Modeling Tennis Match Outcomes for Grand Slams

Sahaja Danthurthy, Ryan Kim, Alyssa Robert, Adarsh Suresh

Introduction

Tennis is a uniquely independent sport with every match acting as a 1-on-1 test for a player. The two major tennis organizations, the ATP (men's tennis) and WTA (women's tennis) have a ranking system that dictates the seeds that players come into when playing tournaments. However, the ranking system does not tell us the full story since there are so many other factors that decide a player's success in a tournament like their expertise on a given surface, the match-up against a given opponent, a player's skill set, and so much more. Due to the number of factors used in determining the outcome of a tennis match, we aimed to tackle the question "What contributes to how close tennis matches are?".

After our analysis, we determined that player rankings do not strongly predict set differential outcomes in the ATP data we used. However, we saw that the rankings have a stronger relationship with game differentials within sets rather than the set differentials. This suggests that players may adjust their effort based on how a set is going to conserve energy or overcome fatigue. The type of surface and handedness of a player also were important factors that influenced how close games were. Overall, the game differential model provided much more useful insights into player performance and consistency rather than the set differential model.

Data

For this project, we used match-level tennis data that was retrieved from Jeff Sackmann's GitHub repository. This repository contains detailed records for both ATP (men's) and WTA (women's) tournaments. For this project, we specifically focused on ATP singles matches from 2000 to 2024. In this dataset, each row represents a single match played during a tournament and includes information such as player names, rankings, match scores, and match details. The response variable we chose for our analysis is the difference in the number of games won between the winner and the loser of each match.

To prepare the dataset for analysis, we first concatenated all of the individual years of ATP singles data together using R. After combining the data, we filtered the dataset to include only matches where the total number of sets played was exactly five. After analysis, we noticed that this left us with mostly just Grand Slams from all the years in the dataset. This step was necessary because it ensured consistency in the structure of the matches since most other ATP tournaments are best-of-three-set matches. We also created a new calculated column that represents the game differential by subtracting the number of games won by the losing player from the number of games won by the winner.

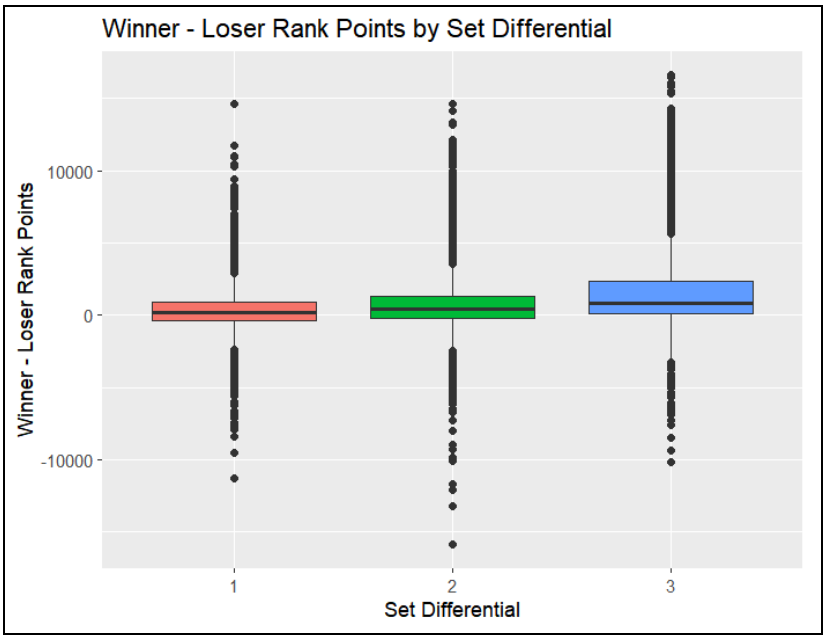

Figure 1: This graph shows the set differential (our initial response variable) compared to the difference in points (ranking points) between the winner and loser.

After processing our data, we conducted an EDA on the data to gain some initial information about the trends in the dataset. As we notice, there is a good amount of overlap between all three intervals showing that the set differential does not initially trend as heavily impacting the difference in rank points between a winner and loser.

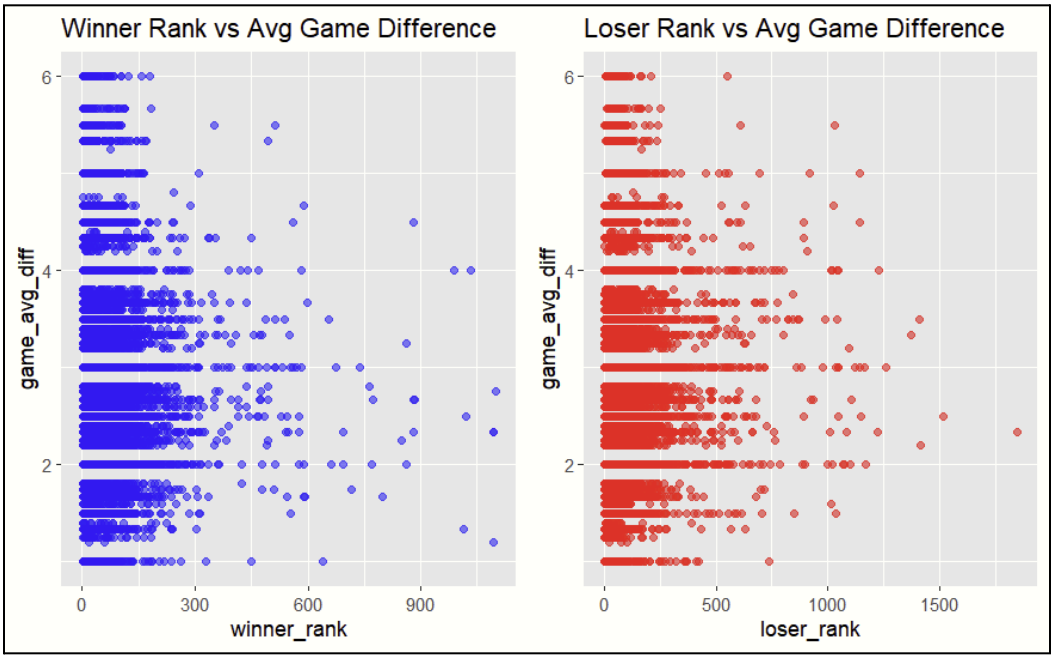

Figure 2: This graph shows the set differential (our initial response variable) compared to the difference in points (ranking points) between the winner and loser.

High ranking (lower rank numbers) players tend to dominate both the high and low average differential rankings, while worse players tend to be towards the middle.

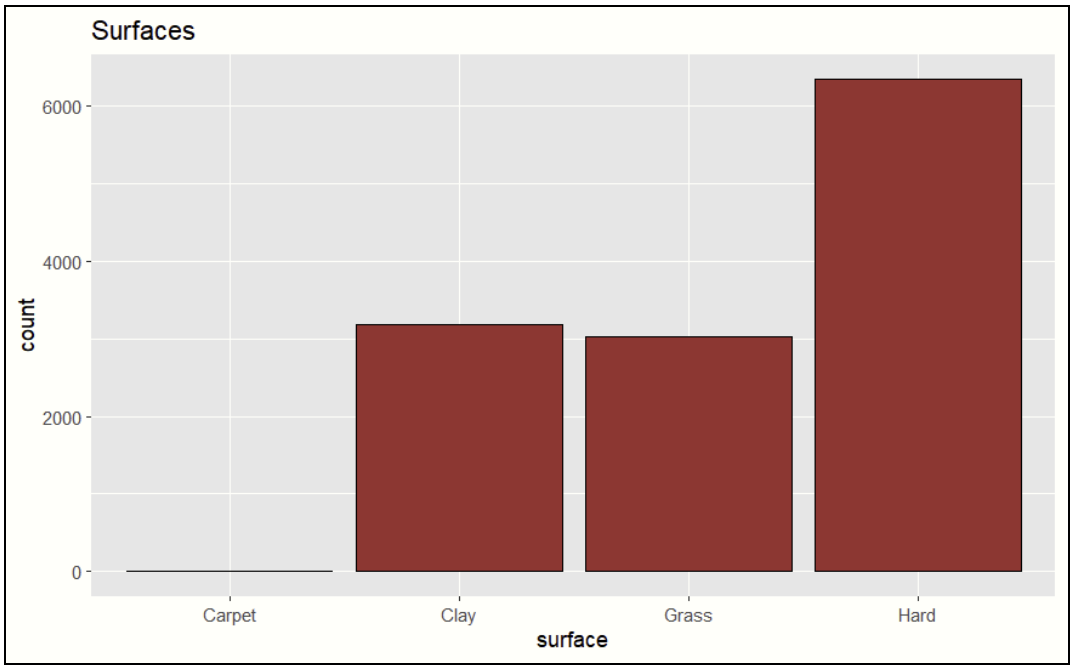

Figure 3: This graph illustrates the breakdown of the surfaces played on amongst the tournaments in this dataset.

The very small fraction corresponding with ‘Carpet’ is due to a singular tournament that used carpet from 2000-2009 when it was outlawed. The rest corresponds accurately with grand slams. Since there is very little data with matches on carpet courts, in our second model we decided to eliminate those matches.

Methods

The model we used for our project was a Bayesian Hierarchical Gaussian Model, which is based on the models discussed in Lecture 18: Modeling Team Ratings and Posterior Predictions. This differs slightly from the model in the demo because instead of assuming that the response variable comes from a Poisson distribution, we assume that it comes from a (Normal) Gaussian distribution. The features with fixed effects were the surface type, winner in-game statistics such as winner aces, winner double faults, winner serve points, winner 1st/2nd serve stats, serve games, and breakpoints, the same in-game statistics for the match loser, player ranks coming into the match, as well as the players' heights and ages. The features with random effects were the name of the player and the handedness of the player, essentially allowing for player-level skill variations in the model.

Bayesian models follow the same 4 assumptions as linear regression: linearity, independence, homoscedasticity, and normality. In addition, we assume that the outcome variable, game differential, itself comes from a Normal distribution. Based on the description of the variable from earlier, this assumption is valid due to both the Central Limit Theorem (we are averaging game differences across sets) and Figure 4 below, which shows an approximately normally shaped curve. The model we have written out is valid because not only does it fold in the fact that in a Bayesian setting, we would expect game differential to be Normally distributed, but we also assume that the game result is a function of external factors like a player's intangible physical skills (height and age) in addition to the player's skill

itself. Using a Bayesian model also gives us posterior distributions and will allow us to quantify uncertainty easily.

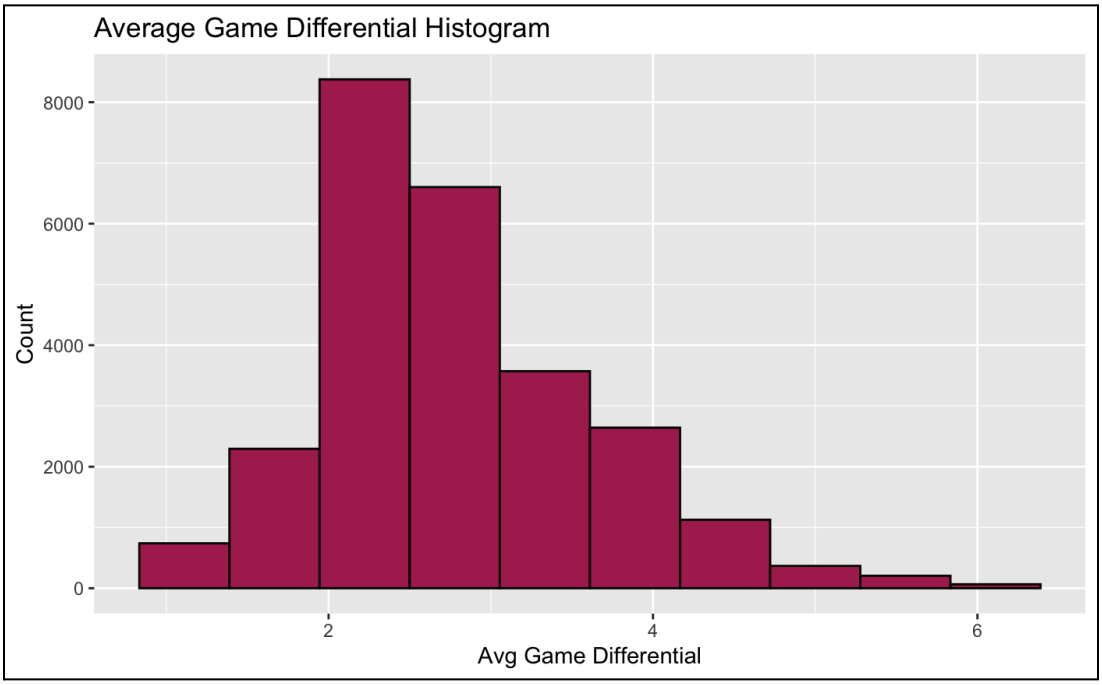

Figure 4: This graph shows the distribution average game differential among our data.
As we can see here, the distribution follows an approximate Bell curve shape, which helps justify the Normality assumption for our Bayesian hierarchical model.

In addition to game differential, we also look at set differential as a potential measure of match outcomes. Similar to the model above, we take the exact same fixed and random effects but try to model the number of sets won by the winner using a Bayesian Hierarchical Negative Binomial Model instead of a Gaussian Model (since set differential falls in the range 1-3 typically). We also did not use a Poisson model because there's no particular reason for the mean and variance to be equal. Though we considered using a multinomial model, we decided against it because we wanted to preserve the hierarchical structure, since set differential is a sum across games and points won (which is continuous rather than categorical), and we weren't too sure about how to implement it. In evaluating the coefficients, posterior distributions, and diagnostics, we can gauge whether our specified model for game differential is in fact the ideal statistical method. We will also use the same posterior predictive checks discussed in lectures/demos to evaluate the performance of our model.

To quantify uncertainty, we will look at the Epistemic uncertainty regarding our parameters. In particular, we can plot the whole posterior distribution for each player and judge whether or not our model finds significant player-level effects by comparing the amount of overlap (if any) between the top and bottom-performing players. This makes sense in the context of tennis since one of our main goals is to isolate the player-level effects and see if their individual skill level plays a significant role in evaluating match outcomes.

Results

After running both the set and game differential model, we checked diagnostic plots for both and confirmed that our R-hat values were not noticeably greater than one and that our MCMC trace plots did not have any concerning trends or distributions.

For the set differential model, we looked at the coefficients by both winning and losing players but did not see any noticeable differences between the players with the top 10 and bottom 10 players' mean coefficient posteriors, which is evident by the median lines overlapping for both the top 10 and bottom 10. Thus, this model led us to conclude that rank does not help that much in predicting set differentials.

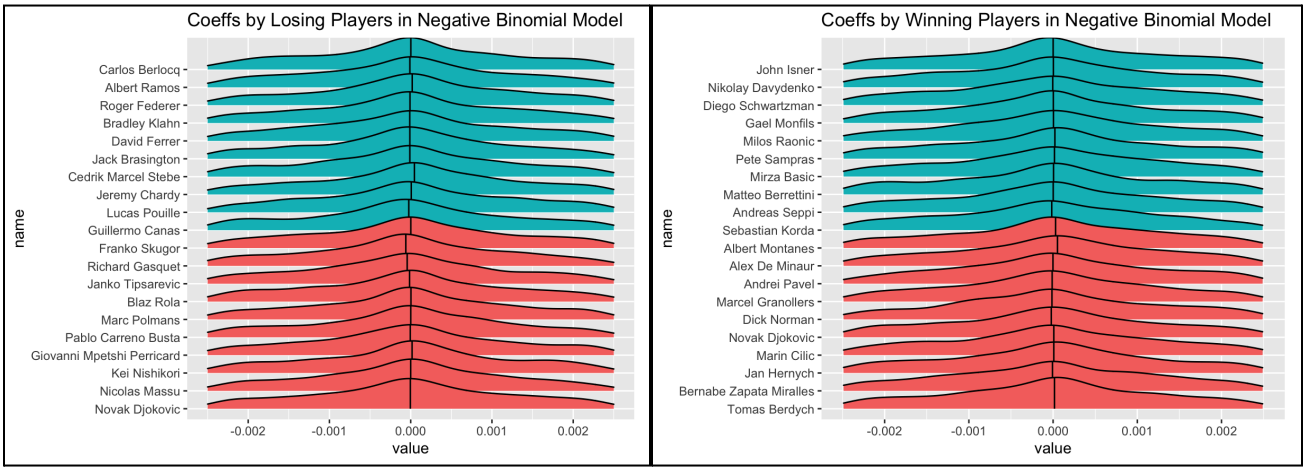

Figure 5: These graphs show top and bottom 10 mean player coefficient posteriors for game score differences in our Negative Binomial Model based on set differences with median lines shown.

For our game differential model, we looked at the coefficient distributions of both the random and fixed effects to compare input terms. In all of these models, positive coefficients indicate players lose by larger margins with the effect present, and negative coefficients indicate that they lose by smaller ones. For random effects, we see that players in both the top and bottom ten tend to have fairly high average rankings (top 100), although there appear to be slightly higher-ranked players in the bottom 10 of the losing player coefficients and winning player coefficients. Negative coefficients for the losing players and positive coefficients for the winning players indicate consistency, and vice versa for inconsistency since consistency is either losing by less or winning by more on average.

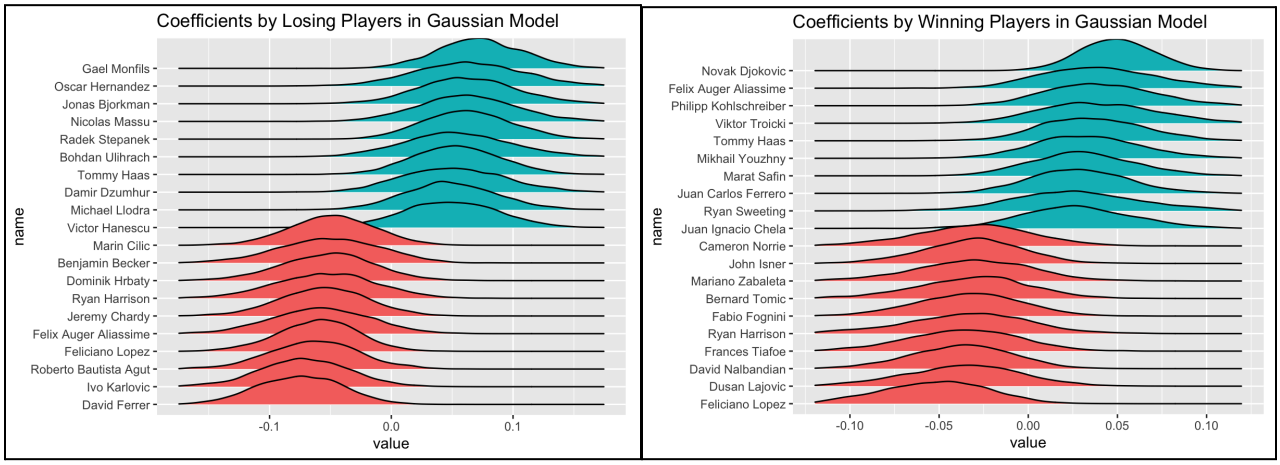

Figure 6: These graphs show top and bottom 10 mean player coefficient posteriors for game score differences in our Gaussian Model based on average game differential within a set.

We can see that overall, the player does not make a large impact on how big their margin of victory will be, but there is still a noticeable difference between the least and most consistent players, although there is still some noticeable overlap between the top 10 and bottom 10 distributions.

For the fixed effects, it looks like winner/loser heights and ages do not have very different posterior distributions. Older players lose by a little bit more than younger players. However, surface type plays a notable role in the margin of victory. Although their distributions are much wider, players on hard surfaces tend to have closer games than the reference surface, clay, while players on grass surfaces tend to have wider margins in their games. Additionally, handedness also has wider margins, with right-handed winners having much lower game score differences than left-handed winners, while right-handed losers have higher game score differences. This may be due to the unfamiliarity that left-handed players bring.

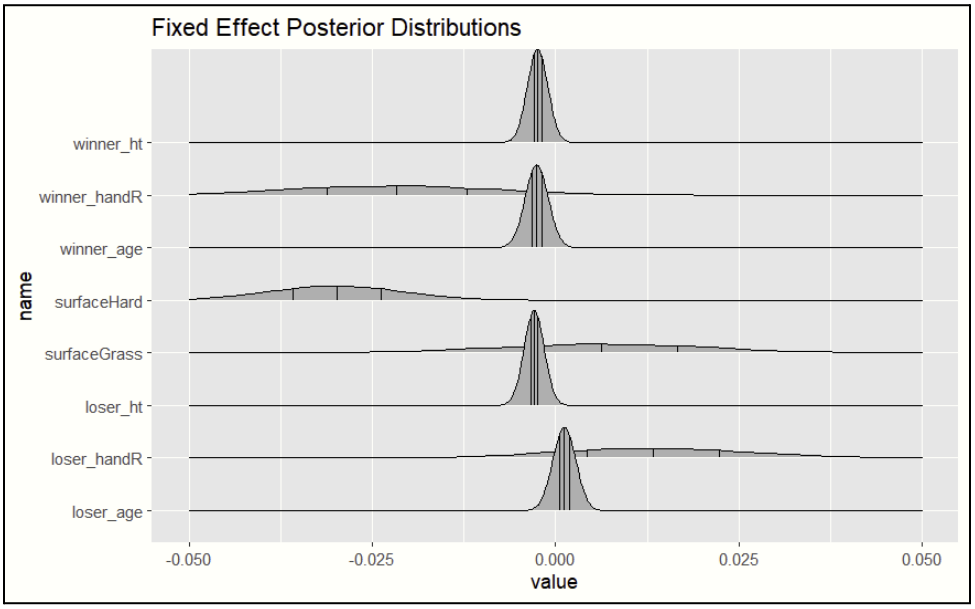

Figure 7: These are the graphs of the handedness random effects and the fixed effects from the hierarchical bayesian model. Many of the distributions appear fairly similar, except for the surfaces and handedness of players, which are much wider, with notably different medians..

Discussion

In conclusion, the game differential model provided stronger insights than the set differential model, although both showed that rankings do not impact performance as much as people think they do. Additionally, while rankings do not seem to predict set differential, they are better at predicting game differential within a set, indicating that when players are winning or losing already, they are more likely to throw away games to conserve energy when they already are assuming they will lose the whole set. This is clearly more evident in game differential within a set, as the results corroborate. Tennis rankings are a cumulative sum of performances over the last 12 months, so the points accrued by a player are a measure of their performance coming into a tournament. However, this does not account for in-tournament performance and momentum which could explain why rankings have a smaller impact on set differential than expected.

In terms of player coefficients, we can see that the game differential model provided much more insight into player-by-player performance than the set differential model. In particular, Figure 5 shows a lot of overlap between the best and worst-performing players, while Figure 6 (game differential model) shows a lot more variance between players. This makes sense because, across games, we have more data to be able to separate the performance across individual players, whereas, across sets, we only have 3 possible outcomes (1, 2, or 3) for any given match. Therefore, we have a wider range to be able to differentiate between players, even though their posterior distributions between the 10 best/worst players overlap slightly.

One limitation of our study is that we only looked at men's tennis matches, meaning that our analysis does not look at any WTA data. Therefore, our analysis and features of importance may not translate into women's tennis so further exploration would be needed there. Additionally, we took data from 2000 to 2024, but we did not account for a year as a factor in our models. Several aspects of the sport, from courts (carpet courts were used before but became banned in 2009) to the increased level of athleticism and improved racquets have changed fundamentally in the last 25 years so it is possible that some of these factors are confounding variables in our analysis.

Some next steps could be to do a more granular analysis, and potentially explore some interaction terms or find a way to weigh recent performance more heavily. Alternatively, we could look at incorporating tracking data to look at these stats on a point-by-point basis, perhaps seeing what makes rallies longer or shorter.