# Opening up the court (surface) in tennis grand slams

*September 15, 2018*

**Abstract**

Tennis grand slams consist of the Australian Open, French Open, Wimbledon, and US Open, which are played on hard (Plexicushion), clay, grass, and hard (DecoTurf) courts, respectively. The surface type may substantially impact ball speed, height, and spin as well as player speed and agility. It is also believed that play style and practice habits may contribute to different results across surface types. For example, Rafael Nadal is thought to be the best clay court player of all time whereas Roger Federer is particularly known for dominance at Wimbledon. On the women's side, Serena Williams once struggled on clay courts but has seemingly transformed her style to perform better on clay courts, but has perhaps suffered on grass as a consequence. In this analysis, we examine the result of the top 100 players in grand slams from 2013-2017 across the four different surfaces. We create a hierarchical model with fixed and random effects to predict the number of points won in a match. We take into consideration player-specific effects, nationality (which is thought to have an effect on play style), sex, ranking, ELO, and game statistics. We assess the fit of our model using standard statistical techniques (e.g. MSE, AIC, BIC, residual diagnostics) in addition to 'common knowledge' factors (for instance, Rafael Nadal should be indicated as a superior clay court player by the model). We compare the results of top 100 players across grand slams to examine the effect of court surface. We also provide an in-depth analysis of Nadal, Federer, and S. Williams.

## Contents



# 1 Introduction

Rafael Nadal is known as the "King of Clay" in tennis, having won 11 out of his current 17 grand slams titles at the French Open, which is played on a clay surface (Jurejko 2018). In contrast, his rival Roger Federer has won his most grand slam titles (8 out of 20) at Wimbledon, which is played on grass. On the women's side, Serena Williams, winner of 24 grand slam titles, has been dominant both on hard court (7 titles at the Australian Open and 6 at US Open) and grass (7 at Wimbledon). This trend extends to other top players, who seem to have better results at some grand slams than others. More broadly, it seems that country of origin has an interaction effect with court type. For example, Spanish players seem to excel on clay courts and Americans have great success at Wimbledon despite grass courts not being of wide use in the USA. It also worth questioning whether the US Open and Australian Open should be grouped together as hard courts despite having different surface compositions (Paxinos 2007). In this paper, we analyze the results of grand slam players from 2013-2017, and we

1. Determine if and how court surface effects players by implementation of a series of nested hierarchical models
2. Examine how Nadal, Federer, and Williams' play differs by surface
3. Assess whether we can group the two hard court surfaces together.

As to issue (1) quantifying the effect of court surface on players, there has not been much written about with regards to tennis. There are materials available in the literature for forecasting the outcome of tennis matches (Klaassen and Magnus 2003; Newton and Keller 2005; McHale and Morton 2011; Kovalchik 2016)] or for assessing whether points within a match are independent and identically distributed (Klaassen and Magnus 2001). (Knottenbelt, Spanias, and Madurska 2012) do take into account surface in their model but do not compare the results of one surface to another. Other sports analyses do take into account surface type such as grass vs. turf in soccer and football. Results from these studies show that surface type does have an effect on the game, either directly or indirectly [Andersson, Ekblom, and Krustrup (2008); Gains et al. (2010);].

We use models that take into account both individual and group effects such as in the Gaussian-process player production basketball model or predicting individual soccer performance (Page, Barney, and McGuire 2013; Egidi and Gabry 2018). Both of those models had success using hierarchical Bayesian models, which we employ in our own models. More specifically, we model the players' expected points in a match based on the player's own characteristics, the court/tournament effects, and the opponent's ranking.

For issue (2) the player analysis of Nadal, Federer, and Williams, we examine whether our model passes the "common sense" tests like how the models in (Thomas et al. 2013) show that commonly well known hockey players also have high status in the model. We also examine whether these players do have surface apparent effects. Few academic papers have been written about Nadal, Federer, or Williams. One paper studies Federer's odds of winning when Nadal suddenly withdrew from Wimbledon showed that Federer was too heavily favored by bookmakers (Leitner, Zeileis, and Hornik 2009). One analysis of Williams shows how she has gotten better with age, even past the point when other greats began to decline, but the study does not look at surface type (Morris 2015).

Finally, for issue (3), we use clustering methods in order to determine which court surface types are more similar to one another.

Readers may object that we are looking at differences between grand slams, which each have their own time period, weather conditions, play time conditions, and "home court effects" instead of differences in surfaces alone. However, (1) grand slam data is the most readily available and most complete which makes it the best choice at the moment for modelling, (2) we adjust for these confounders where we can, and (3) analyzing the difference in the grand slams is still useful as they are considered to be the most prestigious events in tennis.

The rest of this paper is organized as follows. In Section Data we describe our grand slam tennis data. In Section Early Data Analysis we examine the data at a high level and use clustering whether to determine how the courts differ from one another. In Section Methods we describe our hierarchical models we use to determine difference in court surfaces. In Section Results we describe the results of our modelling and also examine the play of Nadal, Federer, and Williams. Finally in Section Discussion, we discuss future work and extensions or our model.

# 2 Data and EDA

## 2.1 Data

The primary data consists of 5080 matches split evenly over the four grand slams and the two leagues: ATP (men's) and WTA (women's). Each match has 80 attributes, many of which are redundant. We focus on the

Table 1: Example of the grand slam data. It includes winner and loser attributes, match attributes, and tournament attributes. Not all attributes are shown here.

| Winner | Tournament | Year | W. IOC | W. Points | W. Rank | L. Points | L. Rank |
|---|---|---|---|---|---|---|---|
| Serena Williams | Australian Open | 2013 | USA | 52 | 3 | 18 | 110 |
| Serena Williams | Australian Open | 2013 | USA | 70 | 3 | 41 | 112 |
| Roger Federer | Australian Open | 2013 | SUI | 95 | 2 | 63 | 46 |
| Roger Federer | Australian Open | 2013 | SUI | 111 | 2 | 86 | 40 |
| Rafael Nadal | Roland Garros | 2013 | ESP | 140 | 4 | 115 | 59 |
| Rafael Nadal | Roland Garros | 2013 | ESP | 113 | 4 | 90 | 35 |

following attributes for both the winner and loser of the match: games won, points won, retirement, break points faced, break points saved, aces, country of origin, and player attributes. Additionally, we take into account the number of sets in a match, the surface type, and round of the tournament. A subset of the data is shown in Table 1.

The secondary data point by point data for grand slam matches. In this data set, each row is a point in a match with details on who won the point, whether a player had a forced or unforced error, winner, ace, or net point win, and serve speed. There is also tournament info such as court, year, and time start. Additional attributes include rally length, winner and final score of the match, retirement, and minutes played. However, these additional attributes are only available for about 1/10 of the data.

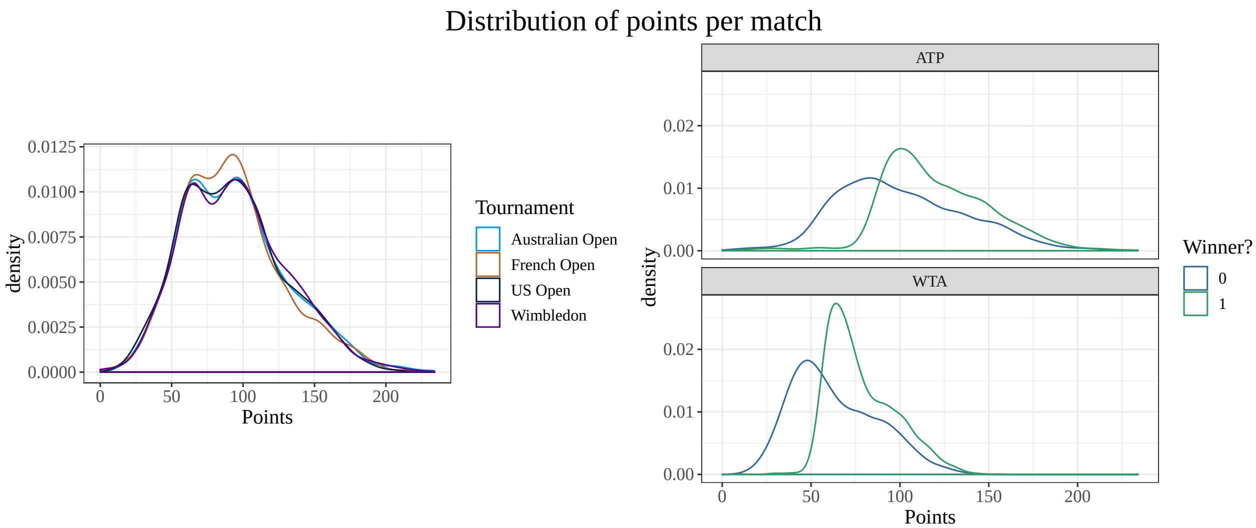
We aggregate the point by point data into match data, summing the total number of errors, winners, aces, net points, and taking the average service speed for each player in the match. We then join the point by point match data with the grand slam data to find the score of the match. We do this by matching on tournament, year, player one, and player two. This final data set consists of 3066 observations (compared to the 5080 in the original grand slams). The median rank of the winners for the original data is 28 and the median rank for the winners of the partial data is 21. The same trend is true for the loser rank. This means that it is more likely for better players to have court by court data, which may effect our analysis. Additionally, we found that no women's point by point data was recorded for 2015.

All data is obtained from Jeff Sackmann's open website via the `R` package `deuce` (Sackmann 2018; Kovalchik 2017). All steps of our analysis from collection to dissemination are freely available online.

## 2.2 Early Data Analysis

### 2.2.1 Examining the distribution of points earned

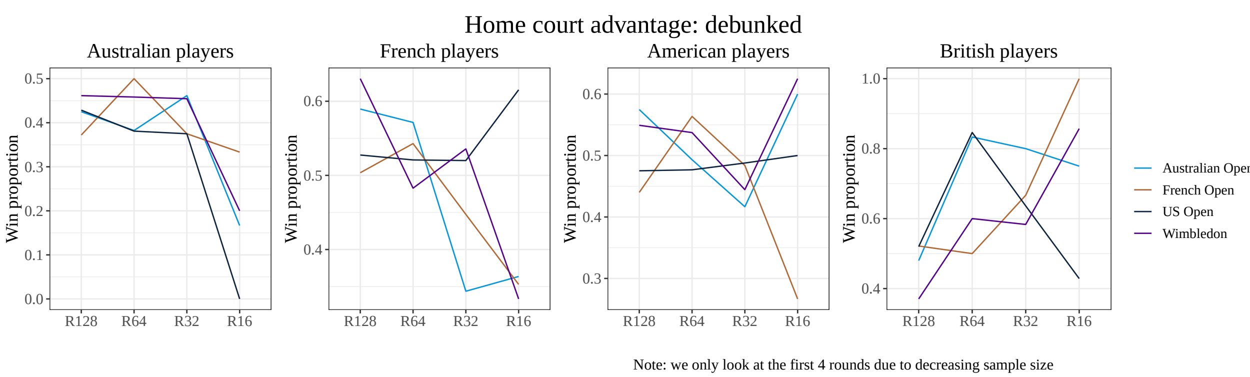
We first examine the distribution of points earned per match to assess normality. The distribution of points earned is approximately normal. This distribution is similar across tournament, with Wimbledon differing slightly from the other grand slams. As expected, there are more points earned in the WTA than the ATP due to the differing numbers of games played. Also unsurprisingly, the winners of the match tended to earn more points than the losers.




### 2.2.2 Home court advantage

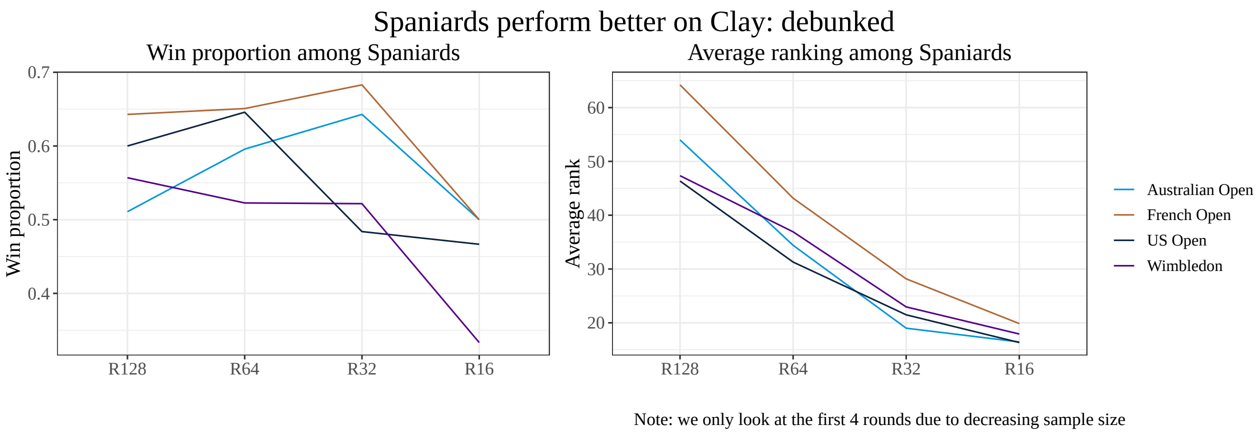
It is commonly thought that there is a home court advantage in grand slam games (SOURCE). In our data we find this to be true (i.e. French players win the French open more than French players win other slams). But, we also know that the home team is given preference for wild card bids (SOURCE) so potentially citizens of a particular country play in "their" tournament more often than they play in other tournaments. We also find this to be true in our data for France, The United States, and Australia (i.e. the proportion of French players in the French Open is greater than the proportion of French players in other slams).

Therefore, we want to see how the proportion of wins for the home country changes across the different tournaments. If there was really a home court advantage, the proportion of French wins each round would be higher at the French Open than the Australian Open, the US Open, and Wimbledon. The same would be true for Australia and the US. But, we see that this isn't the case. After accounting for the number of players from each country, we don't find a home court advantage in the grand slam.




## 2.3 Spaniards on clay

It is also commonly thought that Spaniards play better on clay. We are interested in whether Spaniards win the French Open more than they win in other tournaments. It does appear that Spaniards are winning the French Open more than they are winning other tournaments. But, this result is not significant. In addition, the ranking of Spaniards in the French Open is, on average, higher than other tournaments, which may help explain this common misconception, although this result is not significant either.




## 2.4 Tall players and aces

Why the actual EFF won't fig.width work?

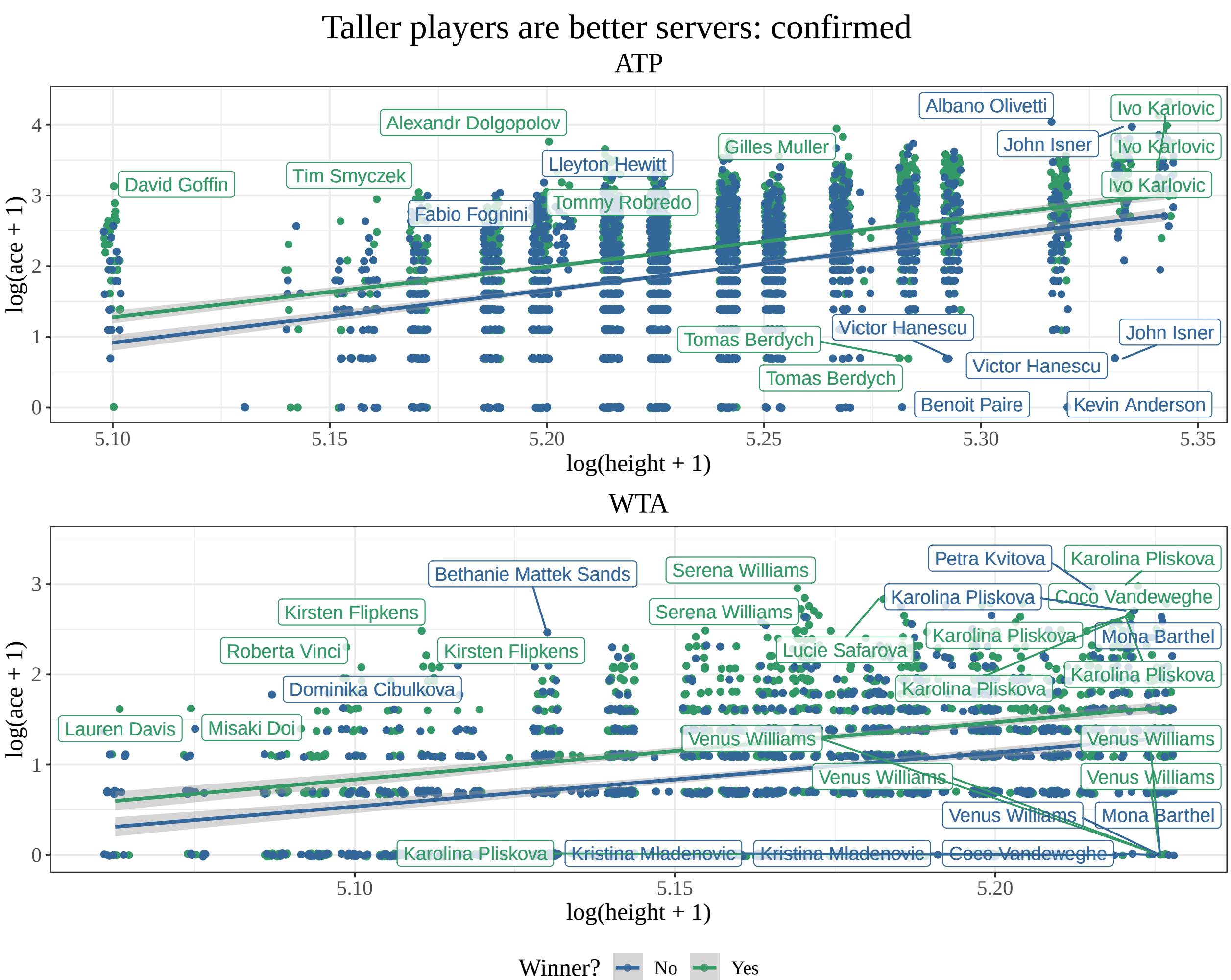


Table 2: Number of matches played for Nadal, Federer, and Williams from 2013-2017 at each of the grand slams.

| Tournament | Nadal | Federer | Williams |
|---|---|---|---|
| Australian Open | 20 | 28 | 30 |
| French Open | 29 | 14 | 23 |
| US Open | 21 | 22 | 26 |
| Wimbledon | 11 | 29 | 21 |
| **Total** | **81** | **93** | **100** |

# 3 Methods

## 3.1 Examining individual players

In Table 2, we display the number of matches Nadal, Federer, and Williams have played from 2013-2017. Over that time span, Nadal won 6 grand slams, Federer won 1, Williams won 8. Despite this, Federer played more total matches in Nadal. All three players were absent for exactly three slams during this time period due to external factors (Nadal: (1 AO, 0 FO, 1 Wim, 1 USO), Federer: (0 AO, 2 FO, 0 Wim, 1 USO), Serena: (0 AO, 1 FO, 1 Wim, 1 USO)). Not unexpectedly, Nadal has the most wins on clay (29), Federer has the most wins at Wimbledon (29) and Williams the most on hardcourt (30 at AO and 21 at USO). Despite having played more matches than Nadal, Federer only has 1 grand slam to show compared to Nadal's six, which indicates that Federer made it deeper into tournaments on average but had difficulties winning the championships. For the WTA, Williams had her second most successful five-year span at the grand slams over this time period, winning 8.

We fit individual linear models for these three individuals (subsetting the data to their matches only), estimating the percent of points won in a match using stepwise regression. The lower model is the percent of points regressed on the opponent rank and indicator variables for court type with the FO as the reference variable. The upper model is the lower model with the additional variables of winner to unforced error ratio (W/UE), average serve speed, percent aces, percent break points, percent net points won, and their interaction effects with court type. We do not display the full best fit models here, but they are available online. We do report the effects ($\pm 2\cdot$ Std. Err.) of significant variables and their interaction terms, adjusting for the other covariates. Since we postulated that Nadal is best on clay, Federer on grass, and Williams on hard court, specifically, AO, we use the FO, Wim., and AO, respectively, as the reference variable in the model. For Nadal and Federer we use forward-backwards stepwise regression, using AIC as the criterion, beginning with the full model. For Williams, we also use forwards-backwards stepwise regression but begin at the low model since she has more missing data.

## 3.2 Nadal: King of Clay: confirmed

Looking at Nadal individually, the model with the lowest AIC has positive coefficients opponent rank, percent of aces, and percent of break points won. It has negative coefficients for AO, USO, and Wim. compared to FO and for percent of net points won. There are no interaction effects in this model.

The standardized residuals plot shown in Figure 1 (along with the other linear regression diagnostics plots not shown here) shows that the model is a good fit for predicting the percent of points won. For the significant variables ($\alpha = 95\%$), while adjusting for the other variables, we expect percent of points won to be $0.034 \pm 0.028$ less at the AO and $0.046 \pm 0.033$ less at Wim compared to the FO; to increase by $1.5 \times 10^{-4} \pm 1.3 \times 10^{-4}$ for a one unit increase in opponent rank; a $0.0012 \pm 5.6 \times 10^{-4}$ increase for a .01 increase in percent of break points won; and to decrease by $8.8 \times 10^{-4} \pm 5.5 \times 10^{-4}$ for a .01 increase in percent of net points won. As such, we see clear evidence Nadal performs better at the FO compared to the other grand slams.

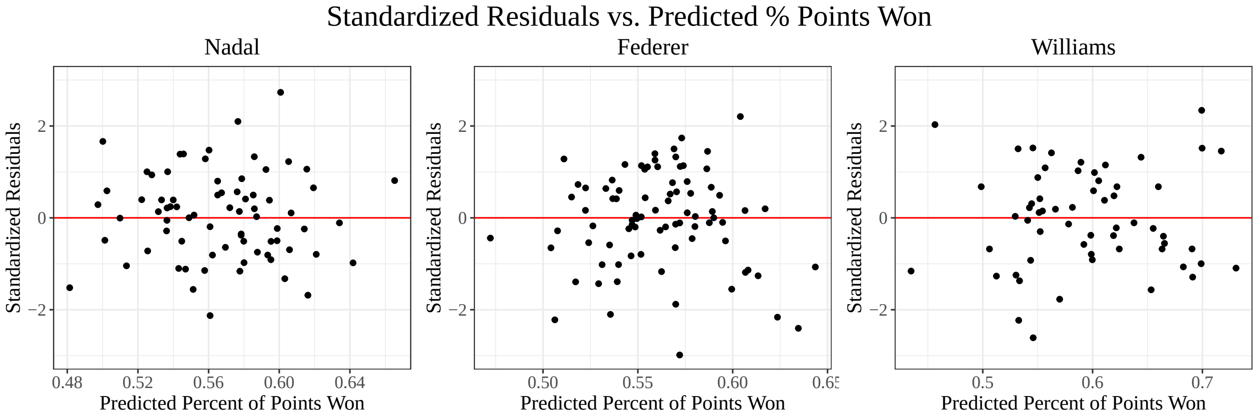



Figure 1: Standardized residuals vs percent of points won best fit model from forward-backwards stepwise regression for Nadal's, Federer's, and William's, respectively.}

### 3.3 Federer: Wimbledon extraordinaire: debunked

For Federer, the model with the lowest AIC has negative coefficients AO, FO, and USO. vs Wim., and W/UE, It has positive coefficients for opponent rank, percent of break points won, and all the interaction terms: court and W/UE.

The standardized residuals plot shown in Figure 1 shows that the model is a fairly good fit for predicting the percent of points won, but almost seem to see two clusters of residuals splitting with predicted point percentage of about 0.63. For the significant variables ($\alpha = 95\%$), while adjusting for the other variables, we expect percent of points won to be $0.04 \pm 0.034$ less at AO compared to the Wim.; $0.058 \pm 0.046$ less at FO compared to the Wim.; $0.007 \pm 0.093$ less at USO compared to the Wim.; to decrease by $0.022 \pm 0.019$ for a one unit increase in W/UE at AO, $0.073 \pm 0.066$ for a one unit increase in W/UE at FO, and $0.041 \pm 0.033$ for a one unit increase in W/UE at USO. This is strong evidence that Federer performs best at Wimbledon but possibly does better at the USO, depending on his W/UE ratio.

### 3.4 Williams: hard court magician only: debunked

For Williams, the model with the lowest AIC is the largest model of the three selected 22 coefficients. Since Williams only has 59 observations, we believe this model is overfitting and so caution against inference with this model. It should also be noted that no women's point by point data was recorded for 2015 so that year is excluded from this model. The coefficients are court, opponent rank, W/UE, average serve speed, percent aces, percent break points won, percent net points won, and interaction effects with court and W/UE, court and average service speed, court and percent aces, and court and break points won.

The standardized residuals plot shown in Figure 1 shows that the model is a fair fit for predicting the percent of points won, but we tend to overestimate the percent of points won in comparison with Nadal. For the significant variables ($\alpha = 95\%$), while adjusting for the other variables, we expect percent of points won to increase by $4.8 \times 10^{-4} \pm 3.2 \times 10^{-4}$ for a 1 unit increase in oponent rank; to increase by $0.034 \pm 0.02$ for a one unit increase in W/UE at FO, to increase by $0.092 \pm 0.045$ for a one unit increase in W/UE at USO, and to increase by $0.0038 \pm 0.066$ for a one unit increase in W/UE at Wim. compared to AO; to increase by $2 \times 10^{-4} \pm 1.6 \times 10^{-4}$ for a .01 increase in percent of aces at the French Open; and to increase by $0.0024 \pm 0.0011$ for a .01 increase in percent of break points won at the French Open.

For this, we do not see that Williams is dominant on the hard courts compared to the other courts. Rather, she has good results on all the courts. Her significant interaction effects show that if she focuses on acing her opponent and capitalizing on breakpoints, then she has a better chance at winning more points than at the Australian Open.

## 3.5 Mixed-effects models for all grand slams

### 3.5.1 Logistic regression for wins

We will use a multilevel logistic regression framework. This allows us to include fixed effects, which remain the same for each player, as well as include player-level effects. Since players appear multiple times in the data, the observations are not independent and including a player-level effect allows us to account for this dependence. It also provides a way to examine individual player effects and assess, for instance, whether Federer is more likely to win at Wimbledon after accounting for other variables.

We initially planned to use points won by an individual as our outcome variable, but found that it may not be an ideal measure of player performance. For instance, players may score few points in a match due to a poor performance, but they may also score few (relative) points if they win a short match. Modeling the number of points won is also complicated by the difference between length of matches for men (best of 5 sets) and women (best of 3 sets). For these reasons, we chose to model whether a player won the match, with the following predictors:

- `ioc_fac`: Country that the player represents
- `tournament`: Which Grand Slam the match was played at
- `late_round`: Indicator for whether the match occured in Round of 16 or later
- `rank`: Rank of player at the time of the match
  - Included on log scale
- `opponent_rank`: Rank of opponent at the time of the match
  - Included on log scale
- `year`: Factor variable with a level for each year included in the dataset (2013-2017)
- `atp`: Indicator that the match was played in the ATP league instead of the WTF league.

(short discussion of models fit, show plots for final model)

### 3.5.2 Linear regression for other stats

Rank and opponent's rank clearly have the largest effect on predicting the winner. Looking at the individual effects for Serena, Nadal, and Federer; they actually fall near the average for all four grand slams. Including rank as an indicator (seeded/not-seeded) substantially decreases the fit of the model, even when individual effects are included. We now look at a similar model, but instead of using 'win' as the outcome variable, we'll look at game statistics that may be more sensitive to court surface – aces, winners, and net points won – and see if individual differences are detectable.

(some plots that show a) Serena/Federer are stronger servers, b) Federer makes fewer unforced errors at Wimbledon, c) something about Nadal)

# 4 Discussion

# 5 References

Andersson, Helena, Björn Ekblom, and Peter Krustrup. 2008. "Elite Football on Artificial Turf Versus Natural Grass: Movement Patterns, Technical Standards, and Player Impressions" 26 (February): 113–22.

Egidi, Leonardo, and Jonah Gabry. 2018. "Bayesian Hierarchical Models for Predicting Individual Performance in Soccer." *Journal of Quantitative Analysis in Sports* 14 (3). De Gruyter: 143–57.

Gains, Graydon L, Andy N Swedenhjelm, Jerry L Mayhew, H Michael Bird, and Jeremy J Houser. 2010. "Comparison of Speed and Agility Performance of College Football Players on Field Turf and Natural Grass." *The Journal of Strength & Conditioning Research* 24 (10). LWW: 2613–7.

Jurejko, Jonathan. 2018. "French Open 2018: Why does 'King of Clay' Rafael Nadal reign supreme?" Edited by BBC Sport at Roland Garros. https://www.bbc.com/sport/tennis/44385223.

Klaassen, Franc J.G.M., and Jan R. Magnus. 2003. "Forecasting the Winner of a Tennis Match." *European Journal of Operational Research* 148 (2): 257–67. https://doi.org/https://doi.org/10.1016/S0377-2217(02)00682-3.

Klaassen, Franc J.G.M., and Jan R Magnus. 2001. "Are Points in Tennis Independent and Identically Distributed? Evidence from a Dynamic Binary Panel Data Model." *Journal of the American Statistical Association* 96 (454). Taylor & Francis: 500–509. https://doi.org/10.1198/016214501753168217.

Knottenbelt, William J., Demetris Spanias, and Agnieszka M. Madurska. 2012. "A Common-Opponent Stochastic Model for Predicting the Outcome of Professional Tennis Matches." *Computers & Mathematics with Applications* 64 (12): 3820–7. https://doi.org/https://doi.org/10.1016/j.camwa.2012.03.005.

Kovalchik, Stephanie. 2017. *Deuce: Resources for Analysis of Professional Tennis Data.*

Kovalchik, Stephanie Ann. 2016. "Searching for the Goat of Tennis Win Prediction." *Journal of Quantitative Analysis in Sports* 12 (3). De Gruyter: 127–38.

Leitner, Christoph, Achim Zeileis, and Kurt Hornik. 2009. "Is Federer Stronger in a Tournamentwithout Nadal? An Evaluation of Odds and Seedings for Wimbledon 2009." *Austrian Journal of Statistics* 38 (4): 277–86.

McHale, Ian, and Alex Morton. 2011. "A Bradley-Terry Type Model for Forecasting Tennis Match Results." *International Journal of Forecasting* 27 (2): 619–30. https://doi.org/https://doi.org/10.1016/j.ijforecast.2010.04.004.

Morris, Benjamin. 2015. "Serena Williams and the Difference Between All-Time Great and Greatest of All Time." Edited by FiveThirtyEight. https://fivethirtyeight.com/features/serena-williams-and-the-difference-between-all-time-great-and-greatest-of-all-time/.

Newton, Paul K, and Joseph B Keller. 2005. "Probability of Winning at Tennis I. Theory and Data." *Studies in Applied Mathematics* 114 (3). Wiley Online Library: 241–69.

Page, Garritt L, Bradley J Barney, and Aaron T McGuire. 2013. "Effect of Position, Usage Rate, and Per Game Minutes Played on Nba Player Production Curves." *Journal of Quantitative Analysis in Sports* 9 (4). De Gruyter: 337–45.

Paxinos, Stathi. 2007. "Australian Open court surface is speeding up." Edited by The Age. https://www.theage.com.au/sport/tennis/australian-open-court-surface-is-speeding-up-20071120-ge6chj.html.

Sackmann, Jeff. 2018. "Tennis Data Repositories." https://github.com/JeffSackmann.

Thomas, A. C., Samuel L. Ventura, Shane T. Jensen, and Stephen Ma. 2013. "COMPETING Process Hazard Function Models for Player Ratings in Ice Hockey." *The Annals of Applied Statistics* 7 (3). Institute of Mathematical Statistics: 1497–1524. http://www.jstor.org/stable/23566482.