Forecasting serve performance in professional tennis matches

Jacob Gollub

Harvard University, Department of Statistics

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-
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.\textsuperscript{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 within a match. Section 4 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, $\pi_{ij}$, 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:

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

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:

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

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

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

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

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 initialize 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.

\[ f_{ij} = f_i + (f_i - f_{av}) - (g_j - g_{av}) \]
\[ f_{ji} = f_i + (f_j - f_{av}) - (g_i - g_{av}) \]

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 \( M_{(y,m)}^i \) represents the set of player i’s matches in year y, month m, we obtain the following statistics\(^4\):

\[ f_i(y, m) = \frac{\sum_{t=1}^{12} \sum_{k \in M_{(y,m+1)}^i} w_{jk}}{\sum_{t=1}^{12} \sum_{k \in M_{(y,m+1)}^i} n_{jk}} \]
\[ g_i(y, m) = \frac{\sum_{t=1}^{12} \sum_{k \in M_{(y,m+1)}^i} n_{jk} - w_{jk}}{\sum_{t=1}^{12} \sum_{k \in M_{(y,m+1)}^i} n_{jk}} \]

w_{jk} denotes the number of service points won by player i in match k and n_{jk} 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 \( M_{(y,m)} \) represents the set of all matches played at tournament v in year y:

\[ f_t(v, y) = \frac{\sum_{k \in M_{(y-1)}^v} w_k}{\sum_{k \in M_{(y-1)}^v} n_k} \]

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 \( M_{(y,m)} \) represents the set of tour-level matches played in year y, month m:

\[ f_{av}(y, m) = \frac{\sum_{t=1}^{12} \sum_{k \in M_{(y-1,m+1)}^i} w_k}{\sum_{t=1}^{12} \sum_{k \in M_{(y-1,m+1)}^i} n_k} \]
\[ g_{av}(y, m) = 1 - f_{av}(y, m). \]

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.\(^4\)

\(^2\)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 return to the Efron-Morris estimator to provide alternative parameters of the form:

\[ f_{ij} = f_i + (f'_i - f_{av}) - (g'_j - g_{av}), \]

\[ f'_{ij} = f_i + (f'_i - f_{av}) - (g'_j - g_{av}). \]

Rather than directly apply the estimator to \( f_{ij}, f'_{ij}, \) 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:

\[ \mathcal{F} = \bigcup_{k \in \mathcal{M}} \mathcal{F}_k \]

\[ \mathcal{F}_k = \{ f_{ik}, f_{jk} \}. \]

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 \( \theta_i \) as follows:

\[ f_i | \theta_i \sim N(\theta_i, \sigma^2_i) \]

\[ \theta_i \sim N(\mu, \tau^2). \]

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

\[ \theta_i | f_i \sim N(f_i + \hat{B}_i(\mu - f_i), \sigma^2_i(1 - \hat{B}_i)) \]

\[ \hat{B}_i = \frac{\sigma^2_i}{\tau^2 + \sigma^2_i}. \]

Normalization coefficient \( B_i \) depends on both \( \tau^2 \), the variance of true service ability, and \( \sigma^2_i \), the variance of \( f_i \) given true service ability \( \theta_i \). 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:

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

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

Then, using maximum likelihood estimation, we estimate \( B_i \) and \( \sigma^2_i \) in order to produce the Efron-Morris estimator:

\[ \hat{B}_i = \frac{\hat{\sigma}^2_i}{\tau^2 + \hat{\sigma}^2_i} \]

\[ \hat{\sigma}^2_i = \hat{f}_i(1 - \hat{f}_i) \]

\[ f'_i = \hat{f}_i + \hat{B}_i(f_{av} - \hat{f}_i). \]

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 \( g'_j \) in place of \( f'_j \), 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:

\[ f'_{ij} = f_i + (f_i - (1 - g'_j)) - (g_j - (1 - f'_j)) \]

\[ f'_{ji} = f_i + (f_j - (1 - g'_j)) - (g_i - (1 - f'_j)). \]

\[ f'_i, g'_i \] represent the average serve and return abilities of player \( i \)'s opponent 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:

\[
\begin{align*}
    f'_i(y, m) = & \frac{\sum_{t=1}^{12} \sum_{k \in \mathcal{M}_i^{y-1,m+t}} n_{jk} \cdot (1 - g'_{ijk})}{\sum_{t=1}^{12} \sum_{k \in \mathcal{M}_i^{y-1,m+t}} n_{jk}} \\
    g'_i(y, m) = & \frac{\sum_{t=1}^{12} \sum_{k \in \mathcal{M}_i^{y-1,m+t}} n_{ik} \cdot (1 - f'_{ijk})}{\sum_{t=1}^{12} \sum_{k \in \mathcal{M}_i^{y-1,m+t}} n_{ik}}.
\end{align*}
\]

Once again, \( n_{jk} \) 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 \( \pi_{ij} \). 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 \( S : (\pi_{ij}, b) \rightarrow (f'_{ij}, f'_{ji}) \), which generates serving probabilities \( f'_{ij}, f'_{ji} \) for both players such that \( P_b(0, 0) = \pi_{ij} \).

As this paper was published in 2002, Klaassen and Magnus produced serve parameters at 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 \( \epsilon \) of our Elo forecast:

```
Algorithm 1 Klaassen-Magnus Elo Serve Parameters

procedure EloServeProbabilities(\( \pi, b, \epsilon \))
    \( f \leftarrow b/2 \)
    diff \( \leftarrow b/4 \)
    currentProb \( \leftarrow 0.5 \)
    while |currentProb - \( \pi \)| > \( \epsilon \) do:
        if currentProb < \( \pi \) then
            \( f \leftarrow f + \text{diff} \)
        else
            \( f \leftarrow f - \text{diff} \)
        currentProb \( \leftarrow \text{matchProb}(f, b - f) \)
    return \( f, b - f \)
```

To generate serve probabilities for a given match, we first compute \( \pi_{ij} \) 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 \( \epsilon \) set to a desired precision level.\(^6\) At each step, we call \( \text{matchProb}(f) \) 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 \( \text{currentProb} \) to \( \text{prob} \) and increment \( f \) by \( \text{diff} \), which halves at every iteration. This process continues until the serve parameters \( f, b - f \) correspond to a win probability within \( \epsilon \) of \( \pi_{ij} \), taking \( O(\log \frac{1}{\epsilon}) \) calls to \( \text{matchProb}(f) \).

Given any pre-match forecast \( \pi_{ij} \), we can produce \( f'_{ij}, f'_{ji} \) consistent with \( \pi_{ij} \), 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 the constraint that \( f_{ij} - f_{ji} = \delta \), 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 serve 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 $f_{av} = .6423$.

<table>
<thead>
<tr>
<th>player name</th>
<th>Daniel Elahi</th>
<th>Ivo Karlovic</th>
</tr>
</thead>
<tbody>
<tr>
<td>service points won</td>
<td>51</td>
<td>3516</td>
</tr>
<tr>
<td>service points</td>
<td>64</td>
<td>4654</td>
</tr>
<tr>
<td>$f_i$</td>
<td>.7969</td>
<td>.7555</td>
</tr>
<tr>
<td>return points won</td>
<td>22</td>
<td>1409</td>
</tr>
<tr>
<td>return points</td>
<td>67</td>
<td>4903</td>
</tr>
<tr>
<td>$g_j$</td>
<td>.3284</td>
<td>.2874</td>
</tr>
<tr>
<td>Elo rating</td>
<td>1585.93</td>
<td>1952.86</td>
</tr>
</tbody>
</table>

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%.

\[
\begin{align*}
\hat{f}_{ij} &= f_i + (f_i - f_{av}) - (g_j - g_{av}) = .8925 \\
\hat{f}_{ji} &= f_i + (f_j - f_{av}) - (g_i - g_{av}) = .8101
\end{align*}
\]

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 $\pi_{ij} = .8995$, 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 serving and return probabilities toward $f_{av}, g_{av}$.

\[
\begin{align*}
f'_i &= f_i + B_i(f_{av} - f_i) = .6599 \\
f'_j &= f_j + B_j(f_{av} - f_j) = .7480 \\
g'_i &= g_i + B_i(g_{av} - g_i) = .3648 \\
g'_j &= g_j + B_j(g_{av} - g_j) = .2935 \\
f'_{ij} &= f_i + (f'_i - f_{av}) - (g'_j - g_{av}) = .7495 \\
f'_{ji} &= f_i + (f'_j - f_{av}) - (g'_i - g_{av}) = .7663
\end{align*}
\]

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 $f_i = .64, \tau^2 = .00176$.](image)
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).

<table>
<thead>
<tr>
<th>player name</th>
<th>Mikhail Youzhny</th>
<th>Nick Kyrgios</th>
</tr>
</thead>
<tbody>
<tr>
<td>service points won</td>
<td>1828</td>
<td>900</td>
</tr>
<tr>
<td>service points</td>
<td>.6176</td>
<td>.6569</td>
</tr>
<tr>
<td>return points won</td>
<td>1145</td>
<td>424</td>
</tr>
<tr>
<td>return points</td>
<td>.3885</td>
<td>.3205</td>
</tr>
<tr>
<td>Elo rating</td>
<td>1941.65</td>
<td>1931.07</td>
</tr>
</tbody>
</table>

From the standard approach, with $f_i = .6583$, 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.

<table>
<thead>
<tr>
<th>player name</th>
<th>Mikhail Youzhny</th>
<th>Nick Kyrgios</th>
</tr>
</thead>
<tbody>
<tr>
<td>$g_i$</td>
<td>.4332</td>
<td>.4486</td>
</tr>
<tr>
<td>$(1 - g_i)\sum n_{ik}$</td>
<td>1677.62</td>
<td>755.44</td>
</tr>
<tr>
<td>$f_i - (1 - g_i)$</td>
<td>.0508</td>
<td>.1055</td>
</tr>
<tr>
<td>$f_i'$</td>
<td>.6982</td>
<td>.7237</td>
</tr>
<tr>
<td>$(1 - f_i')\sum n_{ik}$</td>
<td>889.32</td>
<td>365.59</td>
</tr>
<tr>
<td>$g_i - (1 - f_i')$</td>
<td>.0868</td>
<td>.0441</td>
</tr>
</tbody>
</table>

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:

$f_{ij}' = f_i + (f_i - f_{av}) - (g_j - g_{av}) = .6280$

$f_{ji}' = f_i + (f_j - f_{av}) - (g_i - g_{av}) = .6413$.

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).

<table>
<thead>
<tr>
<th>player name</th>
<th>Kei Nishikori</th>
<th>Gael Monfils</th>
</tr>
</thead>
<tbody>
<tr>
<td>service points won</td>
<td>5709</td>
<td>2345</td>
</tr>
<tr>
<td>service points</td>
<td>.6528</td>
<td>.6637</td>
</tr>
<tr>
<td>return points won</td>
<td>2103</td>
<td>1433</td>
</tr>
<tr>
<td>return points</td>
<td>5229</td>
<td>3608</td>
</tr>
<tr>
<td>Elo rating</td>
<td>2295.56</td>
<td>2140.56</td>
</tr>
</tbody>
</table>

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

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

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

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 $\mathcal{M}_t$, 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, 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 on prior matches. Implementations of all methods in this paper may be found at .

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 for each player for a single match $k$, we obtained $s_{ik}, s_{jk}$ and realized error terms $e_{ik}, e_{jk}$:

$$ s_{ik} = \frac{w_{jk}}{n_{ik}}, s_{jk} = \frac{w_{jk}}{n_{jk}}, $$

$$ e_{ik} = |s_{ik} - f_{ij}|, e_{jk} = |s_{jk} - f_{ji}|. $$

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

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

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 $\pi_{ij}$ 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 constant $b = f_{ij} + f_{ji}$, 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. In Table 1 and Table 2, “EM” denotes a variation that used the Efron-Morris estimator to compute $f_{ij}, f_{ji}$ 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 worse. 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.

Overall, Klaassen-Magnus Elo ratings proved most effective in forecasting serve performance.

---

7Davis cup matches frequently involve lower-ranked players.

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

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

---

**Table 1**

Serve performance prediction of 2014-16 ATP matches

<table>
<thead>
<tr>
<th>Variation</th>
<th>2014 RMSE</th>
<th>2015 RMSE</th>
<th>2016 RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barnett-Clarke</td>
<td>.0846</td>
<td>.0916</td>
<td>.0845</td>
</tr>
<tr>
<td>Barnett-Clarke EM</td>
<td>.0823</td>
<td>.0909</td>
<td>.0823</td>
</tr>
<tr>
<td>Barnett-Clarke surface</td>
<td>.0883</td>
<td>.0968</td>
<td>.0904</td>
</tr>
<tr>
<td>Barnett-Clarke surface EM</td>
<td>.0850</td>
<td>.0944</td>
<td>.0872</td>
</tr>
<tr>
<td>Barnett-Clarke opponent-adjusted</td>
<td>.0829</td>
<td>.0898</td>
<td>.0825</td>
</tr>
<tr>
<td>Barnett-Clarke opponent-adjusted EM</td>
<td>.0821</td>
<td>.0894</td>
<td>.0818</td>
</tr>
<tr>
<td>Klaassen-Magnus Elo</td>
<td>.0804</td>
<td>.0890</td>
<td>.0798</td>
</tr>
<tr>
<td>Common-Opponent</td>
<td>.0957</td>
<td>.1046</td>
<td>.0943</td>
</tr>
</tbody>
</table>

**Table 2**

Match win prediction of 2014-16 ATP matches

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Barnett-Clarke</td>
<td>64.8</td>
<td>.648</td>
<td>65.0</td>
<td>.634</td>
<td>64.6</td>
<td>.663</td>
</tr>
<tr>
<td>Barnett-Clarke EM</td>
<td>65.6</td>
<td>.613</td>
<td>65.2</td>
<td>.609</td>
<td>64.7</td>
<td>.628</td>
</tr>
<tr>
<td>Barnett-Clarke surface</td>
<td>63.3</td>
<td>.705</td>
<td>63.2</td>
<td>.712</td>
<td>62.0</td>
<td>.742</td>
</tr>
<tr>
<td>Barnett-Clarke surface EM</td>
<td>63.5</td>
<td>.628</td>
<td>62.9</td>
<td>.628</td>
<td>62.5</td>
<td>.645</td>
</tr>
<tr>
<td>Barnett-Clarke opponent-adjusted</td>
<td>67.8</td>
<td>.637</td>
<td>68.0</td>
<td>.613</td>
<td>67.4</td>
<td>.641</td>
</tr>
<tr>
<td>Barnett-Clarke opponent-adjusted EM</td>
<td>67.8</td>
<td>.625</td>
<td>68.1</td>
<td>.605</td>
<td>67.6</td>
<td>.632</td>
</tr>
<tr>
<td>Klaassen-Magnus Elo</td>
<td>69.1</td>
<td>.589</td>
<td>69.3</td>
<td>.579</td>
<td>69.7</td>
<td>.594</td>
</tr>
<tr>
<td>Common-Opponent</td>
<td>63.6</td>
<td>.628</td>
<td>64.5</td>
<td>.627</td>
<td>63.4</td>
<td>.650</td>
</tr>
</tbody>
</table>

---

Fig. 2. Number of points played on serve by players $i, j$ in the past twelve months before every match in $M_t$. 

---
610 furthest toward the mean. Since players accumulate
611 return points over time at approximately the same
612 rate, we can assume the estimator normalized \( g_i, g_j \)
613 in a similar manner. If only by reducing errant pre-
614 dictions in this range, the Efron-Morris estimator
615 succeeded in helping all Barnett-Clarke variants more
616 effectively predict serve performance.
617
618 Table 2 details each method’s performance in pre-
619 dicting match winners, where many of the same
620 trends surfaced. Once again, variations with an Efron-
621 Morris estimator outperformed their counterparts. In
622 most cases, the improvement in log loss was greater
623 than that of accuracy, supporting the notion that
624 this estimator mitigates extreme results when faced
625 with limited data. The opponent-adjusted model per-
626 formed several percentage points better than Barnett
627 and Clarke’s original approach across all years,
628 establishing itself as the most effective point-based
629 method surveyed in this exploration. Once again,
630 Klaassen-Magnus Elo ratings predicted outcomes
631 most effectively. However, it is worth noting that its
632 output produced win probabilities identical to the Elo
633 ratings fed into the model. In that regard, its rela-
634 tive performance confirmed results from previous
635 studies noting Elo’s dominance in match prediction
636 (Kovalchik, 2016).
637
638 6. Conclusion
639
640 Although Barnett and Clarke’s approach has long
641 been established in tennis match prediction, there are
642 many ways to forecast serve performance. In order
643 to extract more information from historical match
644 data, we outlined a handful of novel approaches to
645 generating the parameters in their equation. Using
646 an Efron-Morris estimator improved performance
647 across the board, emphasizing the need to address
648 scenarios with limited match data. The opponent-
649 adjusted model demonstrated a new way to quantify
650 strength of schedule through point-level data, result-
651 ing in significantly improved performance over
652 Barnett-Clarke and the Common-Opponent Model.
653 Finally, Klaassen-Magnus Elo ratings synthesized
654 historical serve data with the predictive power of Elo
655 ratings while maintaining the overall serve ability
656 between two players.
657
658 A clear next step to this exploration would involve
659 applying these methods to in-match prediction. With
660 point-by-point datasets available today, we can use
661 these same methods to forecast match outcomes
662 while play is in progress and set similar benchmarks
663 with tour-level match data.
664
665 References
666 Barnett, T. and Clarke, S. R. 2005, Combining player statistics to
667 predict outcomes of tennis matches, IMA Journal of Manage-
668 ment Mathematics, 16(2), 113-120.
670 to model a tennis match. In 6th Conference on Mathematics
671 and Computers in Sport, pages 63-68. Queensland, Australia:
672 Bond University.
673 Barnett, T. J. et al. 2006, Mathematical modelling in hierarchical
674 games with specific reference to tennis. PhD thesis, Swin-
675 burne University of Technology.
676 Bialik, C., Morris, B., and Boice, J. 2016, How we’re forecasting
677 the 2016 u.s. open. http://fivethirtyeight.com/features/how-
678 were-forecasting-the-2016-us-open/. Accessed: 2017-10-30.
679 Efron, B. and Morris, C. 1975, Data analysis using stein’s estimator
680 and its generalizations, Journal of the American Statistical
681 Association, 70(350), 311-319.
683 WH Freeman. Elo, A. E. 1978, The rating of chessplayers,
684 past and present. Arco Pub.
685 Klaassen, F. J. and Magnus, J. R. 2003, Forecasting the winner of
686 a tennis match, European Journal of Operational Research,
687 148(2), 257-267.
688 Klaassen, F. J. G. M. and Magnus, J. R. 2001, Are points in tennis
689 independent and identically distributed? evidence from a
690 dynamic binary panel data model, Journal of the American
691 Statistical Association, 96(454), 500-509.
692 Koontenbelt, W. J., Spanias, D., and Madurska, A. M. 2012,
693 A common-opponent stochastic model for predicting the
694 outcome of professional tennis matches, Computers & Math-
695 ematics with Applications, 64(12), 3820-3827.
696 Kovalchik, S. and Reid, M. 2018, A calibration method with
697 dynamic updates for within-match forecasting of wins in tennis,
698 International Journal of Forecasting.
699 Kovalchik, S. A. 2016, Searching for the goat of tennis win pre-
700 diction, Journal of Quantitative Analysis in Sports, 12(3),
701 127-138.
702 Newton, P. K. and Aslam, K. 2009, Monte carlo tennis: a stochastic
703 markov chain model, Journal of Quantitative Analysis in
704 Sports, 5(3).
706 Spanias, D. and Koontenbelt, W. J. 2012, Predicting the outcomes
707 of tennis matches using a low-level point model,IMA Journal
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:

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

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

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

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:

$$f_{ij} = .6 + \Delta_{ij}^k/2, \quad f_{ji} = .6 - \Delta_{ij}^k/2.$$  

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