Applying Hierarchical Bayesian Models to ATP Data

Horace Shew

Abstract

The ATP tour is a tennis tour for professional men's tennis players managed by the Association of Tennis Professionals. The tour consists of tournaments hosted annually, held all across the world on different surfaces worth different amounts of ranking points, i.e. ATP Masters 1000, ATP 500, ATP 250. In addition, the Grand Slam tournaments also count for ranking points. The tournaments differ in size of the draw, surface, location, and when during the year the tournament takes place. Along with the tournament itself, the attributes of a player and their opponent in a match also factor into his result at a tournament. The current ranking system has remained in place since 2009 and measures player performance over the long term, but we also want to capture short-term fluctuations in player performance. In this paper, we develop a hierarchical Bayesian model to predict player performance that takes these types of factors into account, using public data made available by Jeff Sackmann and Tennis Abstract (Sackmann, 2021). We examine the effects of tournament and player attributes on performance on the tour from 2009-2020. These models provide insight into how match-level, tournament-level, and player-level effects interact to influence player performance.

Introduction

The ATP World Tour features tournaments throughout the year that are held around the world where male professional tennis players compete for ranking points and prize money. Tournaments can have draw sizes ranging from 28 to 128 and are played on different surfaces: hard court (indoor and outdoor), clay, and grass. Often, tournaments played on the same surfaces and in close proximity together are grouped in certain months of the year. For example, clay-court tournaments take place in Europe in March and April while the "US Open Series" (this includes the Canadian Masters, the Cincinnati Masters, the US Open, etc.) takes place in North America during the month of August. In addition, the tour consists of events hosted worldwide on different surfaces annually worth different amounts of ranking points, i.e. ATP Masters 1000, ATP 500, ATP 250. Furthermore, the Grand Slam tournaments and the Olympics also count for ranking points.

Many people believe that certain players perform better depending on these tournament attributes. For instance, Rafael Nadal, one of the greatest players of all time, has dominated on clay throughout his career. Recently, Daniil Medvedev, an up-and-coming talent from Russia who has seemingly reached his prime, has won every ATP Masters 1000 played on a hard court that is held after August (Canada, Cincinnati, Shanghai, Paris) as well as the US Open. Thus, we are interested in investigating the factors that contribute to a player's performance at a certain tournament. This also could provide insight into short-term player trends that are not captured in the current ranking system, which is a better measure for long-term performance. A player's ATP rank also may not be a good measure of likely performance compared to other methods (Williams et al., 2021).

Multilevel models have been applied to other sports. For exmaple, Gerber and Craig fit a mixed effects multinomial logistic-normal model to predict baseball performance (Gerber and Craig, 2021). There have also been random effects models used in tennis to predict player performance at grand slams (Gallagher et al., 2021). Existing literature also utilize a Bayesian regression in the context of tennis, specifically for "developing a mapping between the Universal Tennis Rating (UTR) system and the MDP-based handicaps, so that two amateur players can determine an appropriate handicap for their match based only on their UTRs" (Chan and Singal, 2018). In addition, there have been studies that implement Bayesian hierarchical models to predict the probability of winning a point on serve given surface tournament, and match date using serve and return skill which is assumed to follow a Gaussian random walk (Ingram, 2019).

Along with tournament qualities, certain player attributes may also factor into their performance at certain tournaments. As mentioned above, the ATP ranking is a long-term measure of player performance that is updated each week used to determine qualification and seeding for tournaments. Age is another factor; for example, although the "prime" of a tennis player is believed to be their mid-20s, the "Big Three" (Novak Djokovic, Rafael Nadal, and Roger Federer) have been at the top of the rankings while being in their 30s. However, as these three players have gotten older, they have all missed more tournaments due to injury, allowing more of the younger players to win. Finally, height is another factor to consider as taller players like Matteo Berrettini (6'5") and Alexander Zverev (6'6") have had success on the tour in recent years.

In this study, we analyze the effect of tournament and player attributes on player performance. We do this by developing a multilevel model to estimate the effects of certain variables including tournament surface, draw size, level, player ranking, height, and age. We fit this model to a dataset that includes all ATP and Grand Slam matches from 2009-2020 using the `brms` package in R. After fitting these models, we can examine player performances at certain tournaments on the tour as well.

Data

All data in this study was obtained via a Github repository of Jeff Sackmann (Sackmann, 2021). We used ATP match data from 2009-2021 which also included data from Grand Slams, Davis Cup, and the Olympics. This data set consists of 33,877 matches played on the tour where each match has 49 variables. In our analysis, we focus on the following match attributes: tournament, tournament level, tournament date, surface, draw size, and round, and player attributes: height, age, and rank. In addition, we mutate the data set to include a count of match wins for a player at each tournament. We decided to exclude Davis Cup and Olympic matches as players can lose multiple times in these tournaments which complicates our chosen response variable. After removing these matches from our data set, we have observations for 30,532 matches for 759 tournaments with 857 different players.

The mean age for tournament winners on the tour for this data set is 27.25. However, this number may be skewed due to players like Roger Federer and Rafael Nadal who have continued to win matches in their mid 30s and a lack of younger winners in their teenage years. The youngest player to win a tournament during 2009-2020 was Italian teenager Jannik Sinner who won the Sofia Open in 2020 when he was 19. Meanwhile the oldest player to win a tournament during that same period was Roger Federer who won the 2019 Swiss Indoors when he was 38. We can see that the ATP tour has a large range when it comes to the age of its tournament winners.

The mean rank for tournament winners is 27.43. Once again, this statistic is skewed as the rankings for players are bounded by 1. Thus, there are several players who have won a tournament

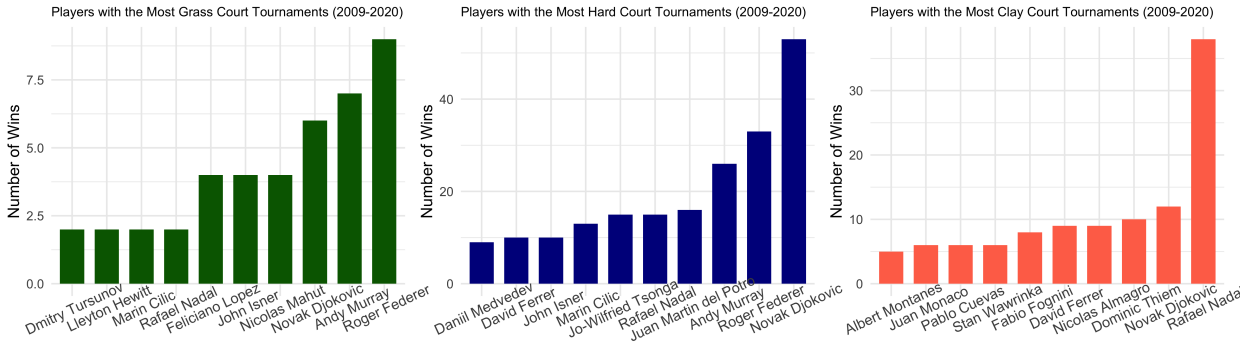


Figure 1: The 10 players with the most titles on each surface from 2009-2020.

while being ranked number 1 in the world. Meanwhile, the worst ranked titlist in this data set was Spaniard Pablo Andujar who won the 2018 Grand Prix Hassan II in Marrakesh while being ranked 355. While no main draw is larger than 128 and qualification for tournaments is based on ranking, there are open slots in the draw for players who are granted wild cards, play in the qualifying rounds of a tournament, or have a protected ranking. While these players seldom win the tournament, it is possible that players like Andujar are able to make a run in smaller tournaments. As with age, we can see that winners on the tour have a large range in terms of their rankings.

While we can analyze the results for the worse ranked players, we focus on the players who consistently play the majority of tournaments throughout the year. More specifically, we look at the "Big Three" who have consistently been top 3 in the ATP rankings in the past decade. In addition, we look at younger players who have started to break through the domination of the "Big Three" including Alexander Zverev and Dominic Thiem. In other words, our model can provide insight into what differentiates the top players on the tour.

Methods

Multilevel structures allow for data to be organized into groups such that coefficients of the model can vary by group. When the data is grouped, the intercept, the slope, or both can vary by group (Gelman and Hill, 2006). For a quantitative response variable y_i belonging to group j with quantitative predictor x_i, a varying-intercept model would have the form

$$y_i = \quad _{j[i]} + \quad x_i + "_i;$$

where $\quad _{j[i]}$ is estimated for each group j. On the other hand, a varying-slope model would look like

$$y_i = \quad + \quad _{j[i]}x_i + "_i;$$

where $\quad _{j[i]}$ is estimated for each group j. Finally, if both the slope and intercept vary by group, the model has the form

$$y_i = \quad _{j[i]} + \quad _{j[i]}x_i + "_i:$$

In the multilevel modeling framework, each level has its own matrix of predictors rather than the regular setup of a data vector $\mathbf{y}$ and matrix of predictors $\mathbf{X}$. In this case, there are two matrices,

one at the group level (indexed by j) and one at the individual level (indexed by i). With a matrix at each level, it is clear what information is available for participants and for items.

In multilevel modeling, modeling takes place at different levels concurrently both within and between individuals. Each level contains clusters that might include variation between observations (in this case, matches) and thus each level has its own sources of variation (De Boeck and Wilson, 2004). The varying coefficients in a multilevel model are termed random effects (Gelman and Hill, 2006). On the other hand, fixed effects are constant across individuals. The fixed effects apply to the population average and separately to each individual (De Boeck and Wilson, 2004).

A model that includes both fixed and random effects is termed a mixed model. The main benefits of multilevel models are that they account for individual and group-level variation, model variation among individual-level regression coefficients, and estimate coefficients for certain groups (Gelman and Hill, 2006). Multilevel modeling can be seen as a compromise between pooling and no pooling models (De Boeck and Wilson, 2004).

Utilizing a mixed-effects approach allows us to analyze fixed effects of variables for every player and random effects at the player-level for different tournaments. Since players compete in different tournaments across different years, players appear in the data set multiple times and the matches are not independent. We can account for this by including a player-level effect. We are also able to examine individual player attributes while analyzing other effects that are assumed to be similar for all players.

One of the difficulties with building a model to analyze player performance on the ATP tour was choosing a response variable. Some potential response variables that we considered, along with identified drawbacks, are listed below:

- $Y_i = 1$ if player $j[i]$ won match i (match level): This is dominated by rank and opponent rank. It is also difficult and computationally expensive to impose a Bradley-Terry type restriction (Bradley and Terry, 1952), i.e. $p_{\text{winner}} + p_{\text{loser}} = 1$, in a multilevel framework.

- Y_i = Total number of wins by player $j[i]$ at tournament $h[i]$: This outcome variable is complicated by draw size.

- Y_i = Proportion of wins by player $j[i]$ at tournament $h[i]$: This variable is equivalent to modeling total wins as in the bullet point above with a Binomial distribution and impossible to be greater than 0 and less than 0.5.

Since tournaments have different draw sizes, taking the maximum of the number of match wins for each tournament was complicated as players need to win 7 matches to win Grand Slams whereas for ATP 250s, a player may only need to win 5 matches to win the tournament. Another complicating factor is the fact that some tournaments include byes for higher seeded players. Thus, it is possible that one player needs to win 5 matches to win a tournament while another player needs to win 6. Due to these complications, rather than using a response variable like match wins, we decided to use an ordered categorical variable: the round that a player reaches in a tournament. In this case, $R_{128} < R_{64} < R_{32} < R_{16} < QF < SF < F$. Although different tournaments do not have the same number of rounds, all players who win the tournament must reach (and win) the final.

Since our response variable is ordered, we implement the adjacent category logistic model. This model utilizes an adjacent category logit which considers the probability of adjacent outcomes m versus $m + 1$ instead of the probability of each category against a baseline. We define the logit as

$$L_m = \log(\frac{p_m}{p_{m+1}})$$

In other words we have that

$$\frac{\log[P(Y_i = m)]}{\log[P(Y_i = m+1)]} = \quad _{j[i]} + \boldsymbol{\beta}_{\boldsymbol{j}[i]}\mathbf{x_i}$$

The model for category probabilities is

$$P(Y_i = m) = \frac{\exp(\quad _{m[i]} + \boldsymbol{\beta}\mathbf{x_i})}{1 + \sum_{k=1}^{c-1}\exp(\quad _{k[i]} + \boldsymbol{\beta}\mathbf{x_i})}; m = 1; : : : ; c-1$$

Here Y_i follows an ordered and adjacent category logit distribution. The outcome is defined where y_i is the round that player $j[i]$ made it to in tournament $h[i]$. There are J players and H tournaments and $I = J \times H$ observations. We can define the logit for each outcome as

$$L_m = \quad _{j[i]} + \quad _0 + \quad _1 \times \text{height} + \quad _2 \times \text{age} + \quad _3 \times \text{rank} + \quad _4 \times \text{draw size} + \quad _5 \times \text{tourney\_level}$$

where $\quad _{j[i]}$ describes the player-level effects:

$$_{j[i]} = \quad _0[j] + \quad _1[j] \times \text{surface}$$

We utilize the R package **brms** to fit complex hierarchical models in a Bayesian framework. **brms** is a package designed to fit Bayesian multilevel models using Stan for Bayesian inference (Bürkner, 2017). Stan is a C++ package that performs Bayesian inference (Stan Development Team, 2018). **brms** uses Markov chain Monte Carlo to draw random samples from the posterior distribution. Stan utilizes Hamiltonian Monte Carlo and No-U-Turn Sampler algorithms to obtain random samples. **brms** supports an extensive range of distributions and link functions, thus allowing us to support different types of models in a multilevel context (Bürkner, 2017).

Results

We examine the effects of player and tournament attributes on player performance using open-sourced ATP data provided by Jeff Sackmann (Sackmann, 2021). We can use this data to visualize how players perform on certain surfaces or at specific tournaments. This allows us to examine how player performance varies given certain conditions (i.e. grass court tournaments, or ATP Masters tournaments). We then develop a Bayesian hierarchical model to examine player and tournament effects. Specifically, we analyze how a player's height, age, and rank, as well as the tournament's surface, level, and draw size affect how far a player advances in a tournament.

The output from **brms** provides us with an estimate for each parameter along with a 95% credible interval. The rest of the output allows us to measure the convergence of the chains and effective sample size. We can see from the estimates that the player attributes have effects with small magnitudes compared to the tournament attribute effects. We can also see that there is a small difference between Intercept[5] and Intercept[6] which correspond to the semifinals and finals. Meanwhile there is a large difference between Intercept[3] and Intercept[3] which correspond to R64 and R32. The `plot` function within **brms** allows us to see the densities of the parameter estimates as well as whether the chains have mixed well as seen in Figure 2. Finally, the `conditional_effects` function provides us with a means of visualizing player or tournament effects.

From our adjacent category model, we can see that age has a positive effect on making it far in a tournament, ranking appears to have a negative effect, and height does not have much effect. Out of these three, ranking makes the most sense as players with a higher ranking (lower numerically) are predicted to have more success given that they have earned ranking points by advancing far in

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
Intercept[1]	-1.9664584	0.0605542	-2.0845771	-1.8481771	0.9998626	9611	7152
Intercept[2]	-1.8512817	0.0570019	-1.9606586	-1.7377080	1.0006818	9008	7211
Intercept[3]	0.5360168	0.0558105	0.4271551	0.6474253	0.9999514	8320	6924
Intercept[4]	0.5739595	0.0587819	0.4611121	0.6924696	1.0001658	8664	6871
Intercept[5]	0.7521819	0.0621800	0.6319821	0.8771539	0.9998953	9182	6890
Intercept[6]	0.7787769	0.0708867	0.6401681	0.9195043	1.0001354	10198	6731
ht	0.0004423	0.0001869	0.0000729	0.0008117	1.0006602	4849	5905
age	0.0161770	0.0017456	0.0127911	0.0196814	0.9996891	9661	7632
rank	-0.0010985	0.0000623	-0.0012209	-0.0009787	1.0004391	7396	6536
tourney_levelF	-0.2324100	0.0694036	-0.3672509	-0.0989649	1.0011214	14325	6172
tourney_levelG	-0.5924190	0.0422261	-0.6757170	-0.5105177	1.0002224	7178	6077
tourney_levelM	-0.5374702	0.0214396	-0.5797027	-0.4971031	1.0008568	6873	6024
draw_size	-0.0196222	0.0004285	-0.0204629	-0.0187784	1.0001656	8677	6409

Table 1: The **brms** output for the fixed effects of our model.

previous tournaments. We can see from our conditional effects plot that a lower numerical rank corresponds to advancing to further rounds in a tournament. In our model, since age positively predicts a player performing well at a tournament, we expect older players to perform better than younger players. This phenomenon can once again be explained by the "Big Three" who have combined to dominate the last decade on the tour. While it may initially seem strange that height does not have a notable effect but not too surprising as tennis is not a sport dominated by especially tall athletes. Additionally, the breakthrough of tall players (above 6'4") winning tournaments is a relatively recent occurrence.

We can analyze the variation in random player effects by extracting the parameter effects from out model using the `ranef()` function. This gives us 3 matrices, one for each surface. Each matrix has 857 rows for each player and 4 columns: estimate, estimated error, the lower bound for the 95% credible interval, and the upper bound. By exponentiating the parameter estimate and plotting 95% credible intervals, we can compare players and their odds ratios for each surface. Looking at the odds ratio of the only Grand Slam winners, we notice these players have quite high odds. Obviously the "Big Three" have higher odds compared to the other Grand Slam winners as they have dominated the tour for the past decade. Comparing the odds ratio for the higher ranked "younger players", we notice that Stefanos Tsisipas (2019 World Tour Finals Champion) has quite a high odds on clay. This could be due to his recent breakout on tour in 2019. We can also notice that the odds ratios for players are lower on grass and hard courts since the intercept (clay) includes both the clay effect and the overall player effect.

Finally, we can look at the group-level effects in the **brms** output. Here, the intercept is the clay court. We can observe that the clay surface effect has a much higher magnitude compared to the other two surfaces. **brms** also includes the correlation estimate between pairs of group-level effects. Looking at these correlations confirms what is believed about court surfaces, i.e. the fast surfaces (grass and hardcourt) are similar and different from the slowest surface which is clay. Grass, especially is quite different from clay. This makes sense as points are quite different when played on clay compared to when they are played on grass. On clay, points are longer and require more defensive skills whereas on grass, points are shorter and favor a more offensive player. Thus, success on each surface requires quite different play styles and is why the the success of the "Big Three" on all surfaces is impressive.

brms Plots

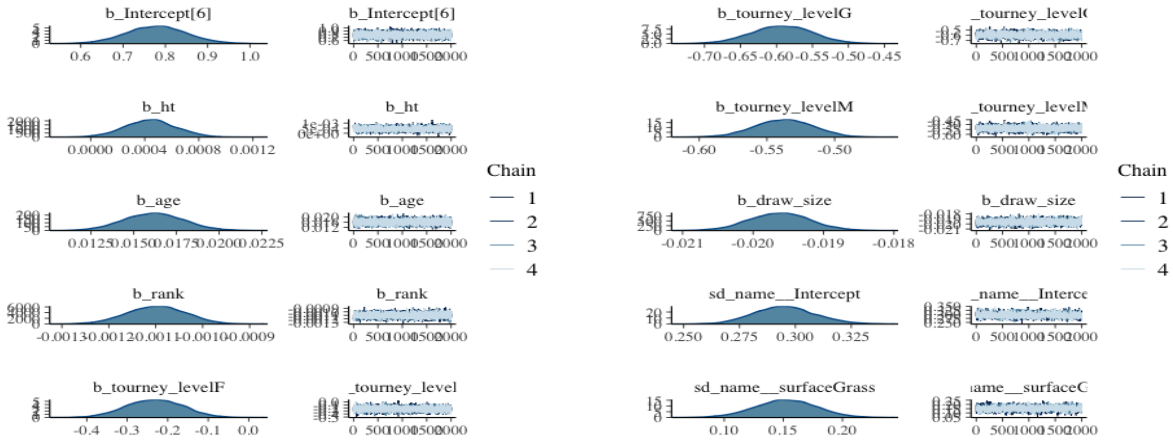
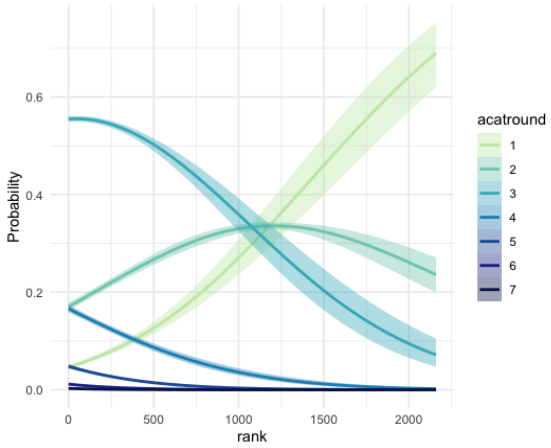


Figure 2: Densities and trace plots for model parameters (via brms `plot()` function), indicating model convergence.

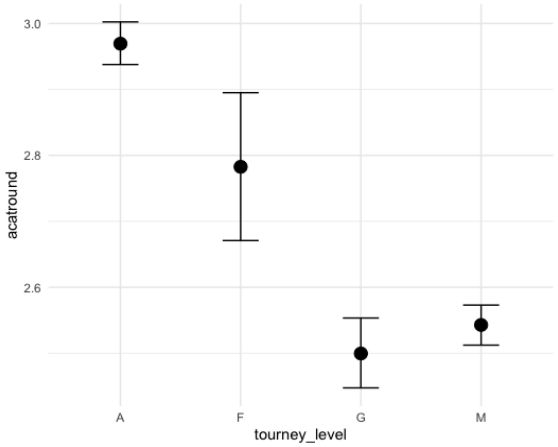
	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
sd(Intercept)	0.2953049	0.0135824	0.2699439	0.3228389	1.000805	2559	3493
sd(surfaceGrass)	0.1521500	0.0257070	0.1009082	0.2020722	1.001116	1832	2722
sd(surfaceHard)	0.1304349	0.0168292	0.0975973	0.1636845	1.000664	1337	2473
cor(Intercept,surfaceGrass)	-0.6229347	0.0904779	-0.8007497	-0.4440428	1.000596	3613	3390
cor(Intercept,surfaceHard)	-0.3089157	0.0833825	-0.4628859	-0.1352975	1.000250	4545	5722
cor(surfaceGrass,surfaceHard)	0.8604382	0.0809747	0.6567168	0.9650497	1.004242	1135	1634

Table 2: The **brms** output for the standard deviations and correlations of the random effects of our model. In this case "Intercept" corresponds to clay.

Conditional Effects Plots



(a) Conditional effects for rank on the response variable.



(b) Conditional effects plot for tournament level on the logit. "A" corresponds to ATP250s and ATP500s, "M" to ATP1000 Masters, "F" to the ATP World Tour Finals, and "G" to Grand Slams.

Figure 3: The `conditional_effects()` function within **brms** outputs these two plots.

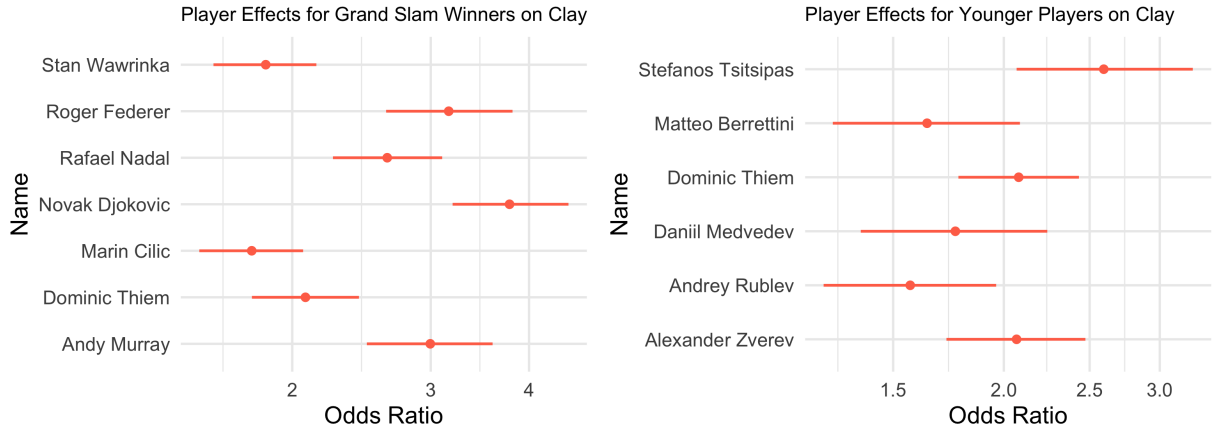


Figure 4: Odds ratio of the random effect for a selection of players with 95% credible intervals on clay.

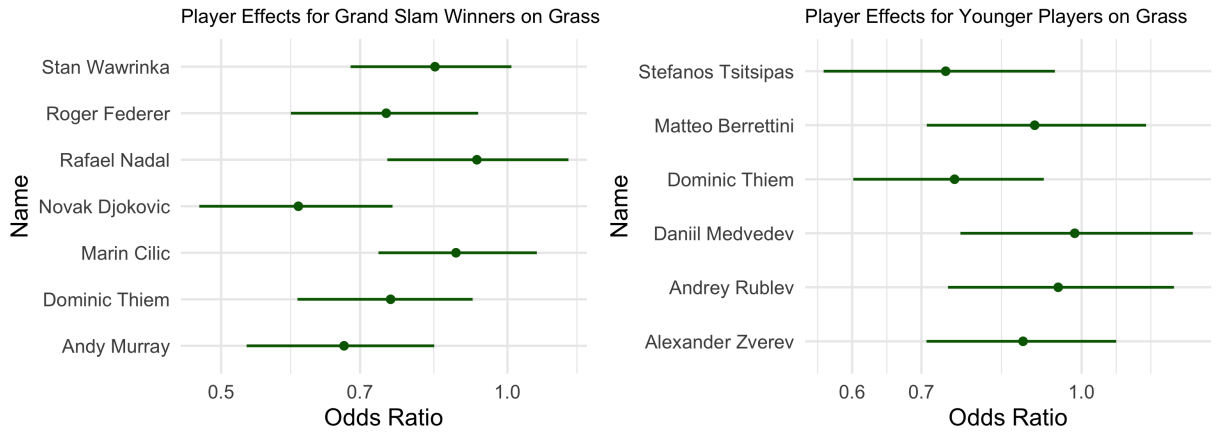


Figure 5: Odds ratio of the random effect for a selection of players with 95% credible intervals on grass.

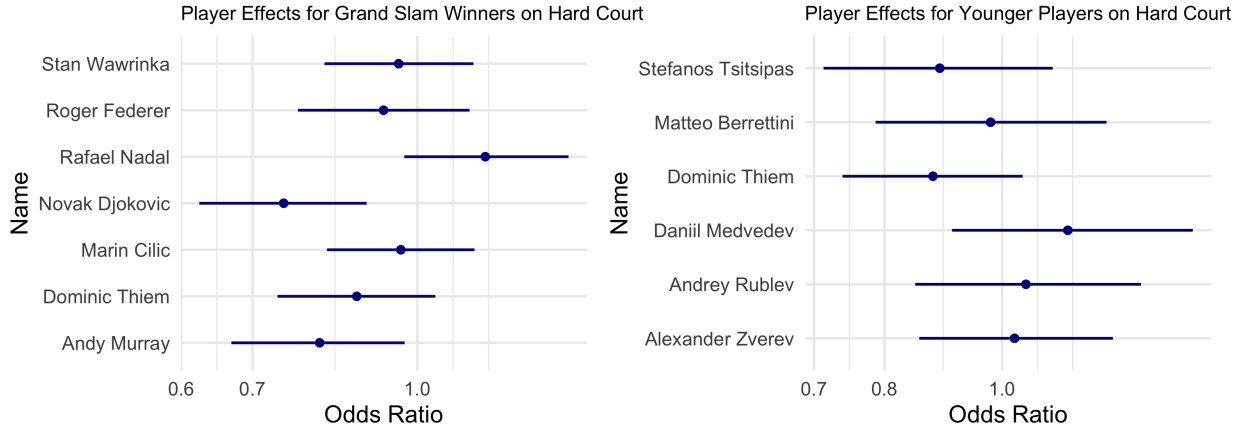


Figure 6: Odds ratio of the random effect for a selection of players with 95% credible intervals on hard court.