Augmenting Tennis Point Stochastic Modeling Utilizing Spatiotemporal Shot Data

Corey Emery

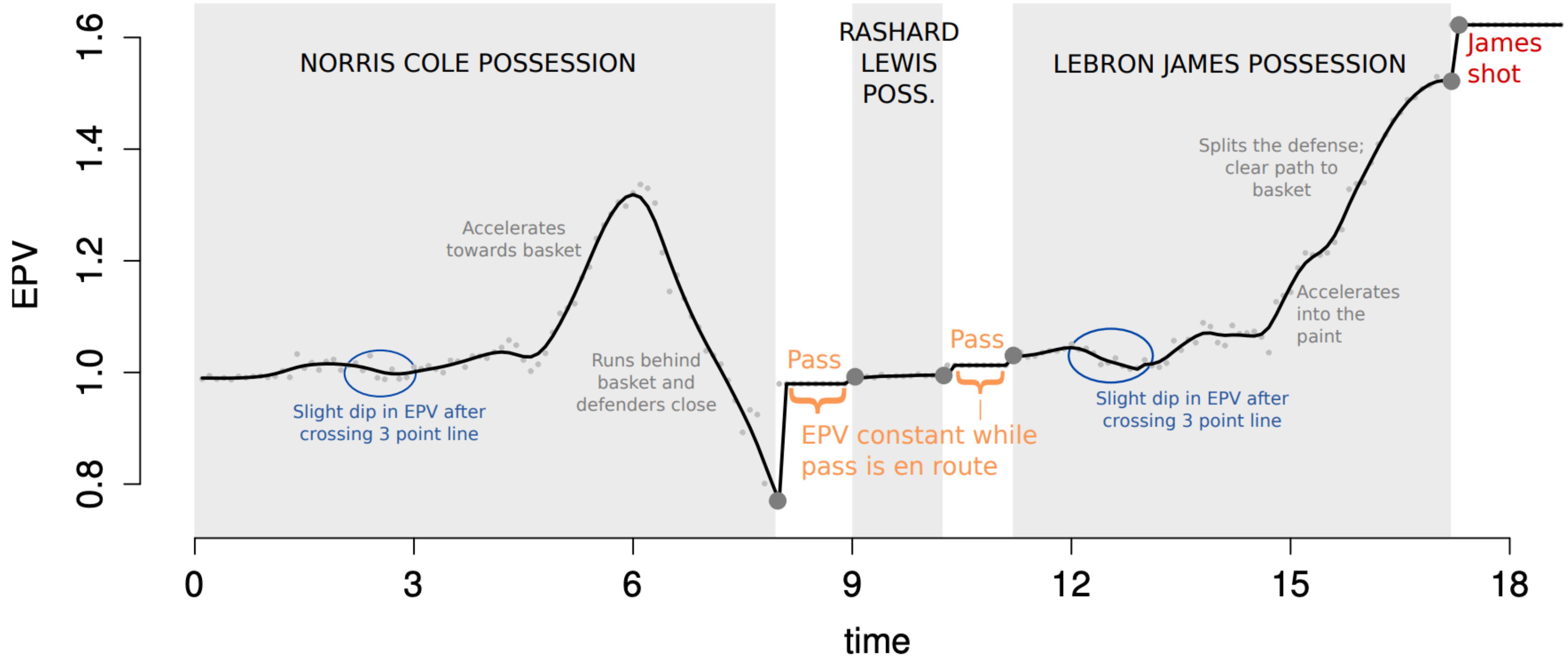
Statistics and Data Science Research Symposium

May 6th, 2020

Roadmap

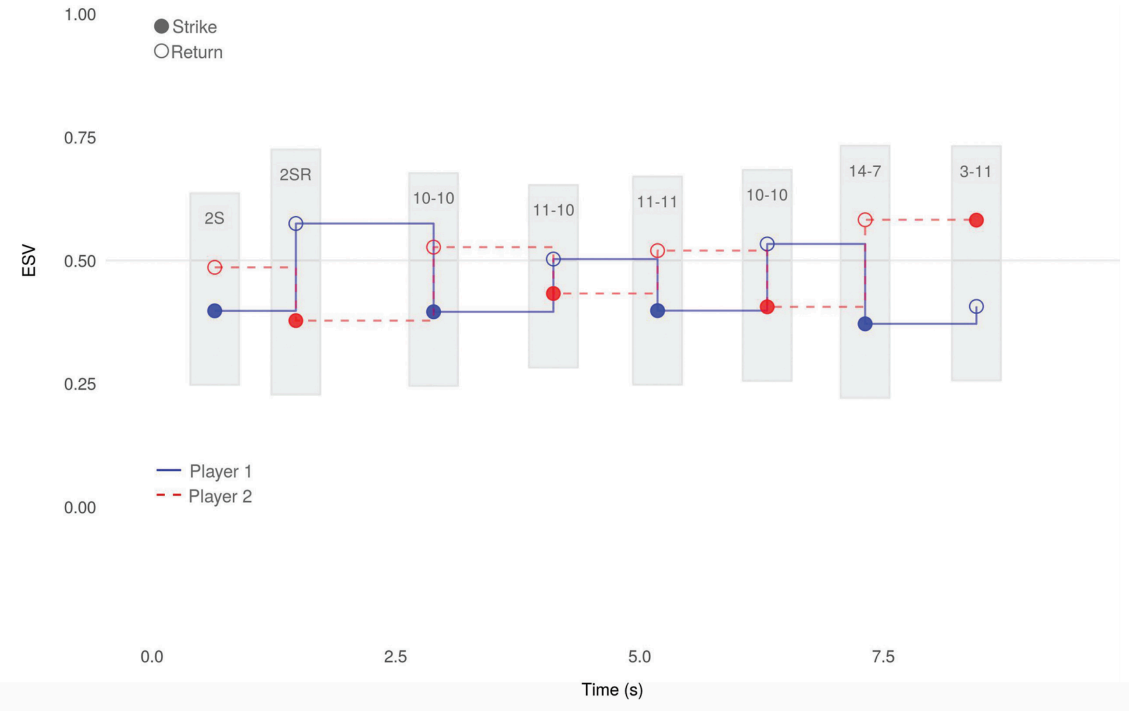
- Background & Inspiration
- Impact & Extensions
- Methodology
- Results
- Future Work

Inspiration

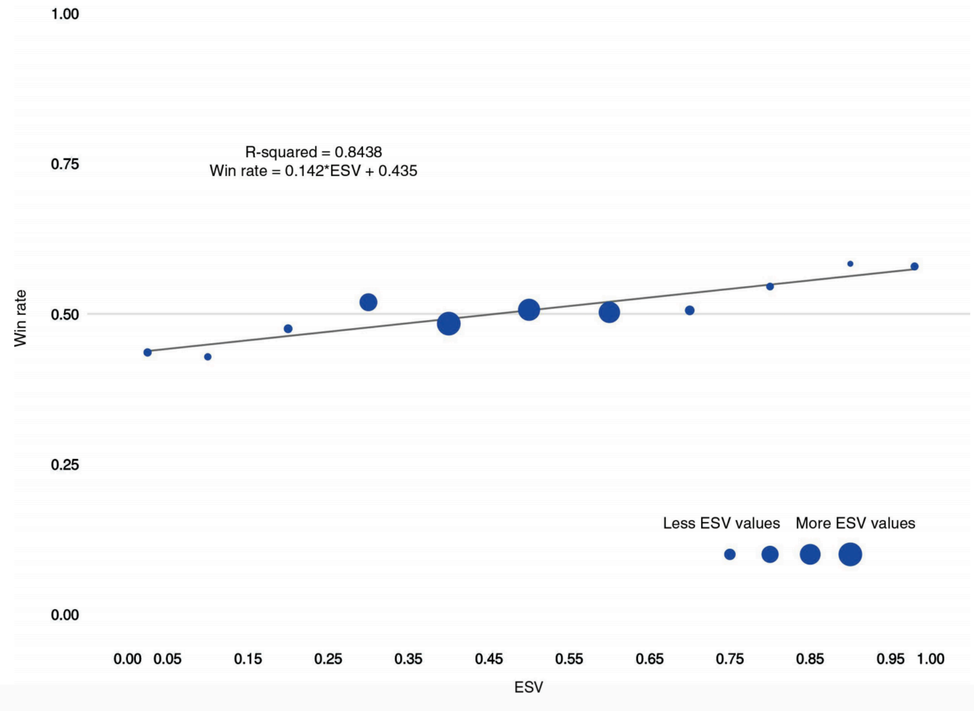


Expected Point Value (EPV). *Cervone et al. (2016)*

Inspiration



Expected Shot Value (ESV). *Floyd et al. (2019)*



ESV Validation. *Floyd et al. (2019)*

What is Expected Shot Value

Expected Shot Value is:

- "How many points a player can expect to score [in a given point] based on the circumstances at the time of a given shot"
- Based on the idea that the relative positioning of players on a tennis court affects the outcome of a point
- Derived from a stochastic model (allows for variation moving forward in time)

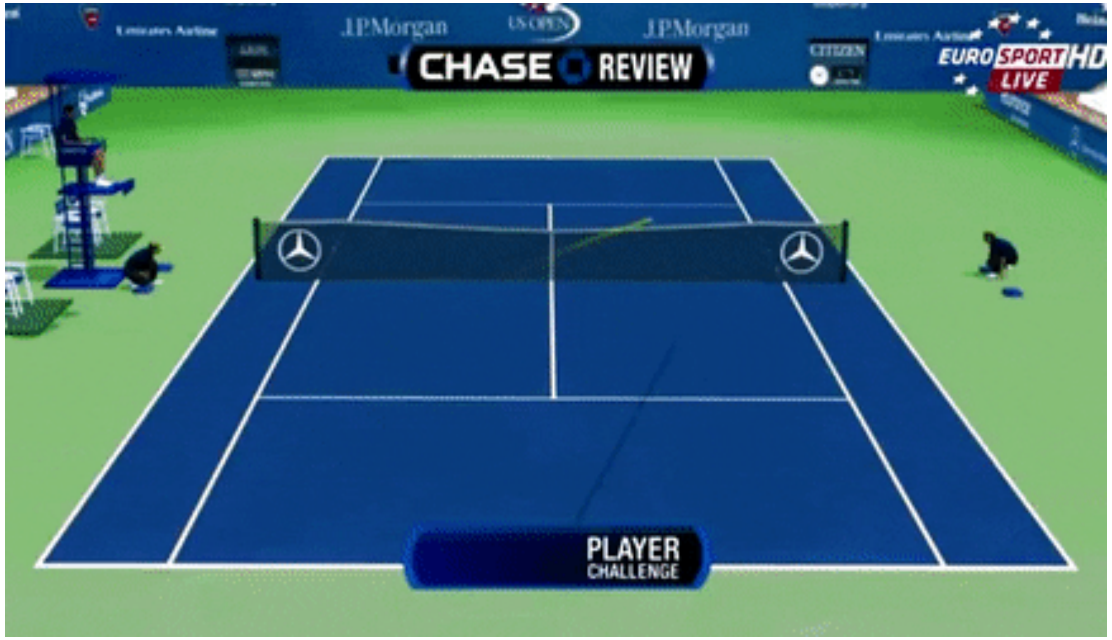
Expected Shot Value is NOT:

- The win-loss probability for both players in a point
  - ESVs are independent between players, so the striker and returner ESVs are not expected to sum to 1 for an individual shot
  - Past information the model is built on is incomplete

Hawk-Eye Impact

Current tennis analytics is primarily high-level statistics (serve, error, and winner percentages, points won, etc.). These ignore a large amount of the information we can learn from a match, including player movement, exertion, and patterns.

Hawk-Eye recovers a lot of lost information. We can use player tracking data to determine the relative position of players on a court (per Floyd's original methodology). We can also use information about speed, spin, and length of match to augment this methodology and explore variations of this model.



Our Goal

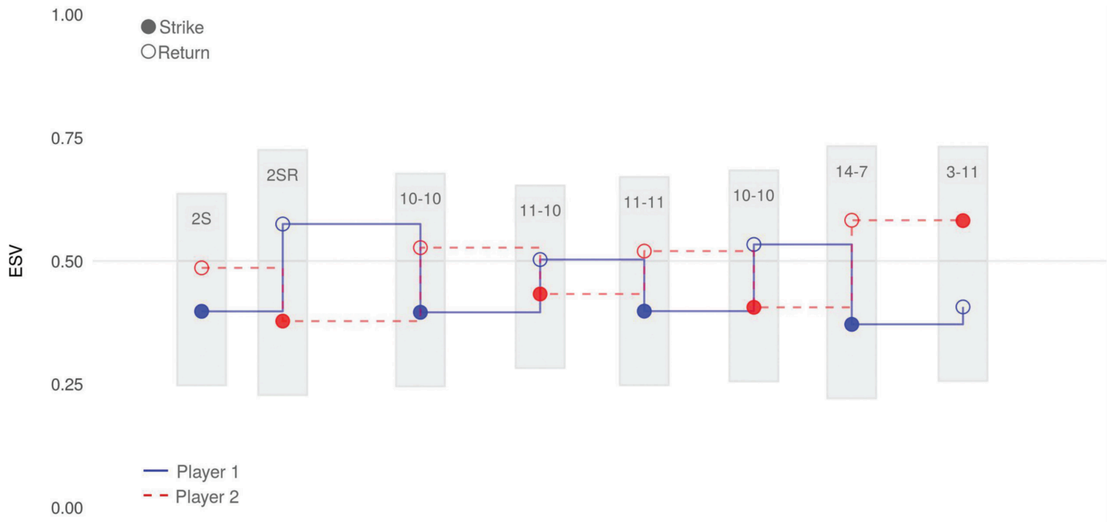
We want to extend the Floyd et al. model to better incorporate the depth of the data that Hawk-Eye offers

We hope to more accurately model the underlying mechanics of and strategy of tennis

Why Do We Care

Understand the Progression of a Tennis Point

See at what shots a player might have had an advantage, missed an opportunity, or mounted a comeback

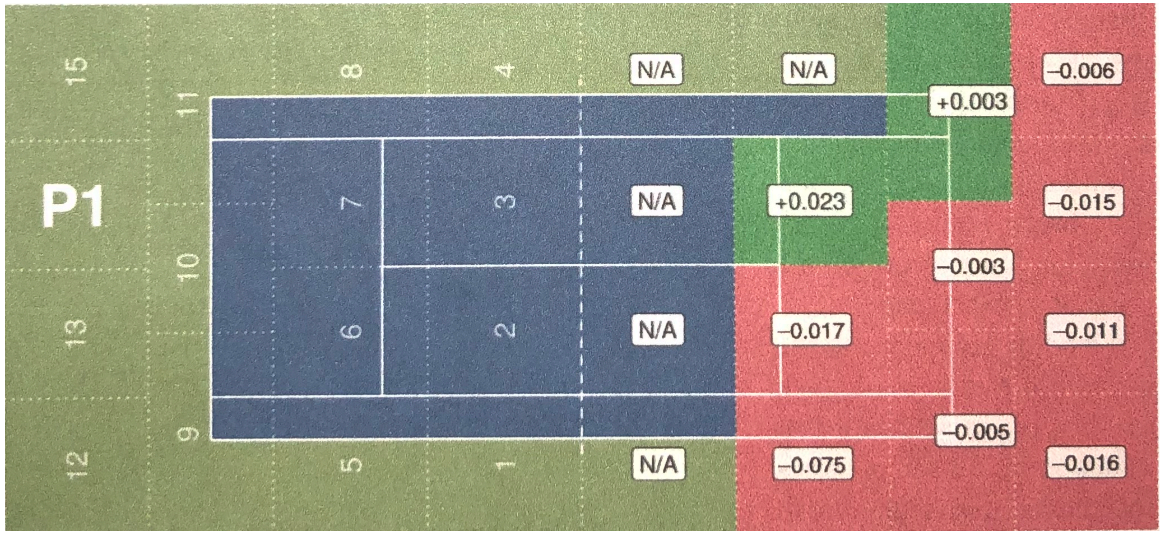


Expected Shot Value over the course of a tennis point.
Floyd et al. (2019)

Why Do We Care

Identify Player Strengths and Weaknesses

Where does a player have the highest success rate, highest failure rate? How do they compare to an "average" player?

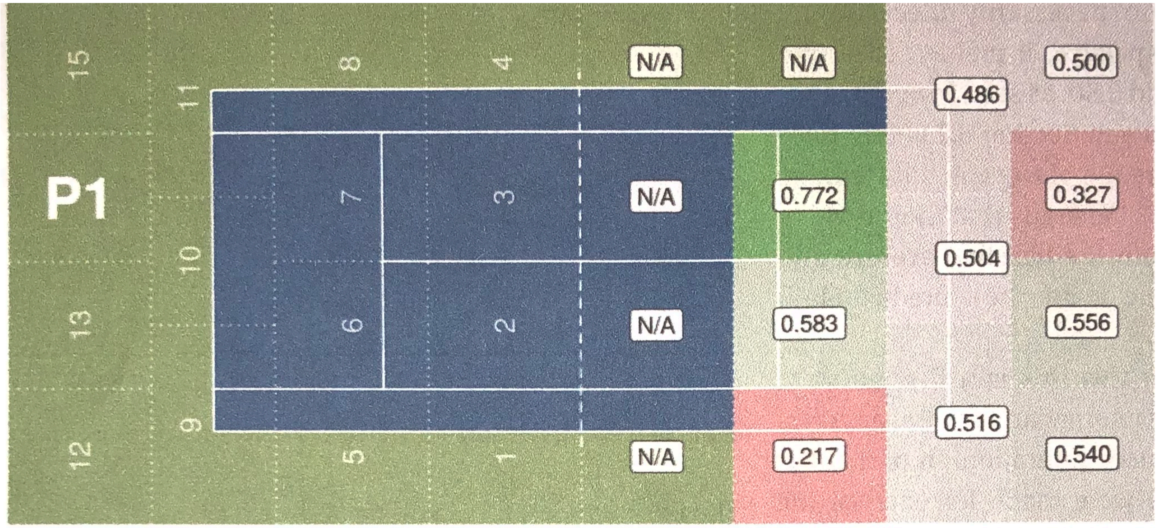


Expected Shot Value difference between a player returning a shot and the dataset average. *Floyd et al. (2019)*

Why Do We Care

Guide In-Match Strategy

Given the player locations, where should a player hit their next shot, where should a player move next? What shots are most effective in these instances?



Return Expected Shot Value for all zones equidistant to the player's current striking position. *Floyd et al. (2019)*

Opportunities for Extension

Features to Incorporate

Shot Type

Shot Speed

Shot Spin

Concepts to Explore

Fatigue

Distance and Direction Ran

Momentum

Terminology

Shot: an individual state where one player strikes the ball to their opponent

Point: a series of shots, ends when a player cannot return the ball or the ball goes out of bounds

Server: the player who hits the first shot of the point

Receiver: the player who receives the first shot of the point

Striker: the player hitting the current shot in a point (not always the same as the server)

Returner: the player receiving the current shot in a point (not always the same as the receiver)

Data

USTA Data Source

Men's and Women's singles match data from the US Open between 2015 and 2018

PRJ: player positions (25 fps)

TRJ: ball location and trajectory

XML: shot characteristics

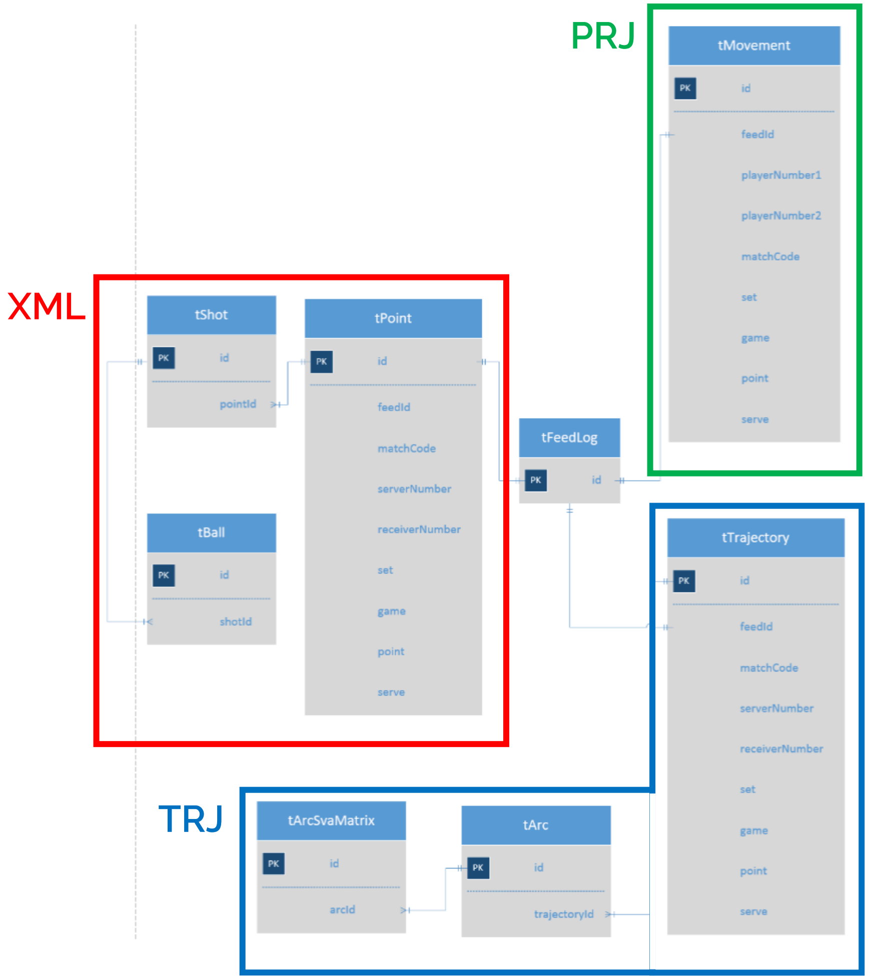
Data Post-Filtering

701 unique matches

366 players, 690 unique matchups

144,695 points (80% of original data)

473,401 shots (73% of original data)

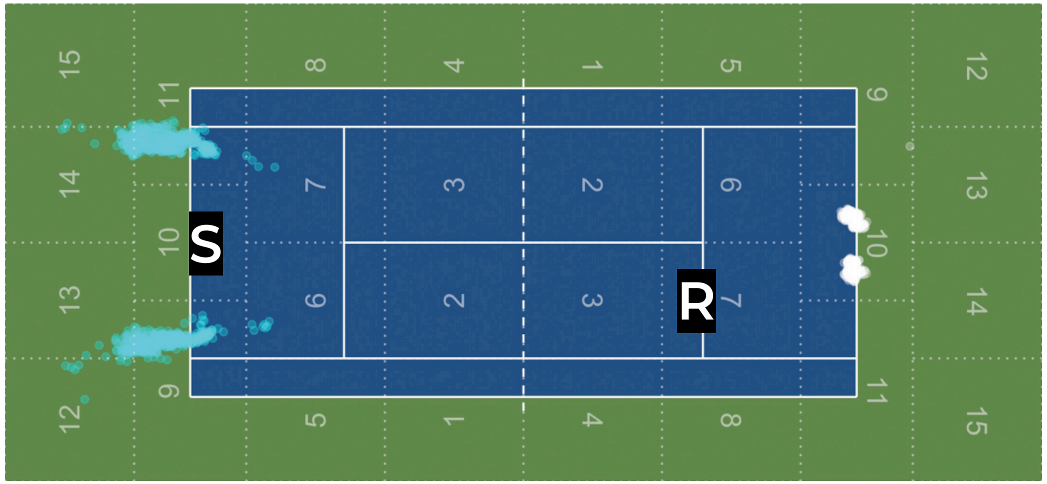


Mathematical Foundation

1. Assigning Strike/Return States

In order to create a Markov Chain model and reduce computational complexity, the tennis court is divided into zones, and states are created from player locations at each shot (usually *striker_zone - returner_zone*)

i.e. the state of this shot would be 10-7



Mathematical Foundation

2. Assigning Strike/Return Categories

This is a description of the shot's outcome/effect on the point, associated with a numeric weight representing expected benefit

Strike Category	Symbol (λ_{SC})	Weight (w_{SC})	Definition
Pure Winner	λ_{PW}	1.00	In-bounds strike which is not returned
Set-Up of Pure Winner	λ_{SUPW}	0.75	Strike by winning player prior to λ_{PW}
Forced Losing Strike	λ_{FLS}	0.75	Strike by winning player prior to λ_{LS}
Non-Impactful	λ_{NI}	0.50	Strike which does not fall into other categorizations
Set-Up of Opponent's Pure Winner	λ_{SUOPW}	0.25	Strike by losing player prior to λ_{PW}
Losing Strike	λ_{LS}	0.00	Strike which goes out-of-bounds or into the net

Table 1: Original Floyd et al. Strike Categories and Weights

Return Category	Symbol (λ_{RC})	Weight (w_{RC})	Definition
Returned Pure Winner	λ_{RPW}	1.00	Returned λ_{PW}
Returned Set-Up of Pure Winner	λ_{RSUPW}	0.75	Returned λ_{SUPW}
Returned Forced Losing Strike	λ_{RFLS}	0.75	Returned λ_{FLS}
Returned Non-Impactful Strike	λ_{R}	0.50	Returned λ_{NI}
Returned Set-Up of Opponent's Pure Winner	λ_{RSUOPW}	0.25	Returned λ_{SUOPW}
Returned Losing Strike	λ_{RLS}	0.00	Returned λ_{LS}
Not Returned	λ_{NR}	0.00	Unable to return

Table 2: Original Floyd et al. Return Categories and Weights

Mathematical Foundation

3. Calculating Transition Probabilities

A matrix that describes the probability of going from the current strike state *u* to the strike/return state *w* on the next shot

$$P_{\mathtt{uw}} = \frac{N_{\mathtt{uw}}}{\sum_{\mathtt{w}'} N_{\mathtt{uw}'}}$$

Mathematical Foundation

4. Calculating Strike/Return Values

These variables quantify the expected number of points a player will receive based on how they've struck/returned the ball previous

$$\mathrm{SV}(\beta^{\mathrm{S}}) = \sum_{\lambda_{\mathrm{SC}} \in \Lambda_{\mathrm{SC}}} w_{\lambda_{\mathrm{SC}}} \frac{|\mathrm{S}_{\beta^{\mathrm{S}},\lambda_{\mathrm{SC}}}|}{|\mathrm{S}_{\beta^{\mathrm{S}}}|}$$

$$\mathrm{RV}(\beta^{\mathrm{S}}) = \sum_{\lambda_{\mathrm{RC}} \in \Lambda_{\mathrm{RC}}} w_{\lambda_{\mathrm{RC}}} \frac{|\mathrm{R}_{\beta^{\mathrm{R}},\lambda_{\mathrm{RC}}}|}{|\mathrm{R}_{\beta^{\mathrm{R}}}|}$$

Mathematical Foundation

5. Putting it Together

Player 2 returning from state 14-7

$^2/_9$ of Player 1's strikes from 14-7 were direct losing strikes, $^1/_9$ were direct winners.

$^6/_9$ strikes continued the point: 1 went to 2-14 (Player 2 SV = 0.750), 2 to 6-11 (SV = 0.438), etc.

$$ESV =$$

$$\frac{2}{9}(1) + \frac{1}{9}(0) + \frac{6}{9}(0.750 + 2 * 0.438 + \ \dots)$$

$$\mathtt{ESV}(\mathtt{c_u}) =$$

$$\sum_{\mathtt{c_w} \in \mathtt{C_{shot}}} \mathtt{RV}(\mathtt{c_w}) P_{\mathtt{uw}} + \sum_{\mathtt{c_w} \in \mathtt{C_{end}}} \mathtt{PV_S}(\mathtt{c_w}) P_{\mathtt{uw}}$$

if the player is the striker

$$\sum_{\mathtt{c_w} \in \mathtt{C_{shot}}} \mathtt{SV}(\mathtt{c_w}) P_{\mathtt{uw}} + \sum_{\mathtt{c_w} \in \mathtt{C_{end}}} \mathtt{PV_R}(\mathtt{c_w}) P_{\mathtt{uw}}$$

if the player is the returner

Our Changes

2. Assigning Strike/Return Categories

This is a description of the shot's outcome/effect on the point, associated with a numeric weight representing expected benefit

By adjusting these categories and their respective weights, we are able to incorporate new features and assign different values to each type of shot

Our Models

Models we Explored

Shot Spin

Shot Speed

Serve Speed

Shot Type and Spin (top right)

Estimated Neutral Shot Weight (middle right)

Speed and Spin (bottom right)

General Setup

Define additional shot categories

Weight beneficial shots as 0.05 higher than the baseline and detrimental shots as 0.05 lower

	λ_{SUPW}/ λ_{RSUPW}	λ_{FLS}/ λ_{RFLS}	λ_{NI}/ λ_{R}	λ_{SUOPW}/ λ_{RSUOPW}
Topspin Shot (-T)	0.8	0.8	0.55	0.3
Forehand Backspin Shot (-C)	0.8	0.8	0.55	0.3
Backhand Backspin Shot (-S)	0.7	0.7	0.45	0.2
Other Shot (-D)	0.75	0.75	0.5	0.25

λ_{SUPW}	λ_{FLS}	λ_{NI}	λ_{SUOPW}
0.75	0.75	0.4333	0.25

λ_{RSUPW}	λ_{RFLS}	λ_{R}	λ_{RSUOPW}
0.75	0.75	0.4262	0.25

	λ_{SUPW}/ λ_{RSUPW}	λ_{FLS}/ λ_{RFLS}	λ_{NI}/ λ_{R}	λ_{SUOPW}/ λ_{RSUOPW}
First Serve, $>$ 110 MPH (-1SF)	0.8	0.8	0.55	0.3
First Serve, $<$ 110 MPH (-1SS)	0.7	0.7	0.45	0.2
Second Serve, $>$ 90 MPH (-2SF)	0.8	0.8	0.55	0.3
Second Serve, $<$ 90 MPH (-2SS)	0.7	0.7	0.45	0.2
Topspin Shot, $>$ 85 MPH (-TF)	0.8	0.8	0.55	0.3
Backspin Shot, $<$ 50 MPH (-BS)	0.7	0.7	0.45	0.2
Other Shot (-A)	0.75	0.75	0.5	0.25

Validation

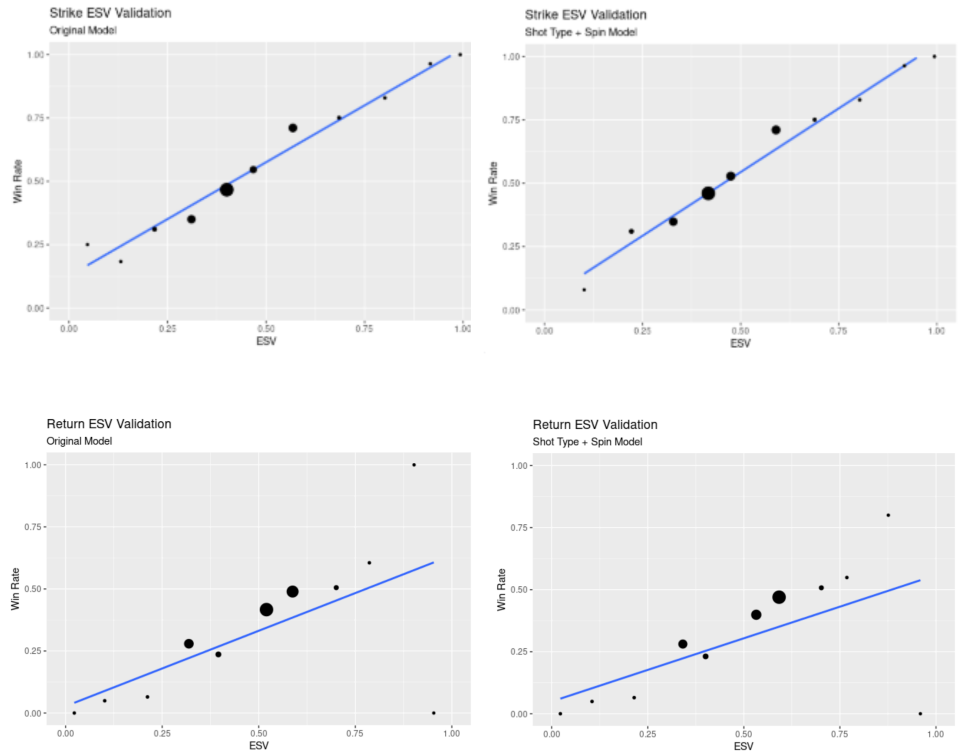
Ensuring a Linear Relationship

Floyd et al. graph the relationship between ESV and Win Rate to determine the validity of their models

We create these plots by pooling shots based on ESV, and calculating win rate as the percentage of shots the striker won

All our plots show a linear relationship, which tells us that these models do not warp the underlying point mechanics

This also uncovers bins that are consistently above and below the trend line, and thus potential structures we are not capturing



Performance

Correlation between Strike ESV and Win Rate

Due to a high baseline R-squared, we did not see a drastic improvement with any of our models

Models that incorporate spin appear to have a higher correlation

Most improvements were due to outlier shots (very high or low ESV) changing bins

Moving forward can attempt a weighted R-square measure to reduce outlier effect

Model	Strike ESV R^2
Shot Type + Spin	0.9820
Spin + Dropshot	0.9815
Spin	0.9814
Estimated Neutral Shot Weight	0.9762
Floyd et al. (Baseline)	0.9758
Serve Speed	0.9758
Full Speed	0.9756
Full Speed + Spin	0.9756
Non-Serve Speed + Spin	0.9755

Examining Residuals

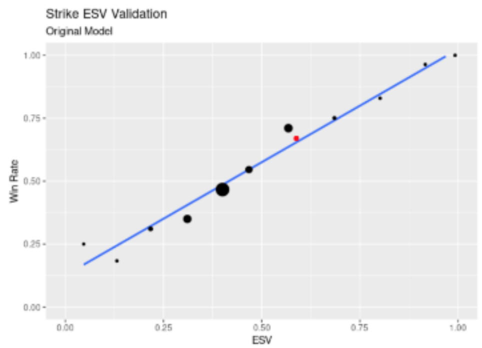
ESV Bin [0.55, 0.65]

The data point for this bin consistently fell above the trend line for all our graphs

This bin includes all the first serves in our data, and is dominated by this shot (of the 110,692 shots, only 6,636 are not serves)

Removing first serves raises ESV from 0.5670 to 0.5881 and drop win rate from 0.7106 to 0.6698, which is in line with the linear model (top figure, red point)

This suggests there is a more complex underlying structure around serves



Location State	Strike ESV	Number of Shots
1S	0.5657	104056
2-5	0.5958	3
2-9	0.5719	102
2-11	0.6141	293
2-13	0.6339	150
3-2	0.5730	84
3-4	0.6375	5
3-9	0.6144	202
3-11	0.5888	240

Analyzing Physical Factors

Implicit Physical Factors

Due to the zoning system used in our model, aspects such as distance and direction ran by a player are already included

We can isolate certain types of movement, then examine the distribution of shot category and mean weight to assess its benefit/detriment

We can deepen this analysis by looking at mean ESV rather than mean shot weight

Movement and Zones	PW	SUPW	FLS	NI	SUOPW	LS	Mean Weight
Baseline to Net: $(9-11) \to (1-8)$	2072	489	900	785	592	1045	0.6212
Backcourt to Net: $(12-15) \to (1-8)$	335	94	204	226	246	329	0.5112
Retreat: $(1-4) \to (9-15)$	13	7	19	85	36	77	0.3544
Cross Court: $9 \to 11$ or $11 \to 9$ or $12 \to 15$ or $15 \to 12$	188	92	280	896	177	488	0.4523

Future Opportunities

Exploring Fatigue and Momentum

Hawk-Eye provides information about match length, shots per point, and how far a player runs over the course of a match

We can also attempt to model streaks of points won or shots hit with an above-average weight to capture momentum

Estimating Category Weights

Much like Floyd et al., we assign category weights subjectively based on our knowledge and assumptions of the sport

Interesting findings could lie in the differences between estimated weights

Thank You

Questions?