Journal of Sports Analytics 7 (2021) 223–233
DOI 10.3233/JSA-200345
IOS Press

Forecasting serve performance in professional tennis matches

Jacob Gollub*
Department of Statistics, Harvard University, Cambridge, MA, USA

Abstract. Many research papers on tennis match prediction use a hierarchical Markov Model. To predict match outcomes, this model requires input parameters for each player's serving ability. While these parameters are often computed directly from each player's historical percentages of points won on serve and return, doing so fails to address bias due to limited sample size and differences in strength of schedule. In this paper, we explore a handful of novel approaches to forecasting serve performance that specifically address these limitations. By applying an Efron-Morris estimator, we provide a means to robustly forecast outcomes when players have limited match data over the past year. Next, through tracking expected serve and return performance in past matches, we account for strength of schedule across all points in a player's match history. Finally, we demonstrate a new way to synthesize historical serve data with the predictive power of Elo ratings. When forecasting serve performance across 7,622 ATP tour-level matches from 2014-2016, all three of these proposed methods outperformed Barnett and Clarke's standard approach.

Keywords: Match prediction, point-based models, Efron-Morris estimator, Elo ratings

1. Introduction

Statistical prediction models have been applied to tennis for decades, often to predict the winner of a match. Laying the groundwork for research on point-based models, Klaassen and Magnus (2001) first tested the assumption that points in a tennis match are independent and identically distributed, or i.i.d. While this was proven false, they concluded deviations are small enough that the i.i.d. assumption provides a reasonable approximation. By estimating each player's probability of winning a point on serve and computing win probability from these parameters, they demonstrated a new way to forecast matches under this assumption (Klaassen and Magnus, 2003).

In the time since, we have seen a variety of ways to forecast serve performance. Barnett and Clarke (2005) estimated these parameters by adjusting historical tournament statistics by the performance of the server and returner, relative to the average player. Assuming that serve and return performances follow a normal distribution on a match-by-match basis, Newton and Aslam (2009) ran Monte Carlo simulations to predict the winner of a match. Adapting Barnett and Clarke's approach, Spanias and Knottenbelt (2012) predicted serve performance by modeling the probabilities of specific point outcomes (e.g. ace, rally win on first serve, etc.) within a service game. In order to reduce bias from variations in strength of schedule, Knottenbelt et al. (2012) then proposed the Common-Opponent Model, which measures performance relative to common adversaries in their respective match histories. More recently, Kovalchik and Reid (2018) demonstrated a way to calibrate serve parameters to the win probability implied by each player's Elo rating.

Building upon previous work in serve performance prediction, this exploration serves two purposes. Following a previous assessment of 11 different pre-match prediction models on 2014 ATP tour-level matches (Kovalchik, 2016), we evaluate serve performance prediction methods across the same dataset as a benchmark for future work. We also introduce sev-

*Corresponding author: Jacob Gollub, Department of Statistics, Harvard University, Cambridge, MA, USA. E-mail: jacobgollub@gmail.com.

ISSN 2215-020X

eral new methods to more robustly compute expected serve performance. When forecasting thousands of matches, the previously stated methods tend to run into problems with limited match data and differences in strength of schedule.[1] To avoid these pitfalls, we explore variations of Barnett and Clarke's approach that specifically address these issues.

In section 2, we review the hierarchical Markov Model and Elo ratings, two of the most often-cited methods for predicting tennis match outcomes. In section 3, we review Barnett and Clarke's formula and explore further approaches to serve performance prediction. Section 3.1.1 introduces a Bayesian estimator to more robustly predict serve performance from limited match data. Section 3.1.2 demonstrates an approach that tracks expected serve and return performances in every match, allowing us to consider strength of schedule over the course of a player's entire match history. Section 3.2 harnesses Elo ratings' predictive power to more accurately predict serve performance while maintaining the overall serve ability between two players. In section 4, we present several case studies to examine how each approach informs predictions for individual matches. In section 5, we evaluate all proposed methods alongside established prediction models. Section 6 summarizes findings and suggests future steps for research with point-based models.

Effective serve performance prediction holds strong implications for both player analytics and in-match forecasting. With more effective forecasts, coaches can better understand how their players match up with opponents on serve and return, and adjust their strategies accordingly. By using the hierarchical Markov Model, we can also compute win probability as a function of each player's expected serve performance from any score. Therefore, improving existing methods will provide means to more confidently predict match outcomes while play is in progress, a problem with significant application to betting markets and real-time sports analytics.

2. Models for match win prediction

Newly proposed methods will require the context of two popular match prediction models. The hierarchical Markov Model computes win probability directly from point-level probabilities, while Elo ratings infer player ability solely from match outcomes.

2.1. Hierarchical markov model

Consider a tennis match between player i and player j, where f_{ij} estimates the probability that player i wins a point on serve against player j in a given match. Given serve parameters f_{ij}, f_{ji} we can calculate the probability that player i wins the match, π_{ij}, from any score. To do so, consider how scores are composed of points, games, and sets. With player i serving to player j, let a_i, a_j represent each player's within-game score. Then we compute the probability of player i winning the current game from score $a_i - a_j$ as follows:

$$P_g(a_i, a_j) = \begin{cases} 1, & \text{if } a_i = 4, a_j \leq 2 \\ 0, & \text{if } a_j = 4, a_i \leq 2 \\ \dfrac{(f_{ij})^2}{(f_{ij})^2 + (1 - f_{ij})^2}, & \text{if } a_i = a_j = 3 \\ f_{ij} * P_g(a_i + 1, a_j) + & \\ (1 - f_{ij}) P_g(a_i, a_j + 1), & \text{otherwise.} \end{cases}$$

Following this approach, one may calculate player i's probability of winning the current set, $P_s(g_i, g_j)$, and then the match, $P_m(s_i, s_j)$, through similar recursive relationships. Barnett et al. (2002) describe the recursion in full detail.

2.2. Elo ratings

Elo was originally developed as a head-to-head rating system for chess players (Elo, 1978). Recently, FiveThirtyEight's Elo variant has gained prominence in the media (Bialik et al., 2016). For a match at time t between player i and player j with Elo ratings $E_i(t)$ and $E_j(t)$, player i is forecasted to win with probability:

$$\hat{\pi}_{ij}(t) = \left(1 + 10^{\frac{E_j(t) - E_i(t)}{400}}\right)^{-1}.$$

For the following match, player i's rating is then updated accordingly:

$$E_i(t+1) = E_i(t) + K_{it} * (W_i(t) - \hat{\pi}_{ij}(t)).$$

[1] Knottenbelt's Common-Opponent model addresses strength of schedule at the cost of reducing the amount of available match data. Kovalchik and Reid's approach indirectly addresses both of these issues through the use of Elo ratings.

$W_i(t)$ is an indicator for whether player i won the given match at time t, while K_{it} is the learning rate for player i at time t.

According to FiveThirtyEight's analysts, Elo ratings perform optimally when allowing K_{it} to decay slowly over time (Bialik et al., 2016). With $m_i(t)$ representing the number of player i's career matches at time t, we update the learning rate as follows[2]:

$$K_{it} = \frac{250}{(5 + m_i(t))^{.4}}.$$

This variant updates a player's Elo rating most quickly when we have no information about them and makes smaller changes as $m_i(t)$ accumulates. To apply this rating system to all ATP tour-level matches, we initalize each player's Elo rating at $E_i(0) = 1500$ and match history $m_i(0) = 0$. Then, we iterate through all tour-level matches from 1968-present in chronological order, storing $E_i(t)$, $E_j(t)$ for each match and updating each player's Elo rating accordingly.[3]

3. Predicting serve performance

A significant portion of research in tennis match prediction concerns estimating each player's probability of winning a point on serve. We present several variations to Barnett and Clarke's approach in Section 3.1 and a way to reconcile Elo ratings with these point-based methods in Section 3.2.

3.1. Barnett-Clarke formula

Given players' historical serve/return performance, Barnett and Clarke (2005) demonstrated a method to calculate f_{ij}, f_{ji}.

$$f_{ij} = f_t + (f_i - f_{av}) - (g_j - g_{av})$$
$$f_{ji} = f_t + (f_j - f_{av}) - (g_i - g_{av})$$

In a match between player i and player j, each parameter estimates one player's probability of winning a point on serve against the other. f_i represents player i's historical percentage of points won on serve, while g_i corresponds to their percentage of points won on return. f_t denotes the percentage of points won on serve at the match's given tournament and f_{av}, g_{av} represent tour-level averages for the percentages of points won on serve and return, respectively.

While Barnett and Clarke's dataset was limited to year-to-date statistics, we may calculate f_i, g_i with the past twelve months of match data for any given match. Where $\mathcal{M}^i_{(y,m)}$ represents the set of player i's matches in year y, month m, we obtain the following statistics[4]:

$$f_i(y, m) = \frac{\sum_{t=1}^{12} \sum_{k \in \mathcal{M}^i_{(y-1,m+t)}} w_{ik}}{\sum_{t=1}^{12} \sum_{k \in \mathcal{M}^i_{(y-1,m+t)}} n_{ik}}$$

$$g_i(y, m) = \frac{\sum_{t=1}^{12} \sum_{k \in \mathcal{M}^i_{(y-1,m+t)}} n_{jk} - w_{jk}}{\sum_{t=1}^{12} \sum_{k \in \mathcal{M}^i_{(y-1,m+t)}} n_{jk}}.$$

w_{ik} denotes the number of service points won by player i in match k and n_{ik} the total number of service points played by player i in match k.

Next, we calculate f_t for a given tournament and year, where $\mathcal{M}_{(v,y)}$ represents the set of all matches played at tournament v in year y:

$$f_t(v, y) = \frac{\sum_{k \in \mathcal{M}_{(v,y-1)}} w_k}{\sum_{k \in \mathcal{M}_{(v,y-1)}} n_k}.$$

w_k and n_k represent the number of service points won and played in match k, respectively.

Finally, we calculate f_{av}, g_{av} where $\mathcal{M}_{(y,m)}$ represents the set of tour-level matches played in year y, month m:

$$f_{av}(y, m) = \frac{\sum_{t=1}^{12} \sum_{k \in \mathcal{M}_{(y-1,m+t)}} w_k}{\sum_{t=1}^{12} \sum_{k \in \mathcal{M}_{(y-1,m+t)}} n_k}$$

$$g_{av}(y, m) = 1 - f_{av}(y, m).$$

Overall, Barnett and Clarke's formula assumes that differences between player serve and return ability are additive. Next, we explore variations to this approach.

3.1.1. Efron-Morris estimator

In the case of players who do not regularly compete in tour-level events, f_i, g_i must be calculated from limited sample sizes. Consequently, match probabilities based on these estimates can be skewed by noise.

[2]The constants in this equation are parameter values that FiveThirtyEight's team chose after fitting this model on decades of tour-level match data.

[3]Tennis' Open Era began in 1968, when professionals were allowed to enter grand slam tournaments.

[4]For current month m, we only collect month-to-date matches.

To address this, we turn to the Efron-Morris estimator to provide alternative parameters of the form:

$$f'_{ij} = f_t + (f'_i - f_{av}) - (g'_j - g_{av})$$
$$f'_{ji} = f_t + (f'_j - f_{av}) - (g'_i - g_{av}).$$

Rather than directly apply the estimator to f_{ij}, f_{ji}, we normalize the serve/return parameters f_i, g_i which constitute Barnett and Clarke's equations.

Decades ago, Efron and Morris (1977) described a method to estimate a group of sample means with unequal variances. The Efron-Morris estimator shrinks sample means toward the overall mean by a magnitude proportional to each sample mean's uncertainty, producing a mean-squared error favorable in expectation to that of Maximum-Likelihood Estimation. While Barnett and Clarke use raw historical averages of serve and return points won, we can instead use this estimator to feed more reliable parameters into their equations. Just as Efron and Morris estimated toxoplasmosis rates across hospitals with uneven populations, we will apply this method to serve performance prediction.

Consider our match dataset $\mathcal{M}$, consisting of all tour-level matches from 1968-present. For each match k between players i and j, we calculate each player's historical percentage of points won f_{ik}, f_{jk} as outlined in section 3.1. Then, $\mathcal{F}$ contains each player's historical serve performance before every match:

$$\mathcal{F} = \textstyle\bigcup_{k \in \mathcal{M}} \mathcal{F}_k$$
$$\mathcal{F}_k = \{f_{ik}, f_{jk}\}.$$

To model serve performance according to a Bayesian distribution, we consider each $f_i \in \mathcal{F}$ to be a random variable that approximates a player's true service ability θ_i as follows:

$$f_i | \theta_i \sim N(\theta_i, \sigma_i^2)$$
$$\theta_i \sim N(\mu, \tau^2).$$

Put together, the above statements imply the following (Efron and Morris, 1975):

$$\theta_i | f_i \sim N(f_i + B_i(\mu - f_i), \sigma_i^2(1 - B_i))$$
$$B_i = \frac{{\sigma_i}^2}{\tau^2 + {\sigma_i}^2}.$$

Normalization coefficient B_i depends on both τ^2, the variance of true service ability, and σ_i^2, the variance of f_i given true service ability θ_i. With $\hat{f}_i$ denoting the observed value of f_i across n_i points, we first estimate mean and variance of true service ability from all matches in our dataset:

$$f_{av} = \frac{\sum_{f_i \in \mathcal{F}} \hat{f}_i}{|\mathcal{F}|}$$

$$\hat{\tau}^2 = \frac{\sum_{f_i \in \mathcal{F}} (\hat{f}_i - f_{av})^2}{|\mathcal{F}| - 1}.$$

Then, using maximum likelihood estimation, we estimate B_i and σ_i^2 in order to produce the Efron-Morris estimator:

$$\hat{B}_i = \frac{\hat{\sigma_i}^2}{\hat{\tau}^2 + \hat{\sigma_i}^2}$$
$$\hat{\sigma_i}^2 = \frac{\hat{f}_i(1 - \hat{f}_i)}{n_i}$$
$$f'_i = \hat{f}_i + \hat{B}_i(f_{av} - \hat{f}_i).$$

When n_i is large, our uncertainty in f_i decreases and shrinkage coefficient $\hat{B}_i$ approaches zero. As n_i gets smaller, $\hat{B}_i$ approaches one and f'_i more closely resembles the average serving ability. By repeating the same calculations with g'_i in place of f'_i, we can obtain Efron-Morris estimators for return ability and then compute f'_{ij}, f'_{ji} from these serve and return estimators.

While Barnett and Clarke's original paper computed f_i, g_i with sample means, using an Efron-Morris estimator will produce more robust forecasts across large datasets, where the amount of available data for a given match varies significantly. In addition, this estimator can be applied to other variations of Barnett and Clarke's approach, as we will observe when evaluating methods in Section 5.

3.1.2. *Opponent-adjusted ratings*

While Barnett and Clarke's equation considers the opponent's serve and return ability, it does not track strength of schedule throughout each player's match history. This is important, as a player's win percentages on serve/return may become inflated from playing weaker opponents or vice versa. In this section, we propose a variation to Barnett and Clarke's equation which replaces f_{av}, g_{av} with opponent-adjusted averages $1 - g'_i, 1 - f'_i$. The equations then become:

$$f'_{ij} = f_t + (f_i - (1 - g'_i)) - (g_j - (1 - f'_j))$$
$$f'_{ji} = f_t + (f_j - (1 - g'_j)) - (g_i - (1 - f'_i)).$$

f'_i, g'_i represent the average serve and return abilities of player i's opponents in the last twelve months. To calculate this, we weight each opponent's serve/return ability by the number of points played throughout player i's match history:

$$f'_i(y, m) = \frac{\sum_{t=1}^{12} \sum_{k \in \mathcal{M}^i_{(y-1,m+t)}} n_{jk} * (1 - g'_{ijk})}{\sum_{t=1}^{12} \sum_{k \in \mathcal{M}^i_{(y-1,m+t)}} n_{jk}}$$
$$g'_i(y, m) = \frac{\sum_{t=1}^{12} \sum_{k \in \mathcal{M}^i_{(y-1,m+t)}} n_{ik} * (1 - f'_{ijk})}{\sum_{t=1}^{12} \sum_{k \in \mathcal{M}^i_{(y-1,m+t)}} n_{ik}}.$$

Once again, n_{ik} denotes the number of service points played by player i in match k. In calculating f'_i, we use n_{jk} to denote the number of return points played by player i in match k.[5] Calculated from historical data at the time each match, f'_{ijk}, g'_{ijk} represent the opponent-adjusted likelihood of player i winning a point against player j in match k on serve and return, respectively.

Since the formula considers each player's opponent-adjusted ratings at the time of each match, we must compute ratings in chronological order. Similarly to Elo, we initialize all players' opponent-adjusted ratings to f_{av}, g_{av} before iterating through all tour-level matches 1968-present and calculating player ratings for each match accordingly.

3.2. Klaassen-Magnus elo ratings

Before Barnett and Clarke's approach, Klaassen and Magnus (2001) suggested a method to infer serving probabilities from a pre-match win forecast π_{ij}. As $b = f_{ij} + f_{ji}$ represents the overall serve ability between two players, they impose the constraint that any new serve parameters f'_{ij}, f'_{ji} must satisfy $f'_{ij} + f'_{ji} = b$. Using this, we may create a one-to-one function $S : (\pi_{ij}, b) \to (f'_{ij}, f'_{ji})$, which generates serving probabilities f'_{ij}, f'_{ji} for both players such that $P_m(0, 0) = \pi_{ij}$.

As this paper was published in 2002, Klaassen and Magnus produced serve parameters from ATP rank-based forecasts. However, given that Elo has since been demonstrated to outperform ATP rank in predicting match outcomes, we apply this method with Elo forecasts.

Even when we impose the constraint $f_{ij} + f_{ji} = b$, our hierarchical Markov Model's match probability equation has no analytical solution to its inverse. Therefore, we turn to the following approximation algorithm to generate serving percentages that correspond to a win probability within ϵ of our Elo forecast:

Algorithm 1 Klaassen-Magnus Elo Serve Parameters

```
procedure EloServeProbabilities(π, b, ε)
    f ← b/2
    diff ← b/4
    currentProb ← .5
    while |currentProb - π| > ε do:
        If currentProb < π then
            f += diff
        else
            f -= diff
        diff = diff/2
        currentProb ← matchProb(f, b - f)
    return f, b, −f
```

To generate serve probabilities for a given match, we first compute π_{ij} as player i's chance of victory against player j given their Elo ratings and f_{ij}, f_{ji} as specified by any Barnett-Clarke variation in section 3.1. Then we run the above algorithm with $\pi = \pi_{ij}$, $b = f_{ij} + f_{ji}$, and ϵ set to a desired precision level.[6] At each step, we call *matchProb()* to compute the win probability from the start of the match if player i and player j had serve parameters $f_{ij} = f$, $f_{ji} = b - f$, respectively. Then we compare *currentProb* to *prob* and increment f by *diff*, which halves at every iteration. This process continues until the serve parameters $f, b - f$ correspond to a win probability within ϵ of π_{ij}, taking $O(\log \frac{1}{\epsilon})$ calls to *matchProb()*.

Given any pre-match forecast π_{ij}, we can produce f'_{ij}, f'_{ji} consistent with π_{ij}, according to our hierarchical Markov Model. While Kovalchik and Reid (2018) recently outlined a similar method for inferring these parameters from Elo ratings subject to difference in serve ability, $f_{ij} - f_{ji} = \delta$, we set the constraint $f'_{ij} + f'_{ji} = b$ to ensure that overall serve ability agrees with historical data. As Klaassen and Magnus (2003) demonstrated, b encodes important information regarding likely trajectories of a match score and the relative importance of service breaks.

[5] The number of player j's service points in match k is equal to the number of player i's return points in match k.

[6] For this project, we set $\epsilon = .001$.

Therefore, keeping f'_{ij}, f'_{ji} consistent with b more naturally lends itself to in-match prediction, a clear future application of methods explored here.

Most importantly, this approach allows us to generate serve parameters from forecasts that are not point-based. While Kovalchik (2016) explored a variety of point-based methods specific to tennis' scoring system, none outperformed Elo ratings in predicting match outcomes. However, Elo ratings alone lack sufficient context to predict outcomes at a point level. Using the above approach, we significantly expand the possibilities when producing serve parameters for a given match. Should methods superior to Elo arise in the future, we may similarly intuit f'_{ij}, f'_{ji} from their match forecasts.

4. Case studies

The following examples illustrate applications of newly proposed methods in several ATP tour-level matches.

4.1. Efron-Morris estimator

To see how the Efron-Morris estimator makes our model robust to small sample sizes, consider the following match. When Daniel Elahi (COL) and Ivo Karlovic (CRO) faced off at ATP Bogota 2015, Elahi had played only one tour-level match in the past year. From a previous one-sided victory, his year-long percentage of service points won, $f_i = .7969$, was abnormally high compared to the year-long tour-level average of $f_{av} = .6423$.

player name	Daniel Elahi	Ivo Karlovic
service points won	51	3516
service points	64	4654
f_i	.7969	.7555
return points won	22	1409
return points	67	4903
g_i	.3284	.2874
Elo rating	1585.93	1952.86

Following Barnett and Clarke's method, we predict Elahi to win 89.25% of points on serve, which eclipses Karlovic's forecast of 81.01%.

$$f_{ij} = f_t + (f_i - f_{av}) - (g_j - g_{av}) = .8925$$
$$f_{ji} = f_t + (f_j - f_{av}) - (g_i - g_{av}) = .8101$$

Given that Karlovic is one of the most effective servers in the history of the game, this estimate seems unrealistic. From the serving stats, our hierarchical Markov Model computes Elahi's win probability as $\pi_{ij} = .8095$, mainly in consequence of only having collected his player statistics for one match. On the other hand, Karlovic becomes a strong favorite when we calculate Elahi's win probability via Elo ratings:

$$d = \frac{E_j(t) - E_i(t)}{400} = \frac{1952.86 - 1585.93}{400} = .9173$$
$$\hat{\pi}_{ij}(t) = (1 + 10^d)^{-1} = .1079.$$

This leads us to further question validity of this approach when using limited historical data. Thus, we turn to the Efron-Morris estimator to shrink Elahi's serve and return parameters toward f_{av}, g_{av}.

$$f'_i = f_i + B_i(f_{av} - f_i) = .6599$$
$$f'_j = f_j + B_j(f_{av} - f_j) = .7480$$
$$g'_i = g_i + B_i(g_{av} - g_i) = .3648$$
$$g'_j = g_j + B_j(g_{av} - g_j) = .2935$$
$$f'_{ij} = f_t + (f'_i - f_{av}) - (g'_j - g_{av}) = .7495$$
$$f'_{ji} = f_t + (f'_j - f_{av}) - (g'_i - g_{av}) = .7663$$

Above, we can see that the Efron-Morris estimator shrinks Elahi's stats far more than Karlovic's, since Karlovic has played many more tour-level matches in the past year. Given f'_i, f'_j, we compute $\pi_{ij} = .4277$. By shrinking the serve and return parameters, our model normalizes Elahi's inflated f_i and demonstrates robustness to small sample sizes.

Figure 1 illustrates how normalization coefficient B_i varies with n throughout our dataset. When n is small, as in Elahi's case, the Efron-Morris estimator

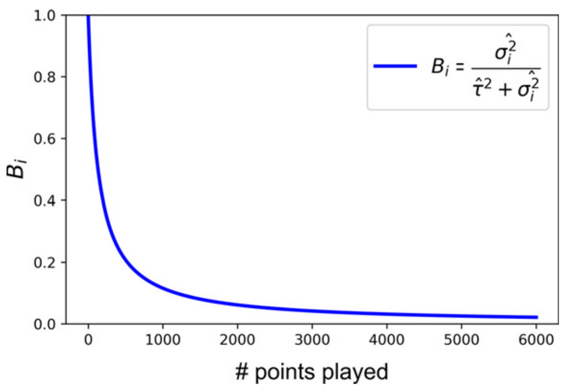

Fig. 1. Strength of B_i against number of points played with $f_i = .64$, $\hat{\tau}^2$=.00176.

strongly shrinks estimates toward the mean. Over the first one thousand points of play, however, B_i sharply declines in strength, approaching zero as match history accumulates further.

4.2. Opponent-adjusted ratings

To illustrate the effect of tracking opponent-adjusted statistics, we consider the 2014 US Open first-round match between Mikhail Youzhny (RUS) and Nick Kyrgios (AUS).

player name	Mikhail Youzhny	Nick Kyrgios
service points won	1828	900
service points	2960	1370
f_i	.6176	.6569
return points won	1145	424
return points	2947	1323
g_i	.3885	.3205
Elo rating	1941.65	1931.07

From the standard approach, with $f_t = .6583$, we compute $f_{ij} = .6446$, $f_{ji} = .6159$. To then calculate opponent-adjusted statistics, we must consider each player's expected performance given their past opponents. As we observe, Kyrgios has faced slightly stronger opponents than Youzhny over the past twelve months.

player name	Mikhail Youzhny	Nick Kyrgios
g_i'	.4332	.4486
$(1-g_i') * \sum n_{ik}$	1677.62	755.44
$f_i - (1-g_i')$	.0508	.1055
f_i'	.6982	.7237
$(1-f_i') * \sum n_{ik}$	889.32	365.59
$g_i - (1-f_i')$	.0868	.0441

In the above table, g_i' represents overall return ability of player i's opponents in the past twelve months. $(1-g_i') * \sum n_{ik}$ then represents the expected number of points won on serve given player i's match history and $f_i - (1-g_i')$ the performance relative to this baseline. Opponent-adjusted serve probabilities f_{ij}', f_{ji}' are calculated as follows:

$$f_{ij}' = f_t + (f_i - (1-g_i')) - (g_j - (1-f_j')) = .6280$$

$$f_{ji}' = f_t + (f_j - (1-g_j')) - (g_i - (1-f_i')) = .6413.$$

Using regular serve parameters, Youzhny is favored to win with $\pi_{ij} = .6410$. With opponent-adjusted serving percentages, we factor in the stronger opponents in Kyrgios' match history and Youzhny's win probability drops to $\pi_{ij} = .4334$. As we see, opponent-adjusted ratings can turn around forecasts when one player has faced tougher opponents over the past twelve months.

4.3. Klaassen-Magnus elo ratings

Finally, we demonstrate the use of Klaassen-Magnus Elo ratings. In the quarterfinals of the 2016 Olympics at Rio de Janeiro, Kei Nishikori (JP) faced off against Gael Monfils (FRA).

player name	Kei Nishikori	Gael Monfils
service points won	3309	2345
service points	5069	3533
f_i	.6528	.6637
return points won	2103	1433
return points	5229	3608
g_i	.4021	.3972
Elo rating	2295.56	2140.56

Though Elo ratings clearly favored Nishikori, serve/return statistics over the past twelve months put Monfils at a slight advantage.

$$f_{ij} = f_t + (f_i - f_{av}) - (g_j - g_{av}) = .6204$$

$$f_{ji} = f_t + (f_j - f_{av}) - (g_i - g_{av}) = .6263$$

To produce serve parameters that reflect Nishikori's Elo advantage and the pair's overall serve ability, we implement the approach from Section 3.3 with $\pi_{ij} = .7094$, $f_{ij} = .6204$, $f_{ji} = .6263$.

$$f_{ij}', f_{ji}' = S(\pi_{ij}, f_{ij} + f_{ji}) = .6451, .6016$$

Now we can forecast match outcomes with the hierarchical Markov Model, using serve parameters that respect each player's Elo rating.

5. Results

We evaluated methods across three years of tour-level matches, including Barnett and Clarke's standard approach and Knottenbelt's Common-Opponent Model, outlined in Appendix A, as benchmarks. Across the board, Klaassen-Magnus Elo ratings fared the best, while Efron-Morris estimators improved performance to a varying degree.

5.1. Dataset

This project drew from a publicly available match dataset (Sackmann, 2018). Match summary statistics cover over 150,000 ATP tour-level matches dating back to 1968. Relevant features included:

- match date, tournament, surface type, player names
- match serve/return statistics

The matches in this repository comprised dataset $\mathcal{M}$, from which we determined players' historical serve/return performance and Elo ratings. While all methods generated serve parameters from historical data, none required a training set for tuning hyper-parameters. Excluding Davis Cup matches,[7] we designated all matches from 2014-16 as test set $\mathcal{M}_t$ and produced serve parameters for each match in $\mathcal{M}_t$ based only on prior matches. Implementations of all methods in this paper may be found at https://github.com/jgollub1/tennis_match_prediction.

5.2. Evaluation

All methods produced parameters f_{ij}, f_{ji} to estimate each player's probability of winning a point on serve in a given match. We evaluated methods by predicting both the winner of every match and each player's percentage of points won on serve. To evaluate serve performance prediction, we considered the RMSE (root-mean-square-error) between each method's parameters and observed performance. By observing the proportion of points won on serve by each player for a single match k, we obtained s_{ik}, s_{jk} and realized error terms e_{ik}, e_{jk}:

$$s_{ik} = \frac{w_{ik}}{n_{ik}}, s_{jk} = \frac{w_{jk}}{n_{jk}}$$
$$e_{ik} = |s_{ik} - f_{ij}|, e_{jk} = |s_{jk} - f_{ji}|.$$

Over test set $\mathcal{M}_t$, a method's RMSE computed to:

$$r = \sqrt{\frac{\sum_{k \in \mathcal{M}_t} e_{ik}^2 + e_{jk}^2}{2|\mathcal{M}_t|}}.$$

To produce match win forecasts, we returned to the hierarchical Markov Model. As a function of a method's parameters f_{ij}, f_{ji} we calculated match win probability π_{ij} recursively from the probabilities of winning sets and games, as described in Section 2.1. Then we computed accuracy and log loss by comparing these forecasts with observed wins or losses for each match in $\mathcal{M}_t$. In measuring accuracy, we classified π_{ij} as a predicted win when $\pi_{ij} \geq .5$ and a loss otherwise.

5.3. Discussion

We evaluated methods across 7,622 ATP tour-level matches from 2014-16. To compute Klaassen-Magnus Elo ratings, we set the constraint $b = f_{ij} + f_{ji}$, where f_{ij}, f_{ji} were computed with the Efron-Morris Estimator, as described in Section 3.1.1. Because a player's performance can vary significantly across surfaces, we included a Barnett and Clarke variation which only considers performance on the given match's surface.[8] In Table 1 and Table 2, "EM" denotes a variation that used the Efron-Morris estimator to compute f_i, g_i in Barnett and Clarke's equation.

Table 1 displays each method's RMSE in predicting the proportion of points won on serve. Of all Barnett-Clarke variations, the surface-based estimator fared worst. This was likely due to decreased sample sizes, as filtering matches by surface limited the amount of available data. Its Efron-Morris variant performed significantly better in all years, suggesting there may be value in a surface-specific approach when bias from limited data is offset. However, this variant only came close to reaching parity with the standard Barnett-Clarke model in the year 2014.

While it was intended to address issues with varying strength of schedule, the Common-Opponent Model underperformed all other methods in predicting serve performance. When two players shared few opponents across their match histories, this approach proved fairly susceptible to bias. Furthermore, while Barnett-Clarke variants considered the past twelve months of data, this approach considered all historical matches, curbing its ability to express recent trends as players' match histories accumulated. On the other hand, opponent-adjusted ratings demonstrated an improvement to both Barnett-Clarke and the Common-Opponent Model. By estimating serve and return ability on a continuous scale, opponent-adjusted ratings allowed us to consider all matches within the last twelve months and avoid a major shortcoming of the Common-Opponent Model.

[7] Davis cup matches frequently involve lower-ranked players.

[8] All matches occurred on hard, clay, or grass courts.

Table 1
Serve performance prediction of 2014-16 ATP matches

Variation	2014 n=(2,488) RMSE	2015 (n=2,540) RMSE	2016 (n=2,594) RMSE
Barnett-Clarke	.0846	.0916	.0845
Barnett-Clarke EM	.0823	.0909	.0823
Barnett-Clarke surface	.0883	.0968	.0904
Barnett-Clarke surface EM	.0850	.0944	.0872
Barnett-Clarke opponent-adjusted	.0829	.0898	.0825
Barnett-Clarke opponent-adjusted EM	.0821	.0894	.0818
Klaassen-Magnus Elo	.0804	.0890	.0798
Common-Opponent	.0957	.1046	.0943

Table 2
Match win prediction of 2014-16 ATP matches

Variation	2014 n=(2,488) Accuracy	Log Loss	2015 (n=2,540) Accuracy	Log Loss	2016 (n=2,594) Accuracy	Log Loss
Barnett-Clarke	64.8	.648	65.0	.634	64.6	.663
Barnett-Clarke EM	65.6	.613	65.2	.609	64.7	.628
Barnett-Clarke surface	63.3	.705	63.2	.712	62.0	.742
Barnett-Clarke surface EM	63.5	.628	62.9	.628	62.5	.645
Barnett-Clarke opponent-adjusted	67.8	.637	68.0	.613	67.4	.641
Barnett-Clarke opponent-adjusted EM	67.8	.625	68.1	.605	67.6	.632
Klaassen-Magnus Elo	69.1	.589	69.3	.579	69.7	.594
Common-Opponent	63.6	.628	64.5	.627	63.4	.650

Overall, Klaassen-Magnus Elo ratings proved most effective in forecasting serve performance. As Elo ratings were already known to estimate player ability more effectively than point-based models, it follows that adjusting serve parameters in accordance with these ratings would significantly improve forecasts. In contrast, Barnett-Clarke variations still appear to leave out important information by ignoring match outcomes and forecasting strictly from point-level data. While point-based models have often proven necessary for predicting more granular outcomes (Barnett et al., 2006), Klaassen-Magnus Elo ratings reconciled point-based models' expressivity with Elo ratings' predictive power. Following the recent work of Kovalchik and Reid (2018), this further demonstrated the effectiveness of synthesizing Elo ratings with point-level models specific to tennis' scoring system.

When applied to Barnett-Clarke variants, the Efron-Morris estimator improved performance to varying degrees. Presumably, this stemmed from robustness with respect to small sample sizes. To better understand its application to our tour-level match dataset, we consider the distribution of points within player match histories.

Figure 2 illustrates the distribution of sample sizes when computing f_i, f_j with an Efron-Morris estimator.[9] Following Barnett and Clarke's original approach, estimates produced with sample sizes from the leftmost range of the distribution were particularly susceptible to bias. However, recalling normalization strength from Figure 1, we know the Efron-Morris estimator reduced bias by shrinking these estimates furthest toward the mean. Since players accumulate return points over time at approximately the same

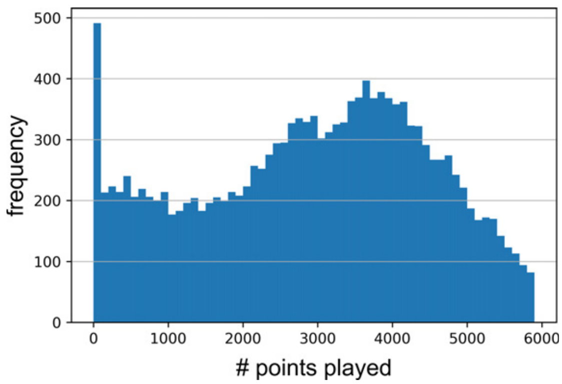

Fig. 2. Number of points played on serve by players i, j in the past twelve months before every match in $\mathcal{M}_t$.

[9]For given match k with serve parameters $f_{ik} = \frac{w_{ik}}{n_{ik}}$, $f_{jk} = \frac{w_{jk}}{n_{jk}}$, we considered n_{ik}, n_{jk}.

rate, we can assume the estimator normalized g_i, g_j in a similar manner. If only by reducing errant predictions in this range, the Efron-Morris estimator succeeded in helping all Barnett-Clarke variants more effectively predict serve performance.

Table 2 details each method's performance in predicting match winners, where many of the same trends surfaced. Once again, variations with an Efron-Morris estimator outperformed their counterparts. In most cases, the improvement in log loss was greater than that of accuracy, supporting the notion that this estimator mitigates extreme results when faced with limited data. The opponent-adjusted model performed several percentage points better than Barnett and Clarke's original approach across all years, establishing itself as the most effective point-based method surveyed in this exploration. Once again, Klaassen-Magnus Elo ratings predicted outcomes most effectively. However, it is worth noting that its output produced win probabilities identical to those of the Elo ratings fed into the model. In that regard, its relative performance confirmed results from previous studies noting Elo's dominance in match prediction (Kovalchik, 2016).

6. Conclusion

Although Barnett and Clarke's approach has long been established in tennis match prediction, there are many ways to forecast serve performance. In order to extract more information from historical match data, we outlined a handful of novel approaches to generating the parameters in their equation. Using an Efron-Morris estimator improved performance across the board, emphasizing the need to address scenarios with limited match data. The opponent-adjusted model demonstrated a new way to quantify strength of schedule through point-level data, resulting in significantly improved performance over Barnett-Clarke and the Common-Opponent Model. Finally, Klaassen-Magnus Elo ratings synthesized historical serve data with the predictive power of Elo ratings while maintaining the overall serve ability between two players.

A clear next step to this exploration would involve applying these methods to in-match prediction. With point-by-point datasets available today, we can use these same methods to forecast match outcomes while play is in progress and set similar benchmarks with tour-level match data.

Acknowledgments

I would like to thank Kevin Rader for his guidance on earlier iterations of this research. I am also grateful to Jeff Sackmann for making all relevant match data publicly available.

References

Barnett, T. and Clarke, S. R. 2005, Combining player statistics to predict outcomes of tennis matches, *IMA Journal of Management Mathematics*, *16*(2), 113-120.

Barnett, T. J., Clarke, S. R., et al. 2002, Using microsoft excel to model a tennis match. In 6th Conference on Mathematics and Computers in Sport, pages 63-68. Queensland, Australia: Bond University.

Barnett, T. J. et al. 2006, Mathematical modelling in hierarchical games with specific reference to tennis. PhD thesis, Swinburne University of Technology.

Bialik, C., Morris, B., and Boice, J. 2016, How we're forecasting the 2016 u.s. open. http://fivethirtyeight.com/features/how-were-forecasting-the-2016-us-open/. Accessed: 2017-10-30.

Efron, B. and Morris, C. 1975, Data analysis using stein's estimator and its generalizations, *Journal of the American Statistical Association*, *70*(350), 311-319.

Efron, B. and Morris, C. N. 1977, Stein's paradox in statistics. WH Freeman.

Elo, A. E. 1978, The rating of chessplayers, past and present. Arco Pub.

Klaassen, F. J. and Magnus, J. R. 2003, Forecasting the winner of a tennis match, *European Journal of Operational Research*, *148*(2), 257-267.

Klaassen, F. J. G. M. and Magnus, J. R. 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), 500-509.

Knottenbelt, W. J., Spanias, D., and Madurska, A. M. 2012, A common-opponent stochastic model for predicting the outcome of professional tennis matches, *Computers & Mathematics with Applications*, *64*(12), 3820-3827.

Kovalchik, S. and Reid, M. 2018, A calibration method with dynamic updates for within-match forecasting of wins in tennis, *International Journal of Forecasting*.

Kovalchik, S. A. 2016, Searching for the goat of tennis win prediction, *Journal of Quantitative Analysis in Sports*, *12*(3), 127-138.

Newton, P. K. and Aslam, K. 2009, Monte carlo tennis: a stochastic markov chain model, *Journal of Quantitative Analysis in Sports*, *5*(3).

Sackmann, J. 2018, Tennis atp. https://github.com/JeffSackmann/tennis_atp.

Spanias, D. and Knottenbelt, W. J. 2012, Predicting the outcomes of tennis matches using a low-level point model, *IMA Journal of Management Mathematics*, *24*(3), 311-320.

Appendix A: Common-Opponent Model

Consider a match between players i and j, who share N common opponents throughout their entire match histories. To quantify their relative performance against opponent C_k, let $spw(i, C_k)$ denote the percentage of service points won by player i against opponent C_k and $rpw(i, C_k)$ the corresponding percentage of return points won. We may quantify player i's advantage over player j with respect to opponent C_k as follows:

$$\Delta_k^{ij} = (spw(i, C_k) - (1 - rpw(i, C_k)) - (spw(j, C_k) - (1 - rpw(j, C_k)).$$

Then we approximate match win probability in terms of this relative advantage.

$$P(i \text{ beats } j \text{ via } C_k) = \frac{M_3(.6 + \Delta_k^{ij}, .4) + M_3(.6, .4 + \Delta_k^{ij})}{2}$$

In the above equation, $M_3(f, g)$ denotes player i's win probability in a best-of-three match with f, g representing their probability of winning a point on serve and return, respectively. To compute π_{ij} via the Common-Opponent Model, we average the win probability over all N common opponents shared:

$$\pi_{ij} = \frac{\sum_{k=1}^{N} P(i \text{ beats } j \text{ via } C_k)}{N}.$$

While Knottenbelt did not originally use this model to predict serve performance, we inferred parameters f_{ij}, f_{ji} from the model's win probability equation as followed:

$$\Delta^{ij} = \frac{\sum_{k=1}^{N} \Delta_k^{ij}}{N}$$

$$f_{ij} = .6 + \Delta^{ij}/2$$

$$f_{ji} = .6 - \Delta^{ij}/2.$$

Following the above steps, we may predict the winner and serve performance of any match using the Common-Opponent Model.