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.

<table>
<thead>
<tr>
<th>Grand Slam</th>
<th>Surface</th>
<th>Known top players</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUSTRALIAN OPEN</td>
<td>DecoTurf</td>
<td>Serena Williams</td>
</tr>
<tr>
<td></td>
<td>(hard court)</td>
<td></td>
</tr>
<tr>
<td>FRENCH OPEN</td>
<td>clay</td>
<td>Rafael Nadal</td>
</tr>
<tr>
<td>WIMBLEDON</td>
<td>grass</td>
<td>Roger Federer</td>
</tr>
<tr>
<td>US OPEN</td>
<td>Plexicushion</td>
<td>Serena Williams</td>
</tr>
<tr>
<td></td>
<td>(hard court)</td>
<td></td>
</tr>
</tbody>
</table>
The grand slams are played on distinct surfaces and may affect player performance.

<table>
<thead>
<tr>
<th>Grand Slam</th>
<th>Surface</th>
<th>Known top players</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUSTRALIAN OPEN</td>
<td>DecoTurf (hard court)</td>
<td>Serena Williams</td>
</tr>
<tr>
<td>FRENCH OPEN</td>
<td>clay</td>
<td>Rafael Nadal</td>
</tr>
<tr>
<td>WIMBLEDON</td>
<td>grass</td>
<td>Roger Federer</td>
</tr>
<tr>
<td>US OPEN</td>
<td>Plexicushion (hard court)</td>
<td>Serena Williams</td>
</tr>
</tbody>
</table>
Two data sets—two perspectives

- Data from Jeff Sackman’s website ([https://github.com/JeffSackmann](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.
Players perform differently at Wimbledon, as displayed by the clustering of purple points.
Spaniards outperform on clay surface

Win proportion among Spaniards

Average ranking among Spaniards

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.
<table>
<thead>
<tr>
<th>Approach 1</th>
<th>Data</th>
<th>Y</th>
<th>Fixed X</th>
<th>Random X</th>
<th>Regression</th>
<th>Conclusions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GS Data</td>
<td>Did win?</td>
<td>league, country, year, late round</td>
<td>surface</td>
<td>Logistic</td>
<td>No significant effects besides rank / opponent rank</td>
</tr>
<tr>
<td></td>
<td>n=10,160</td>
<td>(Yes = 1, No = 0)</td>
<td>op. rank, rank</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data</td>
<td>Y</td>
<td>Fixed X</td>
<td>Random X</td>
<td>Regression</td>
<td>Conclusions</td>
<td></td>
</tr>
<tr>
<td>--------------------</td>
<td>--------------</td>
<td>-------------------------------------------------------------------------</td>
<td>----------</td>
<td>------------</td>
<td>---------------------------------------</td>
<td></td>
</tr>
<tr>
<td>Approach 1</td>
<td>GS Data</td>
<td>Did win? (Yes = 1, No = 0)</td>
<td>league, country, year, late round op. rank, rank</td>
<td>surface</td>
<td>Logistic</td>
<td></td>
</tr>
<tr>
<td></td>
<td>n=10,160</td>
<td></td>
<td></td>
<td></td>
<td>No significant effects besides rank / opponent rank</td>
<td></td>
</tr>
<tr>
<td>Approach 2</td>
<td>GS Data %&gt;% join(PBP)</td>
<td>Aces, Points won at net, UE</td>
<td>league, country, year, late round op. rank, rank</td>
<td>surface</td>
<td>Linear</td>
<td></td>
</tr>
<tr>
<td></td>
<td>n=6,132 (aces)</td>
<td></td>
<td></td>
<td></td>
<td>Significant surface effects for Williams, Federer, Nadal</td>
<td></td>
</tr>
</tbody>
</table>
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.
<table>
<thead>
<tr>
<th>Data</th>
<th>Y</th>
<th>Fixed X</th>
<th>Random X</th>
<th>Regression</th>
<th>Conclusions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Approach 1</td>
<td>GS Data</td>
<td>Did win?</td>
<td>league, country, year, late round op. rank, rank</td>
<td>surface</td>
<td>Logistic</td>
</tr>
<tr>
<td></td>
<td>n=10,160</td>
<td>(Yes = 1, No = 0)</td>
<td></td>
<td></td>
<td>No significant effects besides rank / opponent rank</td>
</tr>
<tr>
<td>Approach 2</td>
<td>GS Data %&gt;% join(PBP)</td>
<td>Aces</td>
<td>league, country, year, late round op. rank, rank</td>
<td>surface</td>
<td>Linear</td>
</tr>
<tr>
<td></td>
<td>n=6,132 (aces)</td>
<td>Points won at net</td>
<td></td>
<td></td>
<td>Significant surface effects for Williams, Federer, Nadal</td>
</tr>
<tr>
<td></td>
<td>n=6,132 (net)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>n=6,132 (UE)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>UE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Approach 3</td>
<td>GS Data %&gt;% join(PBP) %&gt;% filter(player == &quot;{Player}&quot;)</td>
<td>% points won</td>
<td>league, country, year, late round op. rank, rank</td>
<td>% points won</td>
<td>Linear</td>
</tr>
<tr>
<td></td>
<td>n=75 (Nadal)</td>
<td></td>
<td>average service speed, winners, unforced errors, break points won, net points won, etc.</td>
<td>Linear</td>
<td>Significant effects vary by players of interest (Williams, Federer, Nadal)</td>
</tr>
</tbody>
</table>
Federer, Nadal, and Williams: most available data and most detailed individual models

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

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

\[
\text{logit } (P(Y=1 | X)) = B_1 X_{\text{fixed}} + B_2 X_{\text{random}}
\]

- 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: \( \text{E}[\text{(% Points Won)}_{\text{Player}}] = B \times \text{X}_{\text{Player}} \)

- Covariates (\( \text{X} \)) 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 \( P( Y=1 | X ) \) = \( B_1 X_{fixed} + B_2 X_{random} \)

- \( Y \): Winner? (1 = yes, 0 = no)
- \( X_{fixed} \): ATP, IOC, Late round, *Rank, *Opponent Rank, Year
- \( X_{random} \): Court

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

- \( Y \): Number of aces, number of net points won, or number of unforced errors
- \( X_{fixed} \): ATP, IOC, Late round, *Rank, *Opponent Rank, Year
- \( X_{random} \): Court

Model 3: \( E( Y | X ) = B_1 X_{fixed} \)

- \( Y \): % points won by Federer, % points won by Nadal, % points won by Williams
- \( X_{fixed} \): 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.

<table>
<thead>
<tr>
<th>Grand Slam</th>
<th>AUSTRALIAN OPEN</th>
<th>FRENCH OPEN</th>
<th>WIMBLEDON</th>
<th>US OPEN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surface</td>
<td>DecoTurf (hard court)</td>
<td>clay</td>
<td>grass</td>
<td>Plexicushion (hard court)</td>
</tr>
</tbody>
</table>
Federer, Nadal, and Williams: most available data and most detailed individual models

**Federer**
- $E[\text{points won}] \uparrow$ at **@Wimbledon** compared to other slams on average
- $W/UE$ large $\rightarrow$ more $E[\text{points won}]$ at **@US Open** compared to Wimbledon

**Nadal**
- $E[\text{points won}] \downarrow$ as volley points won $\uparrow$

**Williams**
- $E[\text{points}] \uparrow$ **more** for number of aces $\uparrow$ at **@French Open** compared to **@Australian Open**