farther back from the baseline in region 17.

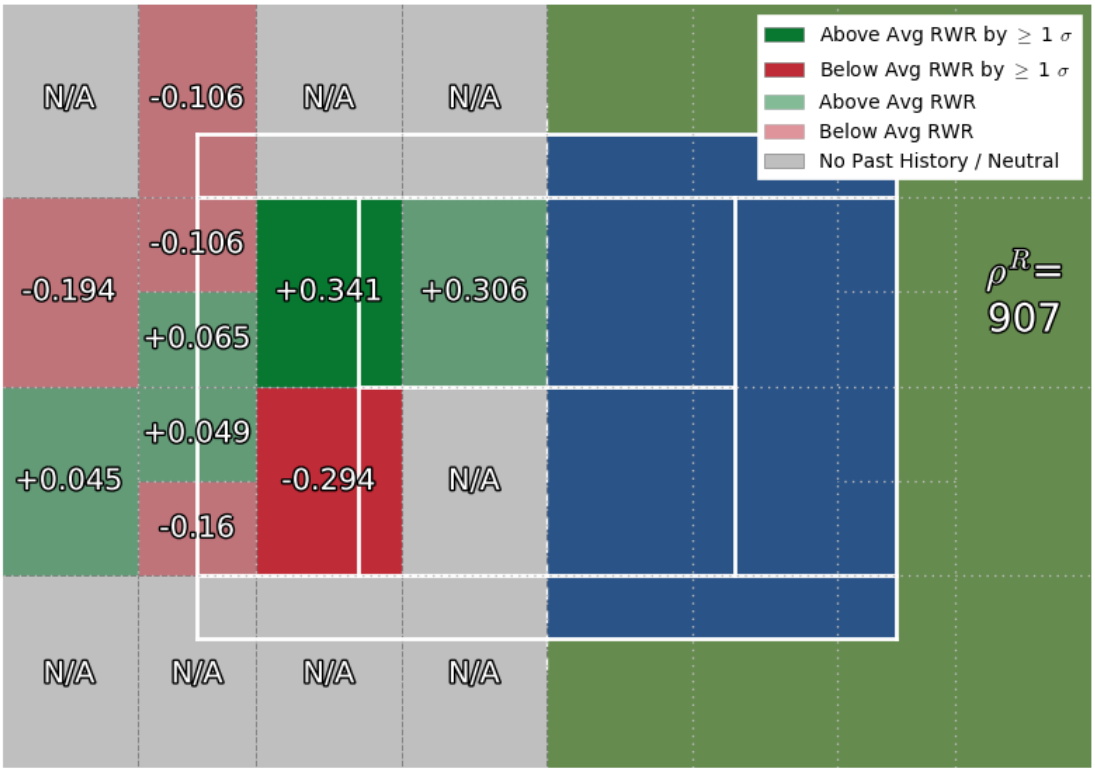

FIGURE 20: Player 907 returning from region 16. Each region on the opposite side of the court color-coded to describe how many standard deviations above/below player 907's return win rate is to the average, if the returner was to set up in that region. The numbers inside the regions describe the net difference between player 907's return win rate and the average for the corresponding return state. The gray regions with "N/A" are regions where no opponent has set up to strike against player 907 when player 907 sets up to return from region 16.

Conversely, Figure 20 depicts how player 907's return win rate fares against the average when he sets up to return in region 16. Which regions that the striker strikes from would give player 907 an above or below average return win rate, from the corresponding return state? If the striker struck from region 7, player 907's return win rate would be above average by at least one standard deviation, and if the striker struck from region 6, player 907's return win rate would be below average by at least one standard deviation.

6 CONCLUSIONS

In Section 5, the potential of the ESWR estimator and its methodology was shown. Sections 5.1 and 5.2 provided justification for the strike win rate and the return win rate metrics, which are the backbones of the main ESWR estimator. Winning players separated themselves from losing players when looking at their strike and return win rates, which is exactly what was desired to be seen, as these are metrics which measure

how much a strike or return contributes to a *winning* point. It should be noted the distinction was more evident with the strike win rate, but there was still a level of separation with the return win rate. Section 5.3 showed the ESWR estimator applied to an individual point, and demonstrated how ESWR can be used to describe the evolution of a point, and how it can be used to "tell the story" of a point. The same metrics have been used for decades to try and communicate to viewers how and why a tennis match was playing out as it was. The ESWR estimator provides potential to give a new, shot-by-shot perspective to players and fans as to why and how a point or match was won. The potential influence on player strategy was discussed in Section 5.4, showing how the ESWR and its methodology can be used to target the opponent's weaknesses and to avoid their strengths. Also, another strength of the strike and return win rate metrics, along with the ESWR estimator, is that they are relatively simple, and could be easily explained to the average tennis fan.

Interesting and insightful comparisons can also be made between players with the methodology shown in this thesis, as players' strike and return win rates were compared to the average. Possibly most importantly, the ESWR estimator and its methodology can guide player strategy in regards to positioning during a match. Everything shown in this Section 5 are examples of how useful and versatile the ESWR estimator and its methodology can be.

7 EXPLORING THE DATA

The metrics described previously in this thesis solely rely on player locations and players' tendencies from those locations, based on weighting past shots to determine how much a given shot would contribute to a winning point. But what if we took it a step further, and started looking at other shot characteristics such as type of shot, shot spin and shot speed? The US Open data contains these variables along with exact spatial coordinates, and they could easily be incorporated into further categorization of strikes. For example, say a player is striking from a given location categorization, and he wanted to know how his strike win rate would go up or down if he tried a slower drop-shot which lands just past the net, rather than a faster strike which lands closer to the opposite baseline? How would the strike win rate fare if he decided to try a strike with back-spin instead of top-spin? What if he decided on a backhand over a forehand? There are many distinct types of strikes in the game of tennis, and this thesis, up to this point, has essentially assumed all strikes were of the same type[8]. In this section we will analyze the strike win rates of well-known strike types, with the use of our five-match sample data.

[8]This is in part because the five-match sample dataset was the only one available for a majority of this research. The access to the entirety of the data was not gained until close to the due date of the submission of this thesis. Hence, it did not seem like a good idea to rework the entire ESWR estimator and its methodology with such little time remaining.

7.1 ANALYZING UNIQUE STRIKE TYPES

The first strike type which will be analyzed will be the "drop shot." The drop shot is a relatively soft-hit strike which bounces low and near the net, making it difficult to for the opponent to return if he is expecting a normal shot and is setting up back on the baseline. If we sort through all strikes in the five-match sample data, and we limit ourselves to ground-strokes which had bounce x-coordinate magnitudes less than or equal to 4.0 meters and greater than or equal to 0.5 meters[9], the 4,474 total strikes in the data went down to 98. The strike win rate of these 98 shots was an above average 0.650. As was previously mentioned, along with bouncing close to the net, another defining characteristic of a good drop shot is a low bounce height. A good way to to do this is to put back-spin on the strike, as back-spin generally slows the ball down and decreases its bounce height. Out of our 98 drop shots, 50 of them had top-spin and the other 48 had back-spin. The strike win rate of the 50 top-spin drop shots was 0.575, and the strike win rate of the 48 back-spin drop shots was 0.729. Based on our small sample, drop-shots are an effective shot, and even more effective when using back-spin.

Another unique strike type which will be analyzed will be volley shots. Volley shots are strikes which are struck before the ball is able to bounce, generally when the striker is close to the net. These kinds of strikes are considered an aggressive tennis strategy. When we only consider volleys, the 4,474 total strikes were cut down to 136. Of these 136 volleys, 52 of them were pure winners. This resulted in volleys having an impressive 0.667 strike win rate. Although it is a small sample, volleys produced a positive strike win rate, which matches expectations.

8 FUTURE WORK

The ESWR estimator and its methodology described in this thesis have demonstrated their potential usefulness when it comes to describing the evolution of an individual point and to finding weaknesses and/or strengths in a player's game. That does not mean the methodology cannot be improved. With more data comes more strikes and more returns, and their corresponding strike states and return states. This could enable us to create a more granular region system, and be even more specific as to which region the striker and the returner reside in. As was mentioned in Section 4.3, the ESWR estimator produces undesirable results if a player has had little to no past history from a certain strike state or return state. Future research could be made to group together similar players, and using those similar players, we could predict a player's probable tendencies from those strike or return states in question.

It is this research's belief that looking at player locations and past player tendencies is enough to accu-

[9]The greater than or equal to 0.5 meters is an attempt to remove any ground-strokes which hit the top of the net, have their trajectory disrupted, and bounce on the opposite side of the court. When they bounce they usually have very low x-coordinate magnitudes. Only these kinds of shots are usually able to obtain x-coordinates with magnitudes less than or equal to 0.5. These are not considered drop-shots, so we will attempt to exclude them from this analysis.

rately calculate ESWR and its related metrics, but it may be beneficial to further categorize shots with more characteristics in future work, such as type of spin and/or whether the shot was a backhand or forehand. Based on the preliminary analysis of drop shots and volleys in Section 7, some distinct strike types will produce high strike win rates, and this should be taken into consideration for future related works. It would also be interesting to see if some players are susceptible to these distinct strike types. This thesis primarily focused on ground-strokes, as all serves were grouped together in strike and return states, but more attention could be given to serves, and what types of serves each player tends to use (i.e. faster serves, serves landing on the edges of the serve box, serves with more spin, etc.). It would also be beneficial to work on the return win rate, and produce more reliable results than the ones found in Section 5.2, which were not as convincing at the strike win rate's results in Section 5.1. It was the intention of this research to keep the weights for both strikes and returns symmetric, but for future works, different weighting may be more appropriate.

Ultimately, more data and bigger sample sizes mean more possibilities for research in the under-researched field of tennis analytics. Hopefully, in the future, the USTA and other tennis organizations will be able to provide researchers with more data and more information to work with. The tennis analytics community is knowledgeable and hungry for more material to work with, as was evident by the numerous scholarly works made on the public 2012 Australian Open dataset. Putting more of this information-rich player-tracking data in the hands of the right people will indubitably come smarter player strategies and more insight for tennis fans everywhere.

9 ACKNOWLEDGEMENTS

Many people contributed to the completion of this thesis, but a select few deserve special recognition. Without your help, this work would not be what it is now. I would like to thank Dr. Matthew Hoffman and Dr. Ernest Fokoué of Rochester Institute of Technology for advising me and helping me along during the writing of this thesis. Also, I would like to thank Faizan Subhani, Jeffrey Zonenshine and the rest of the United States Tennis Association for their support, helpfulness and especially for the access to their US Open data.

APPENDIX

There are five main pieces of the computation of the ESWR estimator. In the following appendices we describe how these are coded. All code in this thesis was written in Python and run on a machine with a single 2.50GHz processor and 4GB of RAM.

APPENDIX A: PLAYER LOCATION TO REGION NUMBER

Data: Let $\vec{x} = [0, 4.9435, 9.887, 13.887, \infty]$ be the vector containing the positive x-coordinates of the vertical lines in the region system. Let $\vec{y}_{\text{general}} = [-\infty, -4.115, 0, 4.115, \infty]$ be the vector containing the y-coordinates of the major horizontal lines in the region system. Let $\vec{y}_{\text{baseline}} = [\infty, -4.115, -2.0575, 0, 2.0575, 4.115, \infty]$ be the vector containing the y-coordinates of the horizontal lines near the baseline area of the region system. Let $x \equiv$ x-coordinate of the player, $y \equiv$ y-coordinate of the player. Also, let

$$A = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ \text{N/A} & \text{N/A} & \text{N/A} & \text{N/A} \\ 15 & 16 & 17 & 18 \end{bmatrix}$$

where the "N/A" portions of A represent missing values of the third row, since it will not be used when we incorporate A.

Result: Takes the coordinates of the player in question and outputs the region he resides in.

if $x < 0$ **then**
  $x = -x$;
  $y = -y$;
end
$n_x = \texttt{sum}(x \geq \vec{x})$;
if $n_x = 3$ **then**
  $n_y = \texttt{length}(\vec{y}_{\text{baseline}}) - \texttt{sum}(y \geq \vec{y}_{\text{baseline}})$;
  Region $= 8 + n_y$;
else
  $n_y = \texttt{length}(\vec{y}_{\text{general}}) - \texttt{sum}(y \geq \vec{y}_{\text{general}})$;
  Region $= \text{A}_{n_x, n_y}$;
end

APPENDIX B: SORTING SHOTS INTO STRIKE SIGNIFICANCES

Data: Let n be the number of shots during a specific point, and let us consider the m^{th} shot in the sequence of shots during the point. Also, let x be the x-coordinate of the bounce location of the final shot of the point, and let y be the y-coordinate of this final bounce location.

Result: Takes the m^{th} shot with regards to the sequence of n shots during a point along with the bounce location of the final shot, and sorts the shot into a strike significance category.

```
if m = n then
    if No final bounce location then
        Strike Significance = "Net";
    else if (x,y) in-bounds then
        Strike Significance = "Pure Winner";
    else
        Strike Significance = "Out-of-Bounds";
    end
else if m = n-1 then
    if No final bounce location then
        Strike Significance = "Forced Net";
    else if (x,y) in-bounds then
        Strike Significance = "Set Up Of Opponent's Pure Winner";
    else
        Strike Significance = "Forced Out-of-Bounds";
    end
else if m = n-2 then
    if No final bounce location or (x,y) out-of-bounds then
        Strike Significance = "Non-Impactful";
    else if (x,y) in-bounds then
        Strike Significance = "Set Up Of Pure Winner";
else
    Strike Significance = "Non-Impactful";
end
```

APPENDIX C: SORTING SHOTS INTO RETURN SIGNIFICANCES

Data: Let n be the number of shots during a specific point, and let us consider the m^{th} shot in the sequence of shots during the point. Also, let λ be the strike significance of the n^{th}, final shot during the point.

Result: Takes the m^{th} shot with regards to the sequence of n shots during a point along with the strike significance of the final shot, and sorts the shot into a return significance category.

if $m = n$ **then**
  if λ = *Pure Winner* **then**
    Return Significance = "No Return";
  else if λ = *Out-of-Bounds or* λ = *Net* **then**
    Return Significance = "Net";
else if m = *n-1* **then**
  if λ = *Pure Winner* **then**
    Return Significance = "Returned Pure Winner";
  else if λ = *Out-of-Bounds or* λ = *Net* **then**
    Return Significance = "Returned Losing Strike";
else if m = *n-2* **then**
  if λ = *Pure Winner* **then**
    Return Significance = "Returned Set Up of Opponent's Pure Winner";
  else if λ = *Out-of-Bounds or* λ = *Net* **then**
    Return Significance = "Returned Forced Losing Strike";
else if m = *n-3* **then**
  if λ = *Pure Winner* **then**
    Return Significance = "Returned Set Up of Pure Winner";
  else
    Return Significance = "Returned";
  end
else
  Return Significance = "Returned";
end

APPENDIX D: FINDING THE NEXT STATE

Data: Let n be the number of shots during a specific point, and let us consider the m^{th} shot in the sequence of shots during the point. Let c_m be the strike state of the m^{th} shot during the point. Also, let PLC_m be the player location combination at the time of m^{th} shot of the point and let λ_m be the strike significance of the m^{th} shot of the point.

Result: Takes the m^{th} shot of a sequence of shots during a point with its strike state c_m, and outputs the next state of the point, c_{next}. This will be used during the formation of the transition probability matrices.

if $m < n$ **then**
  if c_m = *1S* **then**
    $c_{\text{next}} = 1\text{SR}$;
  else if c_m = *2S* **then**
    $c_{\text{next}} = 2\text{SR}$;
  else
    $c_{\text{next}} = \text{PLC}_{m+1}$;
  end
else
  if λ_m = *Pure Winner* **then**
    c_{next} = No Return;
  else if c_m = *1S and* (λ_m = *Out-of-Bounds or* λ_m = *Net)* **then**
    $c_{\text{next}} = 2\text{S}$;
  else if λ_m = *Out-of-Bounds* **then**
    c_{next} = Out-of-Bounds;
  else
    c_{next} = Net;
  end
end

APPENDIX E: FORMATION OF THE TRANSITION PROBABILITY MATRICES

Data: Let us form the transition probability matrix for player i. Recall $\mathcal{C}_{\text{Shot}}$ contains all the location categorizations from the states $\mathcal{C}_{\text{ServerShot}}$ and $\mathcal{C}_{\text{ReceiverShot}}$, and recall $\mathcal{C}$ contains all the possible states. Let $S^i_{\beta_1^S}$ contain all strikes struck by player i from strike state β_1^S, and let $S^i_{\beta_1^S \to \beta_2^S}$ contain all strikes by player i from strike state β_1^S where $c_{\text{next}} = \beta_2^S$.

Result: Takes a specific player i and outputs his personal transition probability matrix, $\mathbf{P}^i$, whose row labels and column labels are elements of $\mathcal{C}$.

$\mathbf{P}^i = 0_{328,328}$;
for $c_1 \in \mathcal{C}_{Shot}$ **do**
  for $c_2 \in \mathcal{C}$ **do**
    $n_1 = \texttt{count}(S^i_{c_1 \to c_2})$;
    $n_2 = \texttt{count}(S^i_{c_1})$;
    if $n_2 = 0$ **then**
      $P^i_{c_1,c_2} = 0$;
    else
      $P^i_{c_1,c_2} = n_1/n_2$;
    end
  end
end
$P^i_{\text{No Return,No Return}} = 1$;
$P^i_{\text{Out-of-Bounds,Out-of-Bounds}} = 1$;
$P^i_{\text{Net,Net}} = 1$;

REFERENCES

[1] A. Bialkowski, P. Lucey, P. Carr, Y. Yue, S. Sridharan, and I. Matthews, *Large-scale analysis of soccer matches using spatiotemporal tracking data*, in Data Mining (ICDM), 2014 IEEE International Conference on, IEEE, 2014, pp. 725–730.

[2] D. R. Brillinger et al., *A potential function approach to the flow of play in soccer*, Journal of Quantitative Analysis in Sports, 3 (2007), p. 3.

[3] B. Bukiet, E. R. Harold, and J. L. Palacios, *A markov chain approach to baseball*, Operations Research, 45 (1997), pp. 14–23.

[4] D. Cervone, L. Bornn, and K. Goldsberry, *Nba court realty*, 2016.

[5] D. Cervone, A. D'Amour, L. Bornn, and K. Goldsberry, *A multiresolution stochastic process model for predicting basketball possession outcomes*, Journal of the American Statistical Association, 111 (2016), pp. 585–599.

[6] A. Franks, A. Miller, L. Bornn, and K. Goldsberry, *Counterpoints: Advanced defensive metrics for nba basketball*, in 9th Annual MIT Sloan Sports Analytics Conference, Boston, MA, 2015.

[7] K. Goldner et al., *A markov model of football: Using stochastic processes to model a football drive*, Journal of Quantitative Analysis in Sports, 8 (2012), p. 1.

[8] K. Goldsberry, *Courtvision: New visual and spatial analytics for the nba*, in 2012 MIT Sloan Sports Analytics Conference, 2012.

[9] K. Goldsberry and E. Weiss, *The dwight effect: A new ensemble of interior defense analytics for the nba*, Sports Aptitude, LLC. Web, (2013).

[10] M. Horton, J. Gudmundsson, S. Chawla, and J. Estephan, *Classification of passes in football matches using spatiotemporal data*, arXiv preprint arXiv:1407.5093, (2014).

[11] J. G. Kemeny, J. L. Snell, et al., *Finite markov chains*, vol. 356, van Nostrand Princeton, NJ, 1960.

[12] M. Lewis, *Moneyball: The art of winning an unfair game*, WW Norton & Company, 2004.

[13] A. Loukianov and V. Ejov, *Markov modeling for a tennis point played*, International Journal of Sports Science and Sports Engineering.

[14] P. Lucey, A. Bialkowski, P. Carr, Y. Yue, and I. Matthews, *How to get an open shot: analyzing team movement in basketball using tracking data*, MIT SSAC, (2014).

[15] P. Lucey, A. Bialkowski, M. Monfort, P. Carr, and I. Matthews, *quality vs quantity: Improved shot prediction in soccer using strategic features from spatiotemporal data*, in Proc. 8th Annual MIT Sloan Sports Analytics Conference, 2014, pp. 1–9.

[16] P. Lucey, D. Oliver, P. Carr, J. Roth, and I. Matthews, *Assessing team strategy using spatiotemporal data*, in Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2013, pp. 1366–1374.

[17] A. Miller, L. Bornn, R. P. Adams, and K. Goldsberry, *Factorized point process intensities: A spatial analysis of professional basketball.*, in ICML, 2014, pp. 235–243.

[18] S. V. Mora and W. J. Knottenbelt, *Spatio-temporal analysis of tennis matches*.

[19] M. Reid and R. Duffield, *The development of fatigue during match-play tennis*, British journal of sports medicine, 48 (2014), pp. i7–i11.

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

[21] 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, 2013.