Martin Ingram*

A point-based Bayesian hierarchical model to predict the outcome of tennis matches

https://doi.org/10.1515/jqas-2018-0008

Abstract: A well-established assumption in tennis is that point outcomes on each player's serve in a match are independent and identically distributed (iid). With this assumption, it is enough to specify the serve probabilities for both players to derive a wide variety of event distributions, such as the expected winner and number of sets, and number of games. However, models using this assumption, which we will refer to as "point-based", have typically performed worse than other models in the literature at predicting the match winner. This paper presents a point-based Bayesian hierarchical model for predicting the outcome of tennis matches. The model predicts the probability of winning a point on serve given surface, tournament and match date. Each player is given a serve and return skill which is assumed to follow a Gaussian random walk over time. In addition, each player's skill varies by surface, and tournaments are given tournament-specific intercepts. When evaluated on the ATP's 2014 season, the model outperforms other point-based models, predicting match outcomes with greater accuracy (68.8% vs. 66.3%) and lower log loss (0.592 vs. 0.641). The results are competitive with approaches modelling the match outcome directly, demonstrating the forecasting potential of the point-based modelling approach.

Keywords: Bayesian modelling; random walk; sports forecasting.

1 Introduction

A wealth of research in tennis, such as Newton and Keller (2005), O'Malley (2008) and Riddle (1988) has shown that if it is assumed that point outcomes on each player's serve in a tennis match are assumed to be independent and identically distributed (iid), the probability of winning a match can be determined from just the probabilities of winning a point on serve for each player. In addition to the match-winning probability, probabilities of winning a service game and winning a set with a given score can be derived from just the serve-winning probabilities. Barnett et al. (2006) derive a wealth of other metrics, including the number of points in a game, tiebreak, set and match.

Although many predictions can be made using the iid assumption, it is only an approximation. Pollard, Cross, and Meyer (2006) analyse patterns of set wins and find evidence that the probability of winning a set varies from set to set. Klaassen and Magnus (2001) analyse 90,000 points from Wimbledon matches between 1992 and 1995 and find deviations from iid behaviour. They suggest however that these deviations are small and that the iid model may still be useful for match prediction. Similarly, Newton and Aslam (2006) investigate a variety of possible non-iid effects using a Monte Carlo simulation and find that the iid model remains robust even when non-iid effects are introduced.

In Kovalchik (2016), the author compares 11 published tennis prediction models by predicting the ATP's 2014 season. Models are broken into three classes: models using the iid model for match prediction, which we will refer to as "point-based models" from now on; models based on regression approaches; and paired comparison models. The best point-based model was found to have lower accuracy and higher log loss than the best regression and paired comparison models.

This paper introduces a new point-based model. The paper's contributions are the following: firstly, it improves on the previous best published point-based model and outperforms the regression models in Kovalchik (2016), coming close to matching the best reported model, an Elo model (Elo 1978) with a customised k-factor devised by the website FiveThirtyEight (Morris and Bialik 2015). Secondly, to the best of our knowledge, it is the first Bayesian hierarchical model presented for predicting tennis matches. The hierarchical model allows the fitting of player and surface-specific skills, as well as tournament-specific adjustments, even when there are few or no observations for some combinations.

***Corresponding author: Martin Ingram,** University of Melbourne, School of BioSciences, Melbourne, Australia; and Silverpond, Melbourne, Australia, e-mail: martin.ingram@gmail.com

2 Methods

2.1 Data

To be able to compare directly with the results obtained in Kovalchik (2016), the ATP's 2014 season is used as a validation set.

Table 1 shows summaries of this validation set, broken down by surface. The dataset was obtained by scraping MatchStat.com. Retirements, walkovers and matches without serving statistics (total points played and total points won on serve for each player) were discarded. Surfaces are broken down into three categories: clay, grass and hard. Hard court matches are most prevalent, representing more than half of the tournaments and matches played, followed by clay courts, followed by grass courts, which only comprise about 10% of total matches played.

Not all players play on all surfaces, with only just more than half participating in at least one grass court match. On average, a player serves 80 points per match, winning 51 of these, or about 64%. This percentage is higher on grass courts (67%) and lower on clay courts (62%). Grass court matches average more service points played, which is likely due to the fact that Wimbledon, which uses the longer best of five sets format, contributes over half of the matches played on grass in the dataset (124).

2.2 Model

2.2.1 Likelihood

In every tennis match, each player serves a number of times n and wins y of those points. We divide each match in the dataset into these two contests on serve, referring to each as a "serve-match". The likelihood for each serve-match i given by:

$$y_i \sim \text{Binomial}(n_i, \theta_i) \tag{1}$$

where θ_i is the serve-winning probability in that serve-match. This assumes that the outcome on each of a player's n_i serves in the match is the result of a Bernoulli trial with the same success probability θ_i, thus making use of the iid assumption.

Table 1: Summaries of the 2014 validation set.

	Clay	Grass	Hard	Overall
Tournaments	22	5	33	60
Matches	736	232	1240	2208
Unique players	186	144	215	267
Average serve points played	78	95	79	80
Average serve points won	48	64	51	51
Fraction of serve points won	62%	67%	64%	64%

In the following, we also divide time into periods. Players' skills are assumed to be constant within a time period. Shorter periods allow the model to adapt more quickly to skill changes but also require a larger number of parameters to be estimated. The shortest period length considered in this paper is one month, and the largest is twelve.

The serve-winning probability θ_i is further broken down as follows:

$$\text{logit}(\theta_i) = (\alpha_{s(i)p(i)} - \beta_{r(i)p(i)}) + (\gamma_{s(i)m(i)} - \gamma_{r(i)m(i)}) + \delta_{t(i)} + \theta_0 \tag{2}$$

Here, the quantities α, β, γ, δ and θ_0 represent the following:

- $\alpha_{s(i)p(i)}$ is server $s(i)$'s serving skill in period $p(i)$
- $\beta_{r(i)p(i)}$ is returner $r(i)$'s returning skill in period $p(i)$
- $\gamma_{s(i)m(i)}$ is server $s(i)$'s additional skill on surface $m(i)$
- $\gamma_{r(i)m(i)}$ is returner $r(i)$'s additional skill on surface $m(i)$
- $\delta_{t(i)}$ is the adjusted serve intercept at tournament $t(i)$, and
- θ_0 is an intercept representing the average player's probability of winning a point on serve.

Breaking down equation (2) into its individual terms, the first term in brackets represents the server's serving skill adjusted by the opponent's return skill, the second represents the difference in skill preferences for the match surface, and the third corrects for tournament variation in the difficulty of winning a point on serve.

Equation (2) builds on the opponent-adjusted model by Barnett and Clarke (2005). They also adjust a player's serve skill by the opponent's return skill and have a tournament-specific intercept, but do not add a surface-specific offset.

2.2.2 Priors

The initial serve and return skills α and β are drawn from a normal distribution:

$$\alpha_{.1} \sim N(0, \sigma^2_{\alpha 0}) \tag{3}$$

$$\beta_{.1} \sim N(0, \sigma^2_{\beta 0}) \tag{4}$$

The initial variance parameters are given hyperpriors:

$$\sigma_{\alpha 0} \sim H(0, 1) \tag{5}$$

$$\sigma_{\beta 0} \sim H(0, 1) \tag{6}$$

where H is the half-normal distribution which assigns non-zero probability only to positive values. This hierarchical prior shrinks players' skills towards the group serve and return priors given by equations (3) and (4).

Similarly, the tournament-specific intercepts and surface-specific preferences are independently drawn from zero-centred normal distributions with half-normal variance hyperpriors:

$$\gamma_{..} \sim N(0, \sigma_\gamma^2) \quad \sigma_\gamma \sim H(0, 1) \tag{7}$$

$$\delta_{.} \sim N(0, \sigma_\delta^2) \quad \sigma_\delta \sim H(0, 1) \tag{8}$$

Finally, the serve and return skills are assumed to follow a random walk over time:

$$\alpha_{.p+1} \sim N(\alpha_{.p}, \sigma_\alpha^2) \tag{9}$$

$$\beta_{.p+1} \sim N(\beta_{.p}, \sigma_\beta^2) \tag{10}$$

$$\sigma_\alpha, \sigma_\beta \sim H(0, 1) \tag{11}$$

Thus, the random walk is assumed Gaussian, with the same variance between periods for each player. This random walk assumption is also made in Glickman (1999) to derive the Glicko paired comparison model.

In the section above, we used unit half-normal distributions as hyperpriors for all variances. Two alternatives were considered: a flat prior giving equal probability to all positive values, and the inverse-gamma distribution, which is often used due to it being the conjugate prior for a normal observation model with known mean but unknown variance (Gelman et al. 2013). We prefer the unit half-normal prior to the flat prior, since large values for variances are unlikely a priori. For instance, serve-winning probabilities above 80%, or 1.38 on the logit scale, are rarely observed, suggesting that the standard deviation in player skills is unlikely to be much larger than this value. The unit half-normal prior expresses this preference for smaller values. The inverse-gamma prior could also be used to assign greater probability to smaller values; however, it places zero mass on variances of zero, which is not always desirable. For instance, the random walk standard deviations σ_α and σ_β may be zero if there is no evidence that players' skills do not vary over time, suggesting that this should not be ruled out.

2.2.3 Model fitting and prediction

The model is fit using the NUTS Hamiltonian Monte Carlo sampler implemented in Stan (Carpenter et al. 2016). To compare the influence of the period length, separate models with period lengths of 12, 6, 3, 2 and 1 month(s) are fit. In addition, for each period length, model fitting is started in 2013, 2012 and 2011 to compare performance. Convergence was assessed using the $\hat{R}$ statistic (Gelman and Rubin 1992), which is reported by Stan; all values were below 1.1. Code to fit the model is available online at: https://github.com/martiningram/tennis_bayes_point_based.

To predict each period of 2014, the model is fit with all data up to but excluding that period. The 4000 posterior samples returned by Stan are used to calculate θ_i for each serve-match, and win expectations are calculated by averaging the iid win expectation obtained by each pair of posterior draws for the serve-winning probabilities of the players involved. Skills for unseen players and tournaments are drawn from their group level prior distributions so that predictions are made for all matches.

For the 2013 start, the 12 month period was skipped, since this would only give the model a single period to fit, making the estimation of period-to-period variation σ_α and σ_β impossible.

2.3 Evaluation

2.3.1 Metrics

The model is evaluated on four metrics: accuracy and log loss for match prediction, and root mean square error (RMSE) on the sum and difference of serve-winning probabilities of each match.

At the match level, accuracy was chosen because it is an intuitive measure, while log loss gives a better sense of the quality of the model's probability estimates.

As shown by Klaassen and Magnus (2003), the match win probability in the iid model is driven almost entirely by the difference of serve-winning probabilities in a match, which we will refer to as $\theta_i - \theta_j$, where θ_i is a player's serve-winning probability in the match, and θ_j is their opponent's. It should be noted that although only serve-winning probabilities are considered, this does not mean that return skills are unimportant; as equation (2) shows, the probability θ_i is assumed to be a function of both player i's serving skill and their opponent j's returning skill, so the sum and difference will involve both players' return and serve coefficients. For in-play forecasts at unequal scores and at later stages in a match, the sum $\theta_i + \theta_j$ becomes important. In addition, Barnett and Clarke (2005) suggest that match length is driven by the sum

of serve-winning probabilities. Because of their different significance in the iid model, the quality of sum and difference predictions are analysed separately.

2.3.2 Baseline models

The model is compared against the best-performing point-based model found in Kovalchik (2016), which is the opponent-adjusted model presented in Barnett and Clarke (2005), as well as the best-performing model overall, FiveThirtyEight's Elo model, which uses a custom non-linear function of the number of matches played by each player to calculate its k-factor (Morris and Bialik 2015).

The opponent-adjusted model's predictions were made by considering the last year of matches leading up to each match, and the tournament average serve-winning probability was calculated by averaging the serve-winning probabilities of all matches played at that tournament in the previous two years. The Elo model was fit using the functional form of the k-factor suggested by the authors and with the fit beginning at the start of the Open Era (1968).

It is worth noting that the Elo model uses the same link function as the model presented in this paper. In the Elo model, the likelihood can be written as

$$p(\text{A wins}|R_A, R_B) = \frac{1}{1 + 10^{(R_B - R_A)/400}} \tag{12}$$

where R_B is player B's Elo rating at the time of the match, and R_A is player A's. The function transforming the difference $R_B - R_A$ to the win probability is a rescaled version of the inverse logit link used in equation (2). Both Elo and the model presented in this paper assume that match outcomes are the result of differences in player skills that vary dynamically over time. However, Elo only considers match win and loss, not point win and loss, and does not assume that players' skills follow a random walk over time. In addition, Elo ratings lack the ability to incorporate covariates such as the surface-specific skills and tournament intercepts.

3 Results

3.1 Performance as a function of period length and start date

Figure 1 summarises the evaluation results at the match level obtained with each combination of starting year and period length. Reducing the period length from 12 months to 2 months appears to improve log loss for all models, but dropping it from 2 months to a single month does not appear to yield further gains. Earlier starts appear to improve performance, with the 2011 curve consistently producing a lower log loss and higher accuracy than the later starts. The model starting in 2011 with 2 month periods has the lowest log loss at 0.592. This model also has the highest accuracy, at 68.8%. The model with a period length of 6 months starting in 2013 has the lowest accuracy (66.9%).

Figure 2 shows the same breakdown, but for the RMSE of the difference $\theta_i - \theta_j$ and the sum $\theta_i + \theta_j$. The trends are very similar, with earlier starts resulting in lower errors, and shorter periods improving performance until a period length of two months, after which the error appears to level off.

Overall, the model with the best figures on all metrics is the model starting in 2011 with 2 month periods.

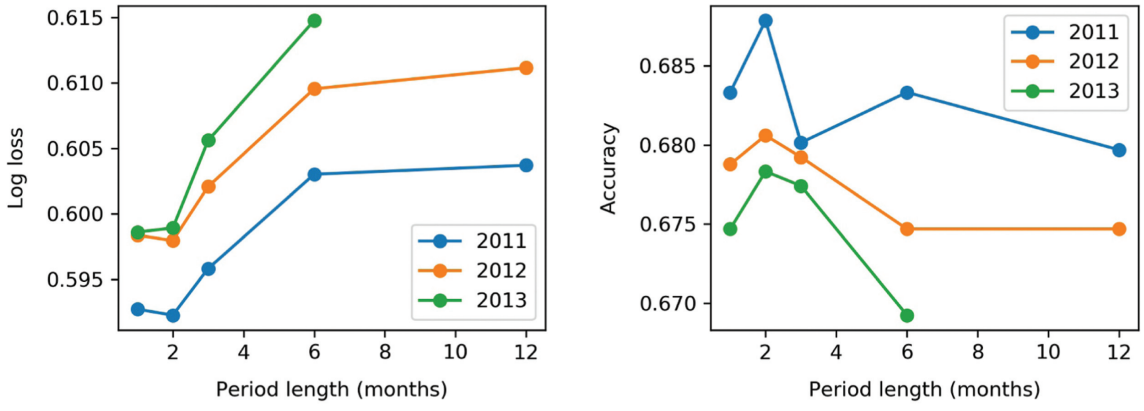

Figure 1: Match-level evaluation results. The left hand plot shows log loss for different period dates and start years, and the right hand plot breaks accuracy down in the same way.

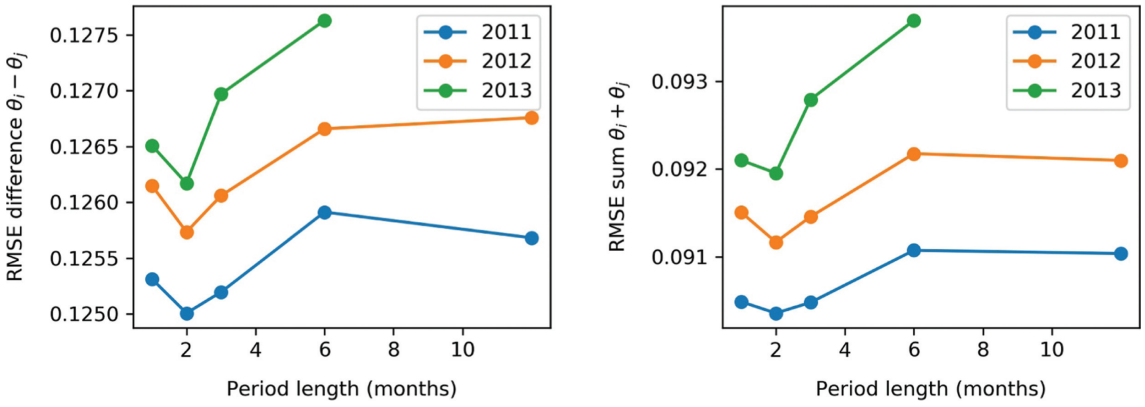

Figure 2: Point-level evaluation results. The left hand plot shows the RMSE for the difference, while the right hand plot shows the RMSE for the sum for different settings of the period length and starting year.

Table 2: Summaries of the 2014 out-of-sample metrics for the best proposed model (2011 start, 2 month periods).

	Accuracy	Log loss	RMSE $\theta_i + \theta_j$	RMSE $\theta_i - \theta_j$
Opponent-adjusted	66.3%	0.641	0.0993	0.130
538 Elo	69.5%	0.586	N/A	N/A
Best proposed model	68.8%	0.592	0.0904	0.125

3.2 Comparison against baseline models

Table 2 compares the metrics of the best model against the baseline models.[1] The model strongly outperforms the opponent-adjusted model on all metrics, with 2.4% higher accuracy, 0.049 lower log loss, and reduced RMSE on both sum and difference of serve-winning probabilities. Compared to Elo, the model's accuracy and log loss are slightly worse. Elo does not predict serve-winning probabilities, hence the missing values in the table for the point-level RMSE scores.

3.3 Posteriors for model coefficients

3.3.1 Group level variance and intercept posteriors

Table 3 summarises the posterior inferences obtained for the group level variance parameters in the best-performing model. With posterior medians of 0.154 and 0.124, respectively, the largest source of variation is the initial uncertainty of players' serve and return skills $\sigma_{\alpha 0}$ and $\sigma_{\beta 0}$. The variation in surface preferences σ_γ is about as large as the variation in tournament-specific intercepts σ_δ. Finally, period-to-period variation of serve skills σ_α is slightly larger than period-to-period variation in return skill σ_β.

Table 3: Posterior summaries of the group level variance parameters and overall intercept θ_0 for the model with fit starting in 2011 and 2 month periods.

	2.5%	25%	median	75%	97.5%
$\sigma_{\alpha 0}$	0.138	0.148	0.154	0.160	0.172
$\sigma_{\beta 0}$	0.109	0.119	0.124	0.130	0.140
σ_γ	0.059	0.064	0.067	0.070	0.076
σ_δ	0.057	0.064	0.069	0.073	0.082
σ_α	0.034	0.037	0.039	0.041	0.044
σ_β	0.023	0.025	0.027	0.028	0.031
θ_0	0.473	0.493	0.503	0.514	0.533

The posterior for the overall intercept θ_0 has a median value of 0.503, indicating that the estimated median win probability on serve for equally-matched players at an average tournament and surface is 62.3%.

3.3.2 Player serve and return posteriors

The model produces serve and return posteriors for each time period, but for the sake of brevity, we focus on the posteriors for one period: the two-month period starting in September 2014. Tables 4 and 5 show posterior summaries for the ten highest serve and return skills estimated for this period. The best servers in this period were estimated to be Ivo Karlovic, Milos Raonic, John Isner and Roger Federer; the best returners were Novak Djokovic, Andy Murray, Rafael Nadal and David Ferrer. Roger Federer and Novak Djokovic are the only players to be in both tables,

1 The accuracy and log loss reported in this paper for Elo is the same as that reported for the 2014 season in Kovalchik (2016). However, Kovalchik reports slightly better performance for the opponent-adjusted model (67% accuracy and 0.63 log loss). These differences are small however and are likely due to small differences in the datasets used.

Table 4: Posterior summaries of the serve-skills α for the servers with the highest median skill in the two-month period starting in September 2014.

	2.5%	25%	median	75%	97.5%
Ivo Karlovic	0.53	0.61	0.66	0.70	0.78
Milos Raonic	0.46	0.53	0.57	0.61	0.69
John Isner	0.41	0.50	0.54	0.58	0.66
Roger Federer	0.39	0.47	0.51	0.55	0.63
Jo-Wilfried Tsonga	0.30	0.39	0.43	0.48	0.57
Novak Djokovic	0.30	0.38	0.42	0.47	0.55
Sam Groth	0.24	0.34	0.39	0.45	0.56
Juan Martin Del Potro	0.18	0.32	0.39	0.46	0.59
Marin Cilic	0.22	0.30	0.34	0.39	0.46
Stan Wawrinka	0.21	0.29	0.34	0.38	0.47

Table 5: Posterior summaries of the return skills β for the highest-median returners during the two-month period starting in September 2014.

	2.5%	25%	median	75%	97.5%
Novak Djokovic	0.34	0.41	0.44	0.48	0.55
Andy Murray	0.29	0.36	0.39	0.43	0.50
Rafael Nadal	0.24	0.32	0.36	0.40	0.48
David Ferrer	0.25	0.32	0.36	0.39	0.46
Kei Nishikori	0.20	0.27	0.31	0.34	0.41
Roger Federer	0.19	0.26	0.29	0.33	0.39
Gael Monfils	0.17	0.25	0.29	0.33	0.41
David Nalbandian	0.07	0.20	0.27	0.34	0.48
Axel Michon	0.03	0.19	0.27	0.35	0.51
Gilles Simon	0.16	0.23	0.27	0.31	0.38

suggesting that they are very effective both when serving and returning.

Figure 3 compares the median serve and return skill estimates, again for the period starting in September 2014. Eight players are highlighted: four with the highest serve skills, and four with the highest returning skills. The very strongest servers – John Isner and Ivo Karlovic – have relatively low return skill estimates. Among the players with the highest return skills, there is some variety: Nadal, Djokovic and Federer have high skill estimates on both serve and return, while Andy Murray and David Ferrer have high median return skills but somewhat lower serve skill estimates.

3.4 Serve and return estimates over time

In the previous section, we focused on the model's estimates in a particular period. To illustrate the model's dynamic aspect, Figure 4 shows the evolution of serve and return skills over time for Roger Federer and Rafael Nadal. Federer generally had the higher serve skill estimate during this period, while Nadal had the higher return skill estimate.

On serve, Federer appeared to decline from February 2012 onwards, reaching a minimum in July 2013 before improving again in early 2014. Nadal shows an opposite trend, generally improving until he reached a peak in July 2013 and declined again. On return, the shifts are less dramatic, although Federer appears to show a slight upward trend.

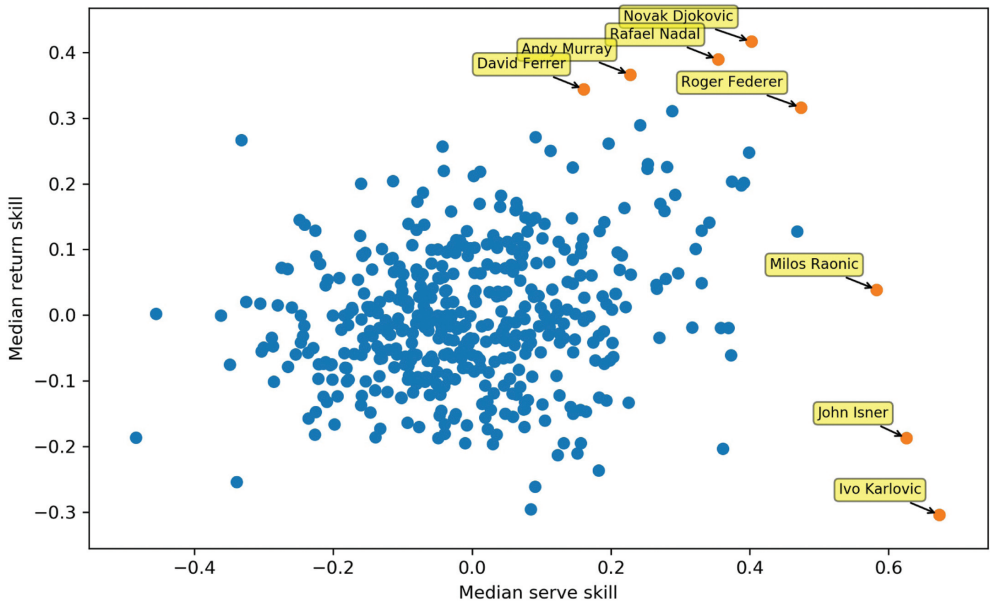

Figure 3: Median posterior serve and return skills plotted against each other. Players with the four highest median serve and return skills are labelled.

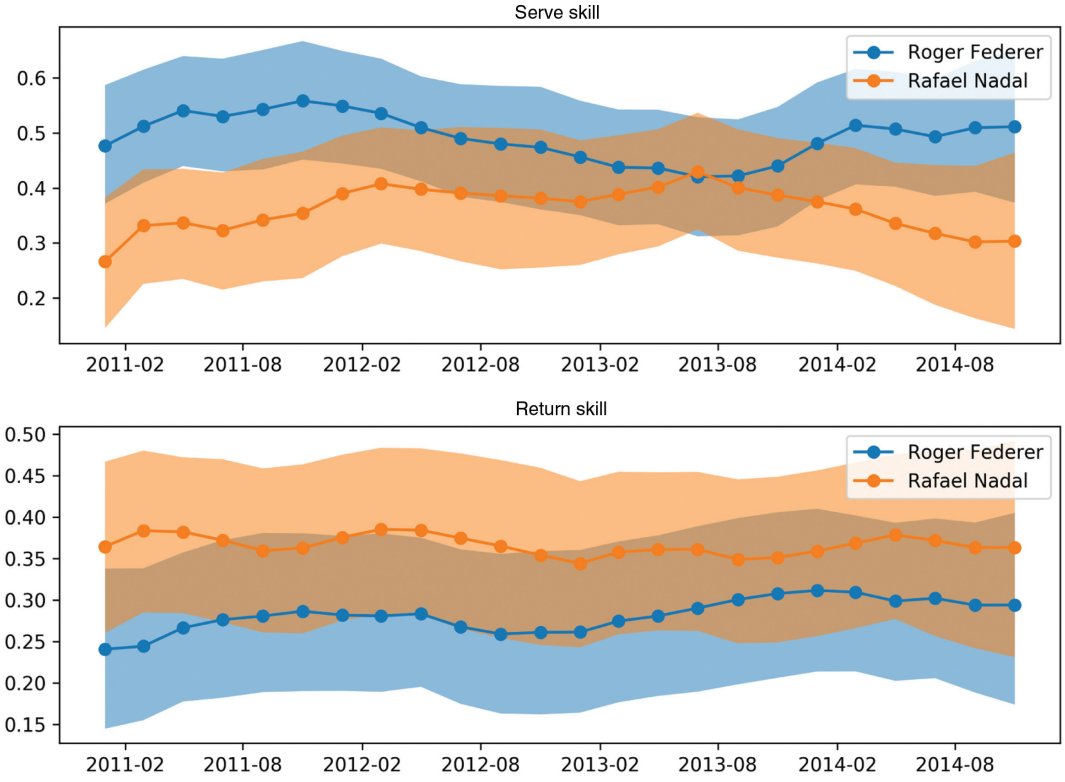

Figure 4: Serve and return skills evolving over time for Roger Federer and Rafael Nadal. The shaded regions denote 95% credible intervals.

The shaded 95% credible intervals are relatively constant throughout. One exception are the final periods for Rafael Nadal, during which the credible intervals widen. Nadal did not play in the period between Wimbledon (starting in the last week of June 2014) and Beijing (starting in the last week of September 2014), and the widening intervals indicate that his lack of play make his skill estimate more uncertain.

3.5 Surface skills

The estimated values for the surface preferences γ on different surfaces are shown in Table 6 for the top 3 highest median skills. On clay, Rafael Nadal has the highest preference, improving his serve and return skill by a median amount of 0.16. On grass courts, Lleyton Hewitt showed the largest boost, and James Blake has the largest median preference for hard courts.

Table 7 shows the lowest surface preferences for each surface. The player with the second-largest preference for clay, Pablo Andujar, has the lowest preference on grass; the player with the highest preference for grass, Lleyton Hewitt, has the third lowest preference for clay. This agrees with conventional wisdom in tennis that clay-court specialists tend to struggle on grass courts, and vice versa.

The range of skills on clay ranges from $+0.16$ (Rafael Nadal) to -0.11 (Benjamin Becker). On grass, the range

Table 6: Top 3 median highest surface skill estimates γ for each surface.

surface		2.5%	25%	median	75%	97.5%
clay	Rafael Nadal	0.08	0.13	0.16	0.18	0.24
	Pablo Andujar	0.06	0.11	0.14	0.17	0.23
	Federico Delbonis	0.04	0.10	0.13	0.16	0.22
grass	Lleyton Hewitt	0.03	0.09	0.12	0.14	0.20
	Brian Baker	0.00	0.07	0.11	0.15	0.22
	Andy Murray	0.03	0.08	0.11	0.14	0.19
hard	James Blake	0.01	0.07	0.10	0.13	0.19
	Novak Djokovic	0.02	0.07	0.10	0.12	0.18
	Somdev Devvarman	0.00	0.06	0.09	0.12	0.18

Table 7: Top 3 lowest surface skill estimates γ for each surface.

surface		2.5%	25%	median	75%	97.5%
clay	Benjamin Becker	−0.21	−0.15	−0.11	−0.08	−0.01
	Lukas Lacko	−0.20	−0.13	−0.10	−0.06	−0.00
	Lleyton Hewitt	−0.19	−0.13	−0.10	−0.07	−0.00
grass	Pablo Andujar	−0.28	−0.21	−0.18	−0.15	−0.08
	Daniel Gimeno−Traver	−0.22	−0.15	−0.11	−0.08	−0.01
	Carlos Berlocq	−0.18	−0.12	−0.09	−0.05	0.01
hard	Jan Hajek	−0.21	−0.14	−0.11	−0.07	−0.01
	Jabor Mohammed Ali Mutawa	−0.23	−0.14	−0.10	−0.05	0.03
	Filippo Volandri	−0.18	−0.12	−0.09	−0.05	0.01

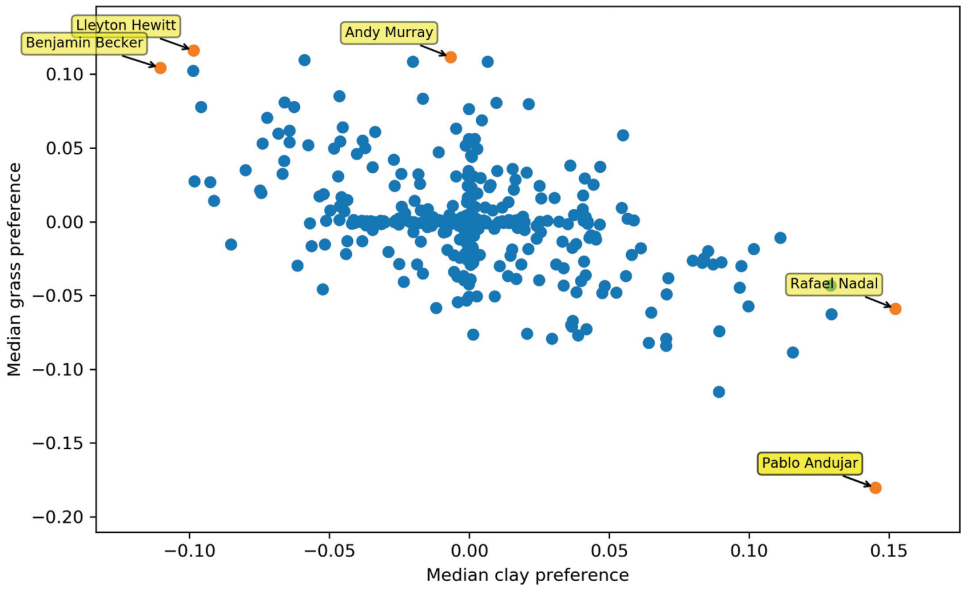

Figure 5: Median clay court preferences plotted against median grass court preferences for the period of September 2014.

is between +0.12 (Lleyton Hewitt) and −0.18 (Pablo Andujar). Preference estimates are less diverse on hard courts, ranging from +0.10 (James Blake) to −0.11 (Jan Hajek).

Figure 5 shows a scatter plot of median clay court preference against median grass court preference. The plot appears to show a negative correlation: players with strong skills on clay tend to have low skills on grass, and vice versa. The cross-shaped accumulation of points near zero shows the effect of the independent hierarchical prior, which independently shrinks players' skills on grass and clay towards zero.

3.6 Tournament intercepts

Figure 6 shows the posteriors obtained for the intercepts of tournaments played in 2014. Generally speaking, and as expected from the summaries in Table 1, clay court tournaments appear to be associated with lower expected probabilities of winning a point on serve, while the opposite is true for grass court tournaments. Hard court tournaments span a wider range, with some intercepts close to those of clay court tournaments (such as Beijing, Acapulco and Indian Wells) and others closer to those of grass court tournaments (such as Marseille and Montpellier). In the case of Acapulco, it should be noted that the tournament was played on clay from 2011 to 2013.

The largest intercepts are observed for the grass court tournaments of Halle and Wimbledon, with median values of +0.15 and +0.14, respectively. The smallest are the clay court tournaments in Monte Carlo and Umag, with median estimates of −0.14 and −0.13.

3.7 Match prediction example: 2014 French Open Final

To illustrate the model's use for prediction, we show how the model predicted the 2014 French Open final, in which

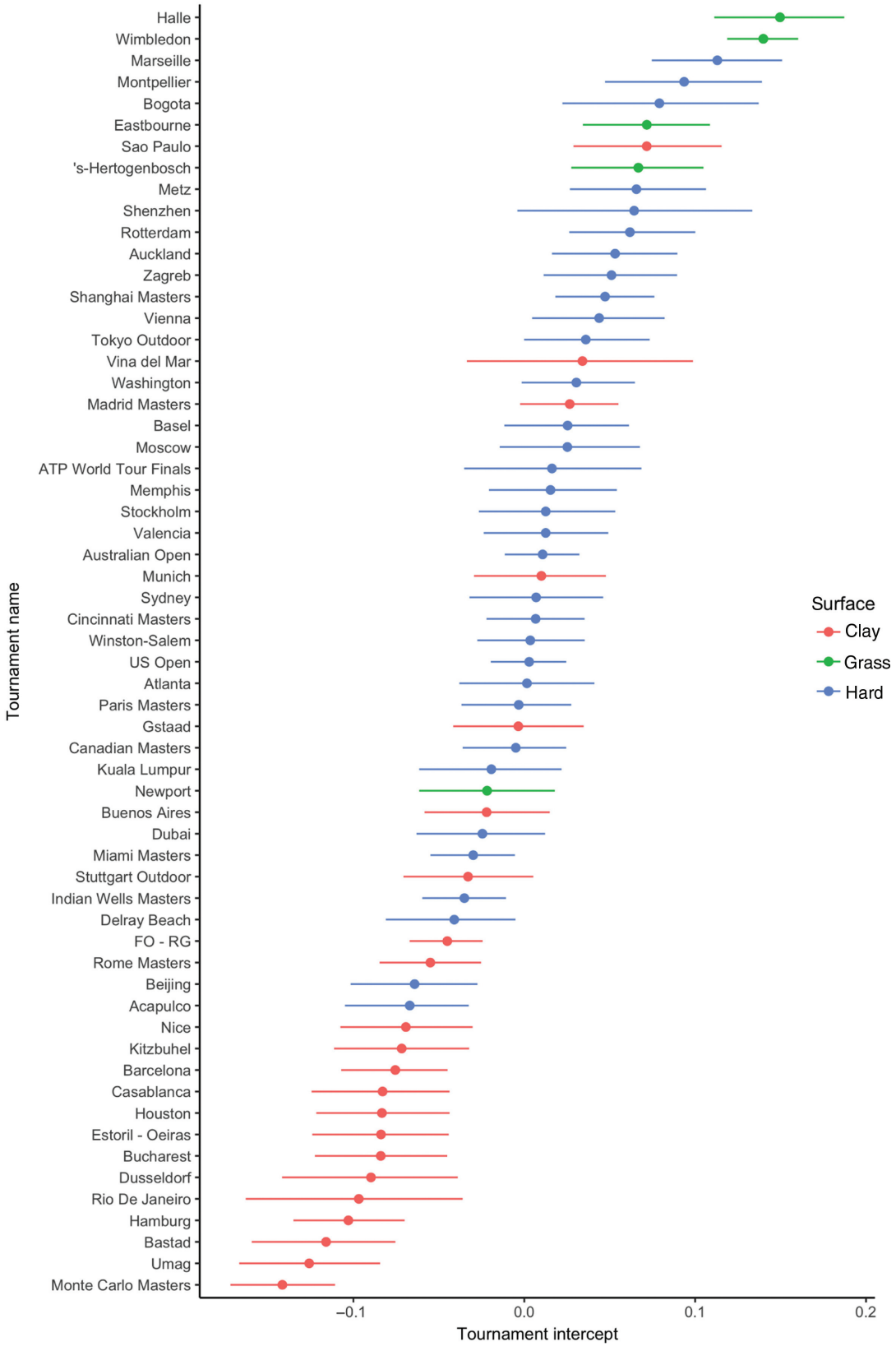

Figure 6: Posterior estimates of tournament-specific intercepts for tournaments played in 2014, coloured by surface, together with their 95% credible intervals.

Rafael Nadal beat Novak Djokovic in four sets, 3-6 7-5 6-2 6-4.

The match took place on the 8th June 2014, so to predict this match using the model with two-month periods, the model is fit up to the start of May 2014. The serve and return skill posteriors for Nadal and Djokovic are then obtained by adding the period-to-period variance to their May 2014 estimates.

Figure 7 shows the posterior estimates for the two players. Novak Djokovic's serve skill was likely to be higher than Rafael Nadal's, and so was his return skill. However, Rafael Nadal's preference for clay was estimated to be much greater than Novak Djokovic's, and so the posterior mode of his predicted serve-winning probability is slightly greater than Novak Djokovic's.

Figure 8 shows the results of using the posterior samples together with the iid model of a tennis match. The iid model suggests that the outcome – Nadal winning in 4 sets – was most likely, and that Nadal's mean win probability was 58%. In the match, Nadal won 66% of points on serve and Djokovic won 60%, both of which fall within the 95% credible intervals shown.

Had Nadal and Djokovic played on grass at Wimbledon instead, the win probabilities would have looked very different. Figure 9 shows the serve probability estimates and the predictions made by the iid model for this hypothetical scenario. On grass, the most likely outcome would be a win by Novak Djokovic in three sets.

4 Discussion

4.1 Strengths and weaknesses of the proposed approach

As the evaluation results in Table 2 show, the proposed model outperforms the previous best point-based model by a large margin. This suggests that, if the goal is to predict serve-winning probabilities, the proposed approach is a better choice. As illustrated in the previous section, the advantage of predicting serve-winning probabilities is that the iid model of a tennis match can be used to predict many aspects of a match, not just win or loss. In addition to the probability of set scores shown, the iid model can also be used to derive win probabilities conditional on the score in a match, allowing it to provide in-match win probabilities (Klaassen and Magnus 2003).

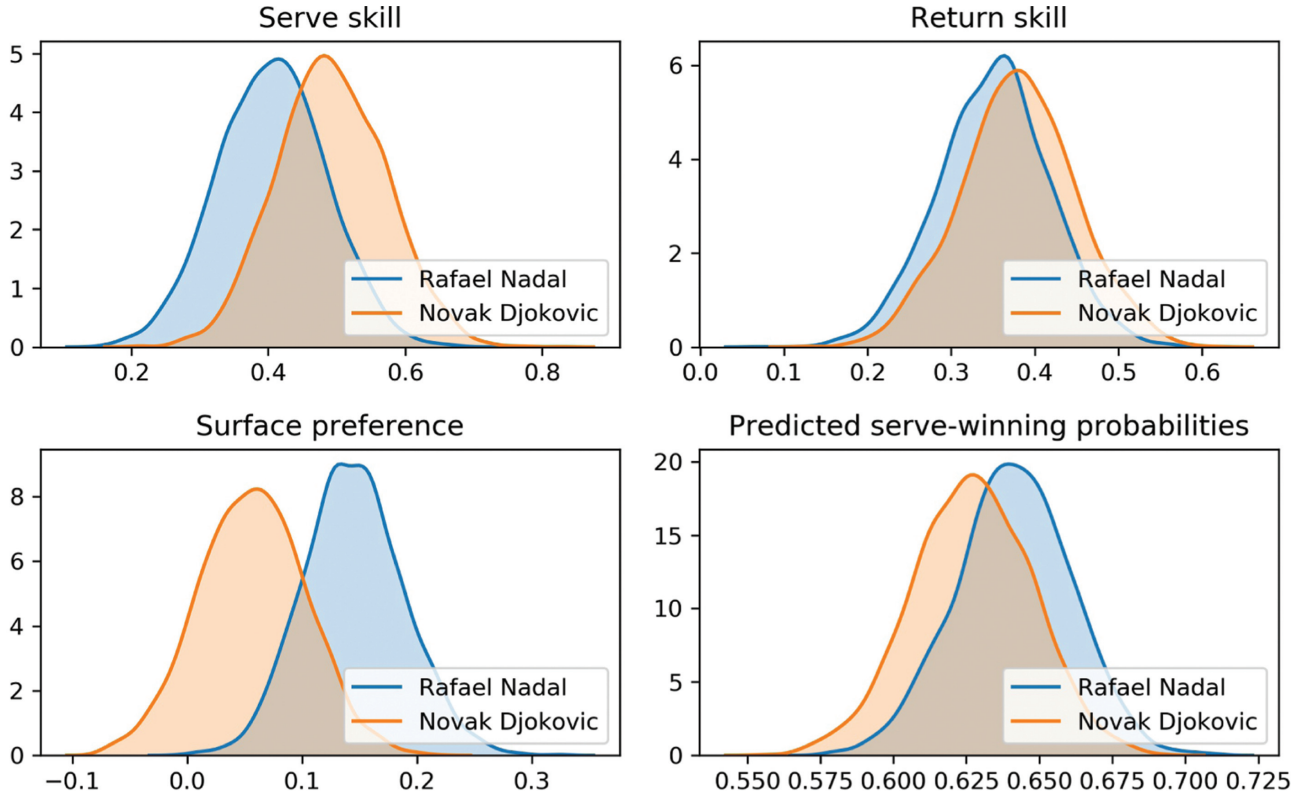

Figure 7: Serve, return and surface skill estimates for Novak Djokovic and Rafael Nadal for the 2014 French Open final, as well as the predictions of their serve-winning probabilities.

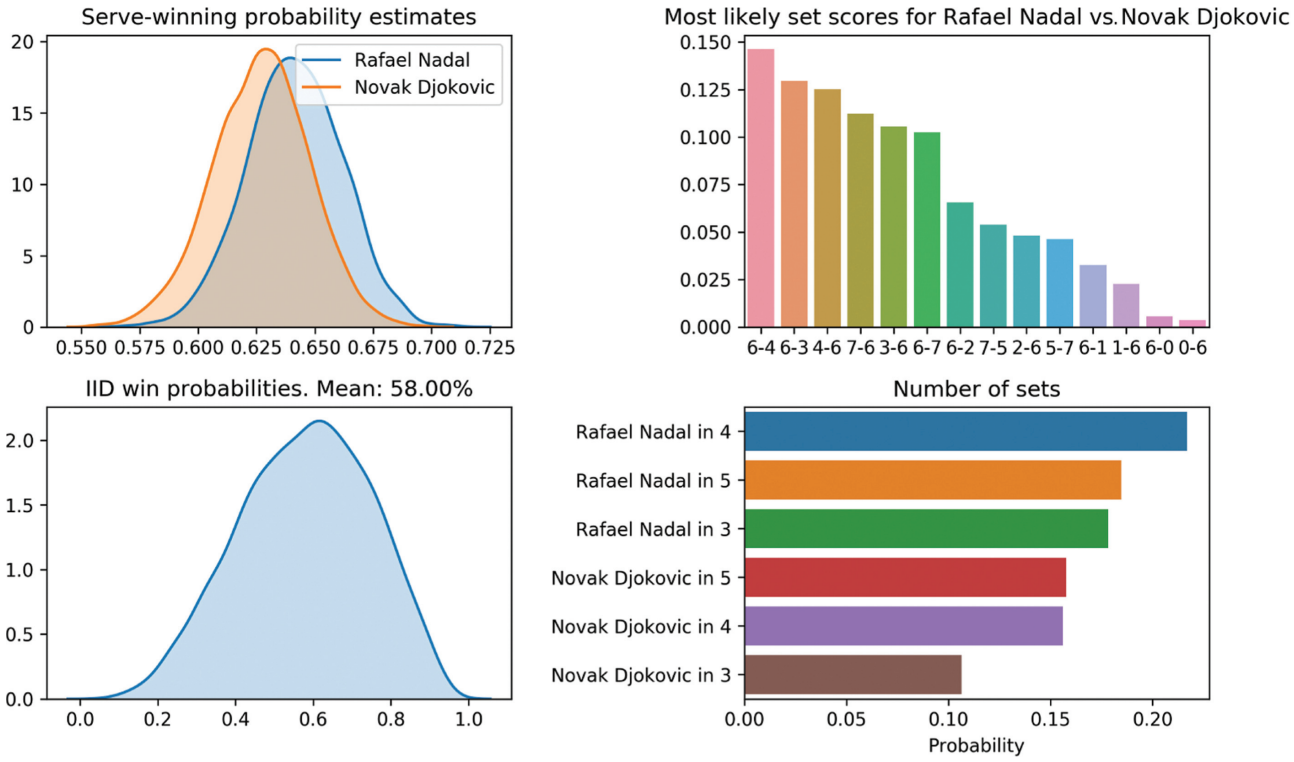

Figure 8: Predictions made for the 2014 French Open final using the iid model.

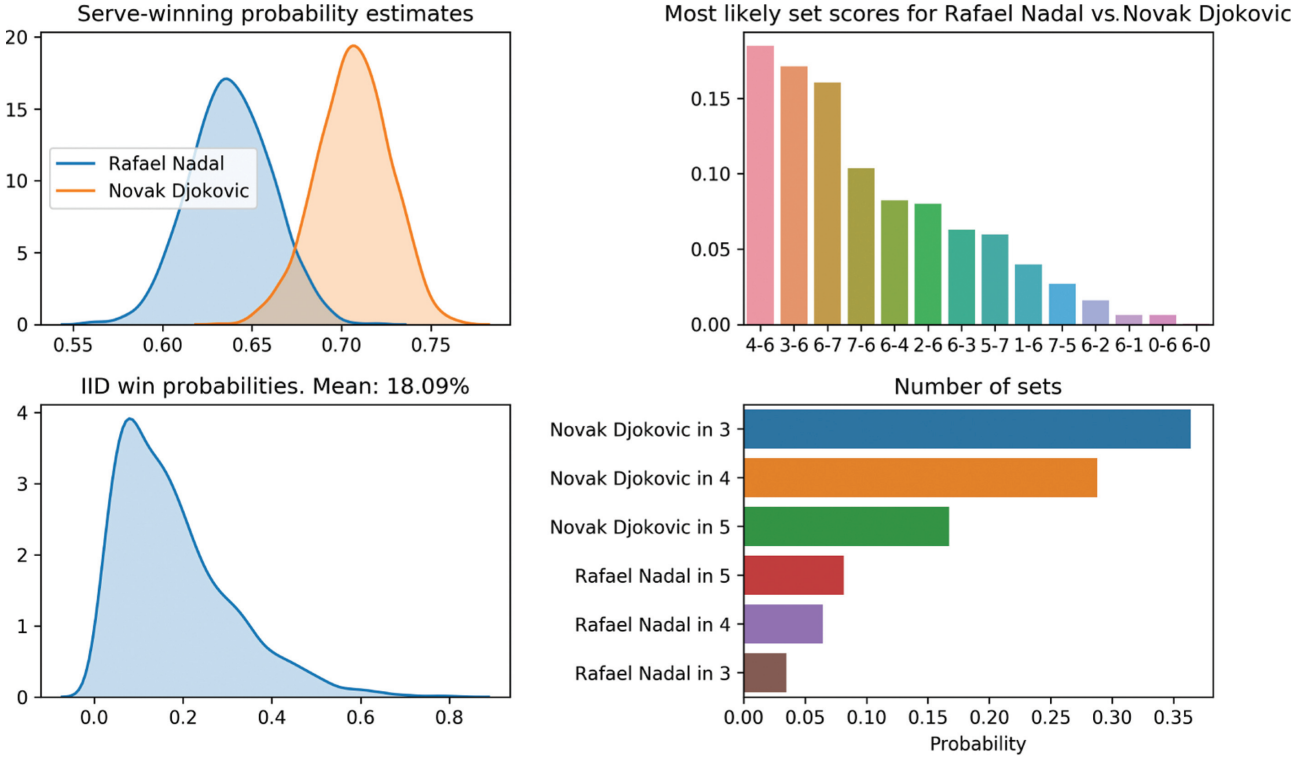

Figure 9: Predictions made for a hypothetical Wimbledon match between Nadal and Djokovic in the period starting March 2014.

Another strength of the proposed model is the interpretability of its results. As shown in the results section, the model estimates player-specific surface effects, allowing players to be characterised by their surface skills. The match prediction examples of Nadal – Djokovic on clay and grass illustrate that the differences between surface skills can be considerable. The tournament-specific intercepts reveal an interesting range of coefficients for tournaments, with some surprising results, such as the clay court tournament in Sao Paulo, which has a higher intercept than some grass court tournaments.

For match outcome prediction, the model performs slightly worse than Elo, despite its additional complexity. A number of reasons may explain why Elo performs so well. Firstly, Elo does not make the iid assumption. It may be that targeting match outcomes directly, rather than serve-winning probabilities, helps it perform better. Secondly, Elo is able to update more quickly than the model proposed. While the shortest period considered for the model in this paper is one month, Elo ratings are updated after each match. Finally, Elo is much faster to fit, allowing it to use much more data: while an efficient implementation of Elo can be calculated with a 1968 start in under a minute, the proposed model (2 months with a start in 2011) takes about 80 minutes to fit, which limits the amount of data that can be used. Figure 1 suggests that earlier starts may lead to reduced error, indicating that more data may also help this model perform better. An approximate solution of the model, using, for example, expectation propagation (Minka 2001) or similar techniques to those applied in the derivation of Glicko (Glickman 1999) could be of interest.

However, it should be noted that Elo only predicts the match outcome and thus cannot make the wealth of predictions concerning the match that point-based models can. A recent paper by Kovalchik and Reid (2018) illustrates how Elo forecasts can be "calibrated" to estimate the serve-winning probabilities, but this can only be done by making additional assumptions about players' serving skills.

4.2 Possible additions to the model

Figure 5 suggests that players who excel on clay courts may be likely to be less proficient on grass courts. A multivariate hierarchical prior allowing for correlation among the surface skills could provide a better fit to the data.

As mentioned, the skill evolution assumes that players' skills change smoothly from one period to the next with equal variance. In some situations, this assumption may not hold; for example, players could sustain injuries, leading to a more rapid decrease in skill than this random walk can accommodate. Allowing the random walk variances to change at each period, as proposed for example in Glicko 2 (Glickman 2001), could handle this situation better.

To improve predictive accuracy, it may also be necessary to move beyond the iid assumption. Using point-level data and indicators such as "set down" and "break point" derived from this data as proxies for pressure, Kovalchik and Ingram (2016) assessed players' deviations from iid behaviour and developed a Monte Carlo model of tennis matches taking these deviations into account. Another effect that could be interesting to explore is whether players tend to tire over the course of the match, and whether this differs by player. For instance, players relying on a strong serve may tire more quickly than those who are particularly skilled at returning, or vice versa. Fitting the proposed model at the point level and including point-level predictors would likely improve predictive accuracy, although it would increase the data set size substantially, since each match could no longer be summarised by the sufficient statistics "points played on serve" and "points won on serve".

4.3 Conclusions

In summary, the proposed model improves on the state of the art in point-based prediction models for tennis and comes close to matching the strong baseline of Elo for match outcome prediction. Future work could include modelling non-iid effects, as well as working on speeding up model fit.

References

Barnett, T. J. 2006. *Mathematical Modelling in Hierarchical Games with Specific Reference to Tennis*. Ph.D. thesis.

Barnett, T. and S. R. Clarke 2005. "Combining Player Statistics to Predict Outcomes of Tennis Matches." *IMA Journal of Management Mathematics* 16:113–120.

Carpenter, B., A. Gelman, M. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. A. Brubaker, J. Guo, P. Li, and A. Riddell 2016. "Stan: A Probabilistic Programming Language." *Journal of Statistical Software* 20:1–37.

Elo, A. E. 1978. *The Rating of Chessplayers, Past and Present*. Arco Pub, p.34.

Gelman, A. and D. B. Rubin. 1992. "Inference from Iterative Simulation Using Multiple Sequences." *Statistical Science* 7:457–472.

Gelman, A., H. S. Stern, J. B. Carlin, D. B. Dunson, A. Vehtari, and D. B. Rubin 2013. *Bayesian Data Analysis* (3rd edition). Chapman and Hall/CRC, pp. 42–43.

Glickman, M. E. 1999. "Parameter Estimation in Large Dynamic Paired Comparison Experiments." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 48:377–394.

Glickman, M. E. 2001. "Dynamic Paired Comparison Models with Stochastic Variances." *Journal of Applied Statistics* 28:673–689.

Klaassen, F. J. and J. 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* 96:500–509.

Klaassen, F. J. and J. R. Magnus 2003. "Forecasting the Winner of a Tennis Match." *European Journal of Operational Research* 148:257–267.

Kovalchik, S. A. 2016. "Searching for the Goat of Tennis Win Prediction." *Journal of Quantitative Analysis in Sports* 12:127–138.

Kovalchik, S. and M. Ingram 2016. "Hot Heads, Cool Heads, and Tacticians: Measuring the Mental Game in Tennis (id: 1464)." MIT Sloan Sports Analytics Conference, March 11-12, Boston, USA, http://www.sloansportsconference.com/wp-content/uploads/2016/02/1464-Hot-heads-cool-heads-and-tacticians.pdf.

Kovalchik, S. and M. Reid 2018. "A Calibration Method with Dynamic Updates for Within-Match Forecasting of Wins in Tennis." *International Journal of Forecasting* 35: 756–766.

Minka, T. P. 2001. "Expectation Propagation for Approximate Bayesian Inference." in *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, Morgan Kaufmann Publishers Inc., 362–369.

Morris, B. and C. Bialik 2015. "Serena Williams and the Difference between All-Time Great and Greatest of All Time." http://fivethirtyeight.com/features/serena-williams-and-the-difference-between-all-time-great-and-greatest-of-all-time/.

Newton, P. K. and J. B. Keller 2005. "Probability of Winning at Tennis i. Theory and Data." *Studies in applied Mathematics* 114:241–269.

Newton, P. K. and K. Aslam 2006. "Monte Carlo Tennis." *SIAM Review* 48:722–742.

O'Malley, A. J. 2008. "Probability Formulas and Statistical Analysis in Tennis." *Journal of Quantitative Analysis in Sports* 4:15.

Pollard, G., R. Cross, and D. Meyer 2006. "An Analysis of Ten Years of the Four Grand Slam Men's Singles Data for Lack of Independence of Set Outcomes." *Journal of Sports Science & Medicine* 5:561.

Riddle, L. H. 1988. "Probability Models for Tennis Scoring Systems." *Applied Statistics* 37: 63–75.