Analyzing the Effect of Surface in Tennis Grand Slams

Kayla Frisoli, Shannon Gallagher, and Amanda Luby

September 14, 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.

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; S. A. 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

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

2013; Egidi and Gabry 2018). Both of those models had success using hierarhcical 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 oponent'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

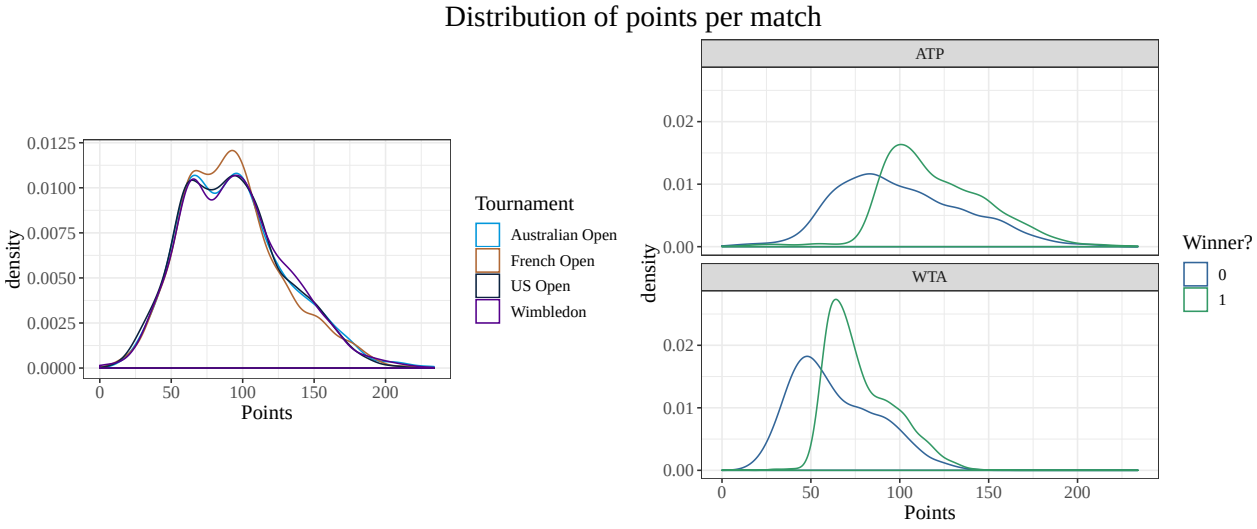
The data consists of 5080 matches split evenly over the four grand slams and the two leagues: ATP (men's) and WTA (women's). Each match has 80 attributes, many of which are redundant. We focus on the following attributes for both the winner and loser of the match: games won, points won, retirement, break 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 data is obtained from Jeff Sackmann's open website via the `R` package `deuce` (Sackmann 2018; S. 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

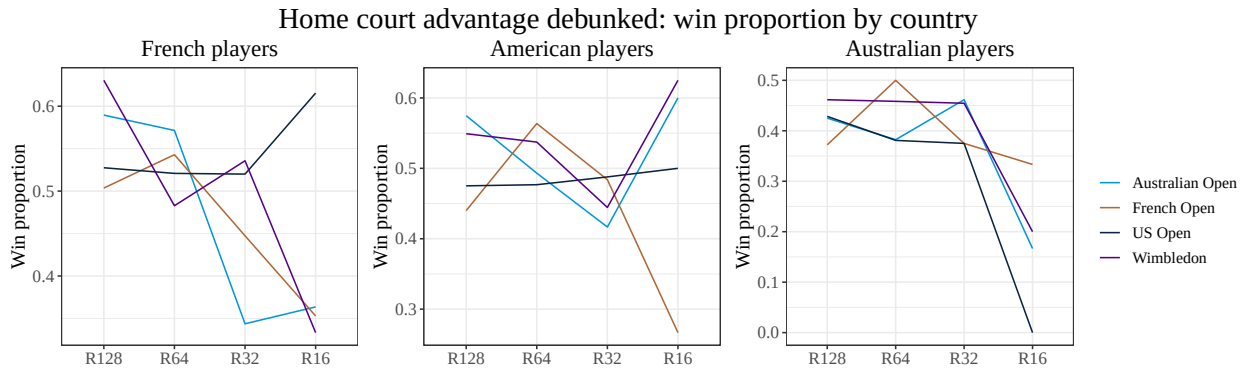
We first examine the distibution of points earned per match. We find that the distribution is similar across tournament, with Wimbledon differing slightly from the other grand slams. As expected, there are more points earned in the WTA than the ATP due to the differing numbers of games played. Also unsuprisingly, 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.



Note: we only look at the first 4 rounds due to decreasing sample size as the tourr

3 Methods

4 Results

5 Discussion

6 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. 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. doi:10.1198/016214501753168217.

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. doi:https://doi.org/10.1016/S0377-2217(02)00682-3.

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. doi: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. doi: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.