Analyzing the Effect of Surface in Tennis Grand Slams

Kayla Frisoli, Shannon Gallagher, and Amanda Luby

September 09, 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 [cite]. 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 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 [cite]. In this paper, we analyze the results of grand slam players from 2013-2017, and we

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.

Tournament	Year	W. Country	W. Points	Winner	W. Rank	L. Points	Loser
US Open	2014	USA	58	Serena Williams	1	31	Taylor Townsend
US Open	2013	ESP	136	Rafael Nadal	2	112	Philipp Kohlschreiber
Australian Open	2015	ESP	91	Rafael Nadal	3	51	Mikhail Youzhny
US Open	2013	USA	65	Serena Williams	1	41	Yaroslava Shvedova
Australian Open	2017	USA	73	Serena Williams	2	60	Lucie Safarova

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) the hiearchical modelling, we use models similar to [cite] and [cite]. 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. We use a Bayesian method because in [cite,cite] they had success in similar situations.

For issue (2) the player analysis of Nadal, Federer, and Williams, we examine whether our model passes the "common sense"" tests (e.g. these players should win often when contrasting them to opponents) like in [ventura something] and whether these players do have surface apparent effects.

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

Readers may object that we are looking at difference 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 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, minutes played, and round of the tournament. A subset of the data is shown in Table 1.

The data is obtained from Jeff Sackman's open website [cite] via the `R` package `deuce` [cite]. All steps of our analysis from collection to dissemination are freely available online and are compiled into an `R` package http://github.com/shannong19/courtsports.

2.2 Early Data Analysis

3 Methods

4 Results

5 Discussion

6 References