Opening up the court (surface) in tennis grand slams

Kayla Frisoli, Shannon Gallagher, and Amanda Luby
Department of Statistics & Data Science
Carnegie Mellon University

CMSAC -- October 20, 2018

Tennis, anyone?

Roger Federer
@ **WIMBLEDON**

GOAT?

20 grand slam titles

Grass extraordinaire
(8 GS @ Wim.)

Tennis, anyone?

Rafael Nadal
@ **FRENCH OPEN**

GOAT?

17 grand slam titles

King of Clay
(11 GS @ FO)

Tennis, anyone?

Serena Williams
@ **US OPEN**
@ **AUSTRALIAN OPEN**

GOAT?

24 grand slam titles

Jack of all trades
(7 GS @ USO)
(6 GS @ AO)

The grand slams are played on distinct surfaces and may affect player performance.

Grand Slam	Surface	Known top players
AUSTRALIAN OPEN	DecoTurf (hard court)	Serena Williams
FRENCH OPEN	clay	Rafael Nadal
WIMBLEDON	grass	Roger Federer
US OPEN	Plexicushion (hard court)	Serena Williams

The grand slams are played on distinct surfaces and may affect player performance.

Grand Slam	Surface	Known top players
AUSTRALIAN OPEN	DecoTurf (hard court)	Serena Williams
FRENCH OPEN	clay	Rafael Nadal
WIMBLEDON	grass	Roger Federerer
US OPEN	Plexicushion (hard court)	Serena Williams

Are these real or perceived effects?
Do these effects vary by player?

Two data sets- two perspectives

- Data from Jeff Sackman's website (https://github.com/JeffSackmann)
- Accessed via the R `deuce` package (Kovalchik, S 2017)

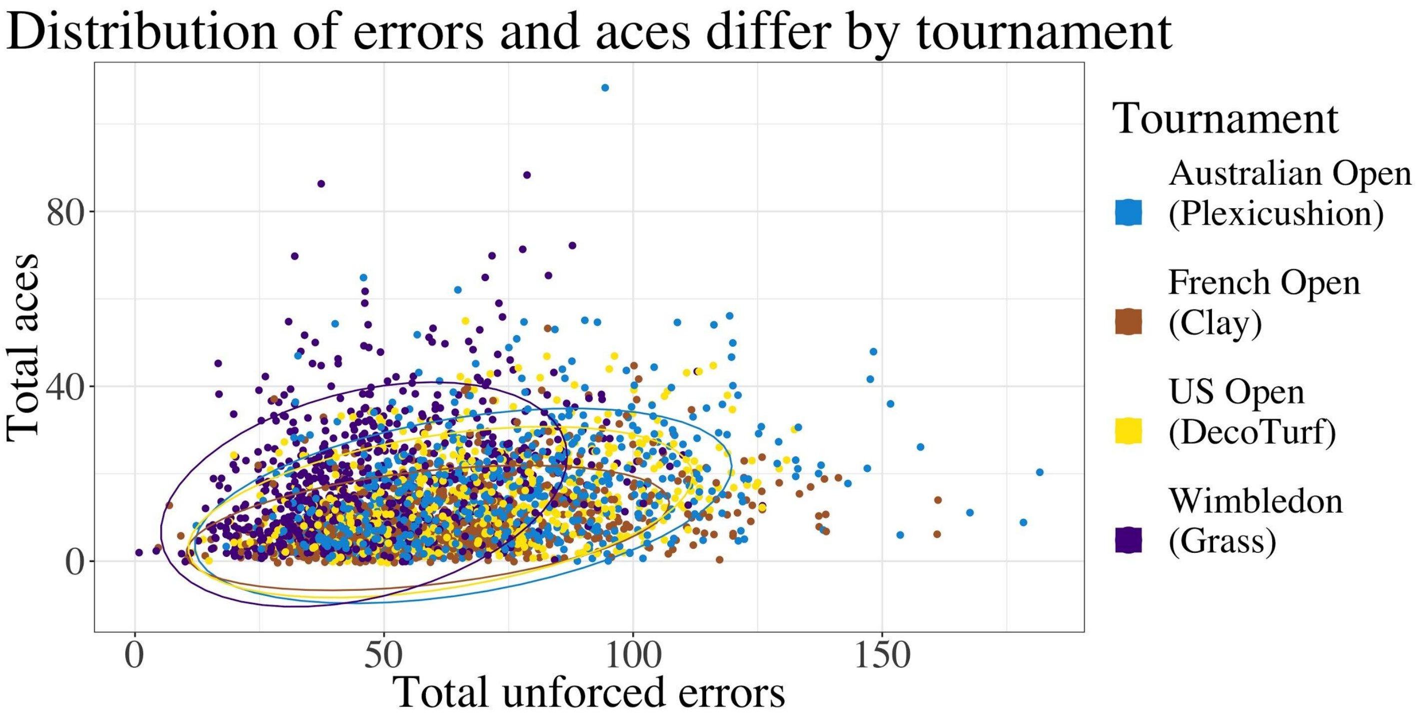
Grand Slam Results (GS)

- 2013-2017
- One row = one match
- 5080 matches
- 4 GS, 7 rounds each
- Match and game scores

Grand Slam Point by Point (PbP)

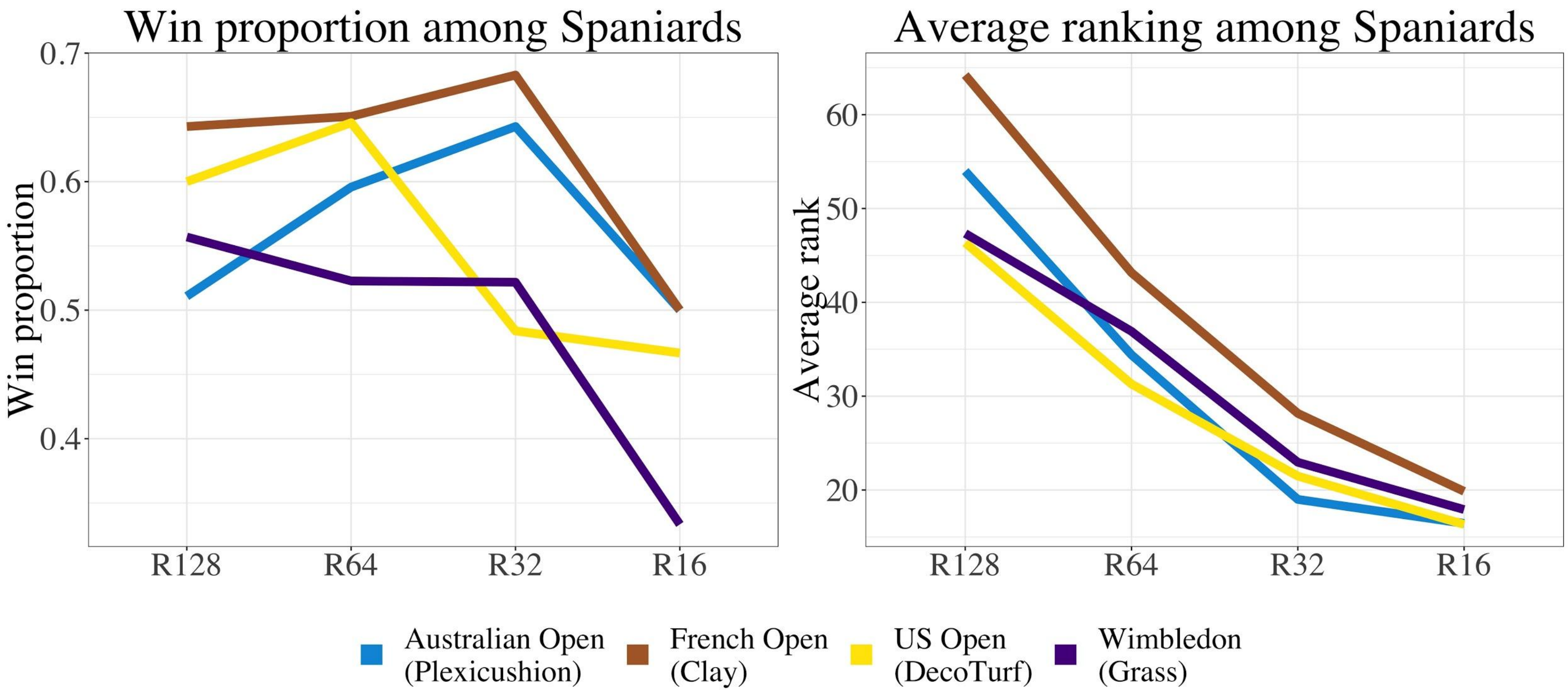
- 2013-2017
- One row = one point
- 720,465 points from 3066 matches
- Missingness
- Additional variables: winners, aces, unforced errors (UEs), etc.

Exploring tournament differences



Players perform differently at Wimbledon, as displayed by the clustering of purple points.

Spaniards outperform on clay surface



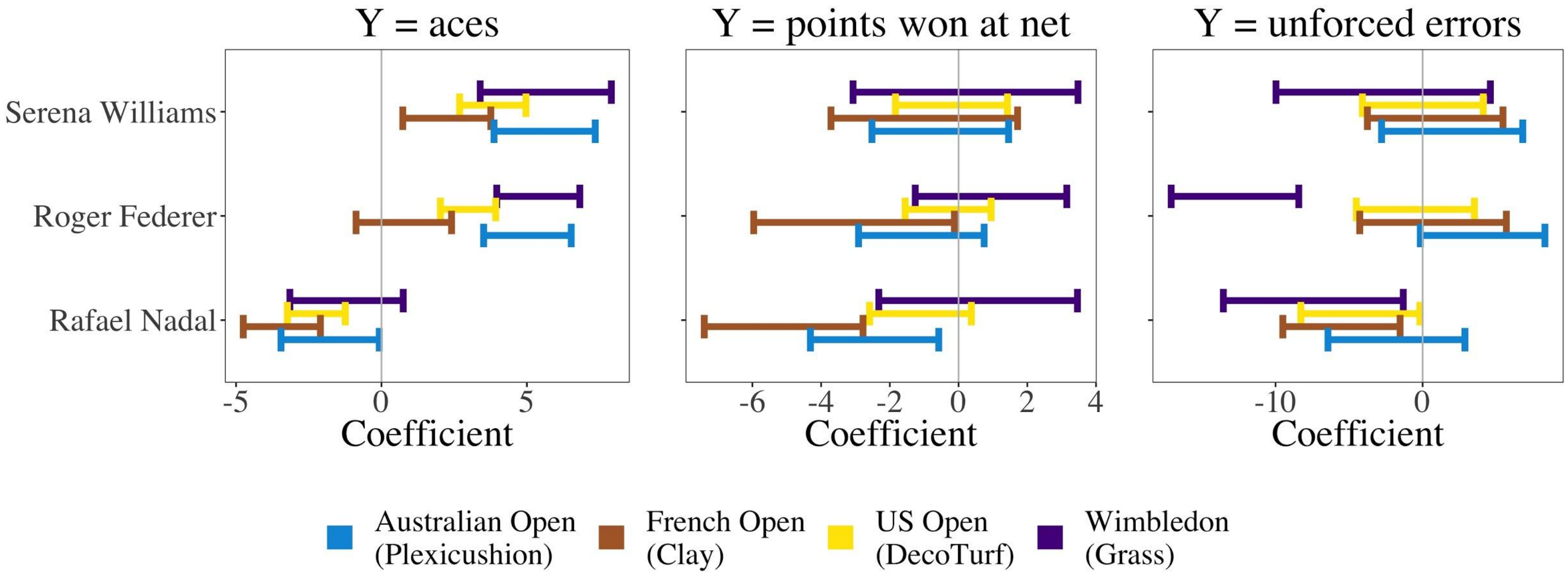
Spanish players win more at the French Open, despite their, on average, worse rankings.

We build a series of models to assess the match effects of court surface and individual players

	Data	Y	Fixed X	Random X	Regression	Conclusions
Approach 1	GS Data *n=10,160*	Did win? (Yes = 1, No = 0)	league, country, year, late round op. rank, rank	surface	Logistic	No significant effects besides rank / opponent rank

	Data	Y	Fixed X	Random X	Regression	Conclusions
Approach 1	GS Data *n=10,160*	Did win? (Yes = 1, No = 0)	league, country, year, late round op. rank, rank	surface	Logistic	No significant effects besides rank / opponent rank
Approach 2	GS Data %>% join(PBP) *n=6,132* (aces) *n=6,132* (net) *n=6,132* (UE)	Aces Points won at net UE	league, country, year, late round op. rank, rank	surface	Linear	Significant surface effects for Williams, Federer, Nadal

Player effects vary by court surface



Williams and Federer have more **aces** in general, and most on grass and hard court

Federer makes far fewer unforced errors on **grass** compared to others and himself

	Data	Y	Fixed X	Random X	Regression	Conclusions
Approach 1	GS Data *n=10,160*	Did win? (Yes = 1, No = 0)	league, country, year, late round op. rank, rank	surface	Logistic	No significant effects besides rank / opponent rank
Approach 2	GS Data %>% join(PBP) *n=6,132* (aces) *n=6,132* (net) *n=6,132* (UE)	Aces Points won at net UE	league, country, year, late round op. rank, rank	surface	Linear	Significant surface effects for Williams, Federer, Nadal
Approach 3	GS Data %>% join(PBP) %>% filter(player == “{Player}”) *n=75* (Nadal) *n=83* (Federer) *n=59* (Williams)	% points won	league, country, year, late round op. rank, rank average service speed, winners, unforced errors, break points won, net points won, etc.		Linear	Significant effects vary by players of interest (Williams, Federer, Nadal)

Federer, Nadal, and Williams: most available data and most detailed individual models

Player	Model Finding	Interpretation
Federer	Expected % points won at **US Open** **greater than** at **Wimbledon** **if** W/UE large	On average, better at **Wimbledon** but given **peak performance**, better at **US Open**
Nadal	Expected % points won **decreases** as % of points won at net **increases**	Indicative of a change of strategy
Williams	Expected % points won at **French Open** **greater than** at **Australian Open** **if** % of aces increases by 1%	Serving well at **French Open** is more important than serving well at **Australian Open**

Conclusions

- Surface effects are not apparent until we utilize tennis-specific features (e.g. unforced errors, aces) and vary across players
- With full, feature rich player data, we can make more interesting conclusions for individual players (e.g. Williams, Federer, Nadal)
- Our data are only available for some matches -- need more, detailed tennis data for modeling lower-tier players
- We are in talks with the Chief Technology Officer for the US Tennis Association

Game. Set. Match.

https://github.com/shannong19/courtsports

Kayla Frisoli

@stat_frizz
http://stat.cmu.edu/~kfrisoli

Shannon Gallagher

@shannonkgallagh
http://stat.cmu.edu/~sgallagh/

Amanda Luby

@amandaluby
http://stat.cmu.edu/~aluby/

Game. Set. Match.

Modeling win probability: only rank is signif.

- Outcome: Wins

- Predictors: ATP, IOC, Late round, **Rank**, **Opponent Rank**, Court, Year

- $\text{logit}\ (P(\ Y=1 \mid \mathbf{X})) = B_1\mathbf{X}_{\mathbf{fixed}} + B_2\mathbf{X}_{\mathbf{random}}$

- No significant player-level effects

Does surface matter? For whom?

- Do results differ across the three surface types (grass, clay, hard)?
  Yes.

- How useful is including tennis specific features
  (e.g. winners, aces, unforced errors)?
  Quite useful.

- Are there player-level effects in performance on different surfaces?
  Only when looking at tennis-specific outcomes

Modeling of Individuals: Details

- Linear regression: $E[(\% \text{ Points Won})_{\text{Player}}] = B\mathbf{X}_{\text{Player}}$
- Covariates (**X**) include opponent ranking, surface type, average service speed, winners, unforced errors, break points won, net points won, etc.
- Models fit using forward-backwards stepwise regression
- Best model for each player chosen with AIC

Logistic Model (GS data): logit (P(Y=1 | $\mathbf{X}$)) = $B_1\mathbf{X}_{\mathbf{fixed}} + B_2\mathbf{X}_{\mathbf{random}}$

Y	Winner? (1 = yes, 0 = no)
$\mathbf{X}_{\mathbf{fixed}}$	ATP, IOC, Late round, ***Rank**, ***Opponent Rank**, Year
$\mathbf{X}_{\mathbf{random}}$	Court

Linear Model: E (Y | $\mathbf{X}$) = $B_1\mathbf{X}_{\mathbf{fixed}} + B_2\mathbf{X}_{\mathbf{random}}$

Y	Number of aces, number of net points won, or number of unforced errors
$\mathbf{X}_{\mathbf{fixed}}$	ATP, IOC, Late round, ***Rank**, ***Opponent Rank**, Year
$\mathbf{X}_{\mathbf{random}}$	Court

Model 3: E (Y| $\mathbf{X}$) = $B_1\mathbf{X}_{\mathbf{fixed}}$

Y	% points won by Federer, % points won by Nadal, % points won by Williams
$\mathbf{X}_{\mathbf{fixed}}$	opponent ranking, surface type, average service speed, winners, unforced errors, break points won, net points won, etc.

The grand slams are played on distinct surfaces and may affect player performance.

Grand Slam	**AUSTRALIAN OPEN**	**FRENCH OPEN**	**WIMBLEDON**	**US OPEN**
Surface	DecoTurf (hard court)	clay	grass	Plexicushion (hard court)

Federer, Nadal, and Williams: most available data and most detailed individual models

Federer

- E[points won] ↑ **@Wimbledon** compared to other slams on average
- W/UE large → more E[points won] **@US Open** compared to Wimbledon

Nadal

- E[points won] ↓ as volley points won ↑

Williams

- E[points] ↑ **more** for number of aces ↑ **@French Open** compared to @**Australian Open**