Int. J. Appl. Math. Comput. Sci., 2021, Vol. 31, No. 3, 525–538
DOI: 10.34768/amcs-2021-0036

A COMBINATORIAL APPROACH IN PREDICTING THE OUTCOME OF TENNIS MATCHES

ANA ŠARČEVIĆ [a,*], MIHAELA VRANIĆ [a], DAMIR PINTAR [a]

[a]Faculty of Electrical Engineering and Computing
University of Zagreb
Unska 3, HR-10000 Zagreb, Croatia
e-mail: {ana.sarcevic,mihaela.vranic,damir.pintar}@fer.hr

Tennis, as one of the most popular individual sports in the world, holds an important role in the betting world. There are two main categories of bets: pre-match betting, which is conducted before the match starts, and live betting, which allows placing bets during the sporting event. Betting systems rely on setting sports odds, something historically done by domain experts. Setting odds for live betting represents a challenge due to the need to follow events in real-time and react accordingly. In tennis, hierarchical models often stand out as a popular choice when trying to predict the outcome of the match. These models commonly leverage a recursive approach that aims to predict the winner or the final score starting at any point in the match. However, recursive expressions inherently contain computational complexity which hinders the efficiency of methods relying on them. This paper proposes a more resource-effective alternative in the form of a combinatorial approach based on a binomial distribution. The resulting accuracy of the combinatorial approach is identical to that of the recursive approach while being vastly more efficient when considering the execution time, making it a superior choice for live betting in this domain.

Keywords: binomial distribution, final score prediction, independent and identical distribution, predictive model.

1. Introduction

Predicting the outcome of sporting events has always attracted the attention of a large number of people, from sports professionals and bookmakers to the general population. With the advancement of the Internet, betting on sports event outcomes has seen a dramatic surge in popularity. The Internet provides a more dynamic and practical way of betting while also offering the opportunity to place bets for an ongoing sporting event (so-called "live betting"). The European Gaming and Betting Association (EGBA) claims that, because of this technological change, Europe's online gambling market is growing at about 10% per year, faster than land-based gambling. According to the EGBA, the economic size (or gross gaming revenue) of the EU online sector is expected to rise from €22.2 billion in 2018 to €29.3 billion in 2022 (EGBA, 2020).

Due to its nature, live betting differs significantly from pre-match betting. In the latter, a common business practice is to calculate the initial odds based on historical data and adjust them according to the new knowledge of betting houses. In the former, there is a wealth of other information to consider, all of which may affect the change in the betting odds. For example, any change in the score might result in a change in odds, and those changes need to be both as accurate as possible as well as executed very quickly, often milliseconds after each score change.

One of the more popular betting sports is tennis. It is enjoyed by millions of viewers, who watch numerous matches throughout the year. Additionally, large and easily accessible datasets make tennis an attractive candidate for research in scientific papers. The nature of tennis itself also contributes to the popularity of tennis match modeling. Tennis is an example of a sport with a strongly defined structure and a rigid scoring system, making it relatively easy to model its matches in the form of discrete stochastic processes, a typical example of which is the Markovian process. Due to this fact, tennis is often categorized as a *discrete* or *Markovian sport*. Another typical example of such a sport is volleyball, also

*Corresponding author

a frequent subject of scientific papers.

Betting on a winner is only one type of popular sports betting. In addition to this, it is possible to bet on the exact score by which the match will end, on the total number of games or sets to be played, or on handicap (the difference in the final number of games or sets).

Previous works published on the subject of predicting the outcome of tennis matches mainly focus on predicting the winner of the match (Pollard, 1983; Liu, 2001; Newton and Keller, 2005; O'Malley, 2008; Croucher, 1986; Barnett and Clarke, 2002; Barnet *et al.*, 2006; Wozniak, 2011), and several papers focus on predicting the final score (Barnett and Brown, 2012). Among the more popular models for predicting the above are hierarchical tennis models. These leverage the structure of the scoring system in tennis and model the matches as Markov chains with transition probabilities obtained from historical player service statistics. Using two recursion approaches (forward recursion and backward recursion), it is possible to estimate the probability of a player winning the match at any time (live) or predict the final score of the match. However, the effectiveness of methods that rely on recursive expressions is questionable because of the computational complexity of these expressions. This paper proposes a more resource-effective alternative in the form of a combinatorial approach used to evaluate the final score of a match, which can then be used to predict the winner of the match and estimate the total number of points, games, and sets that will be played in the match. The approach has been validated in a manner common for such problems, and it has been demonstrated that, without compromising the accuracy of recursive approaches, the combinatorial approach significantly reduces the time required to generate the results.

Figure 1 gives an overview of the approach proposed in the paper. The model consists of two levels of formulae: a game level and a set level. Each receives as inputs the probabilities of winning points on the players' own services. These probabilities are called p and q, and are calculated from historical data. The probabilities of winning points on their own service can be updated with current data from the ongoing match in order to obtain more accurate statistics. Individual formulae are devised for predicting the final score in a game or set. In order to predict the final score in a match, it is necessary to combine game and set formulae. After evaluating the final result, it is possible to form predictions for the winner of the match, the total number of points, or the handicap.

The rest of this paper is organized as follows. In Section 2 related work and current solutions of the problem are presented. Section 3 gives a brief description of the scoring rules in tennis. Understanding the scoring rules is essential for comprehension of the rest of the paper. Section 4 describes the hierarchical combinatorial approach. Mathematical expressions are given and they are supported by examples. Section 5 evaluates the performance of the combinatorial approach, and Section 6 concludes the paper.

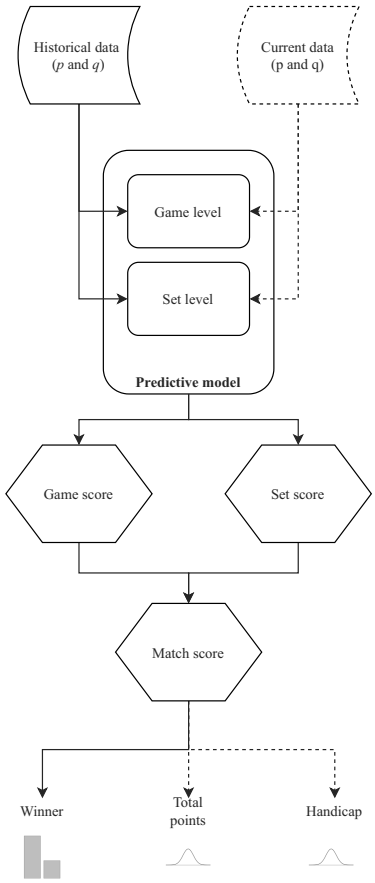

Fig. 1. Approach overview.

2. Related work

From a theoretical standpoint, tennis is a very attractive sport for modeling because it only takes two players into consideration and only one outcome is possible. It does not take into account complex interactions within the team, and, given the high popularity of the sport, there is also a large amount of readily available data. In recent years, Jeff Sackmann has released the largest library of tennis datasets on GitHub (`https://github.com/JeffSackmann`). Modeling tennis matches has therefore become an extremely popular subject in scientific papers over the last few years.

The scientific literature is still mostly focused on pre-match models for predicting the outcome of tennis matches. There are several different approaches to pre-match modeling. The first attempts to model tennis matches used player rankings (Boulier and Stekler, 1999; Clarke and Dyte, 2000; Klaassen and Magnus, 2003;

Radicchi, 2011). In addition to tennis, similar research has been conducted in other sports, some of which can be applied to tennis (Boulier and Stekler, 2003; Lebovic and Sigelman, 2001; Dangauthier *et al*., 2007; Glickman, 2001). In addition to the official ranking systems, in pre-match modeling, Elo ranking models also stand out. Although Elo was originally developed as a chess player rating system (Elo, 1978), it can also be used as a match prediction tool in many other sports (Hvattum and Arntzen, 2010; Leitner *et al*., 2010; Ryall and Bedford, 2010; Carbone *et al*., 2016). When talking about the pre-match group of models, it is important to mention both hierarchical models and Bradley–Terry models (Bradley and Terry, 1952; McHale and Morton, 2011; Glickman, 1999; Baker and McHale, 2014; 2017). The evaluation of 11 different pre-match models was done by Kovalchik (2016). The paper proved that the FiveThirtyEight.com Elo rating method (Silver and Fischer-Baum, 2015; Morris *et al.*, 2016) outperforms other approaches and shows the best performance. Kovalchik and Reid (2019) extended this method to live betting.

As mentioned, hierarchical Markov models are very popular in the modeling of tennis matches. From the perspective of any player, tennis is a game that involves a lot of repetition. A player is constantly exposed to a situation where they have to score a point under roughly similar rules and conditions. A match is divided into sets, which are further broken down into games. Each game is won after earning enough points. By winning a certain number of games a player wins a set, and finally after winning a specified number of sets a player wins the match. It is because of this nature of the sport that a tennis match can be easily described through hierarchical Markov models. Much of the literature on modeling tennis matches uses this approach. Schutz (1970) describes a tennis match through a Markov chain with constant probabilities of transition between states. The states in such a model represent the result in a game/set/match (depending on what level of the hierarchy it is done at), while the transitions between states are constant probabilities of winning points/games at one's own service or set. Pollard (1983) presented an analytical approach to calculate the probability of winning a game or set and the expected number of points/games to be played in a match. Liu (2001), Newton and Keller (2005) or O'Malley (2008) give equivalent hierarchical expressions to estimate the probability of winning games, sets and matches only based on the probability of winning points at their own service, using different approaches. These formulae can be used to estimate the likelihood of the outcome of a particular match level before the match starts (pre-match). Croucher (1986) studied the conditional probability of winning a game from any score (live). Barnett and Clarke (2002) as well as Barnett *et al*. (2006) present recursive formulae to predict the winner and duration of each level of a tennis match live. Due to the computational complexity of the recursive expressions presented, Wozniak (2011) (based on the pre-match solution proposed by O'Malley (2008)) offers an analytical solution for the calculation of the same probabilities. If one is to predict the likelihood of a particular score in a tennis match, recursive expressions can be found in the literature (Barnett and Brown, 2012). Our paper completes the research area that deals with the construction of predictive models based on the identical and independent point distribution assumption, which is as of yet left unexplored. We offer expressions that can also be used to estimate the likelihood of a particular score at any point in the match. These expressions are much faster than recursive ones. The importance of the terms is also compounded by the fact that the same terms can also be used to calculate the probability of winning at any given moment in the match with the same precision as the approaches offered in the literature, but at a faster rate.

Given that the likelihood of winning points at one's service has proven to be a major and crucial factor in trying to estimate the likelihood of winning a game, set or match, several scientific articles have offered different approaches to pinpoint that parameter (Barnett and Clarke, 2005; Newton and Aslam, 2009; Spanias and Knottenbelt, 2013; Knottenbelt *et al*., 2012). The probabilities of winning points at one's service have proven very important throughout history to model sports such as squash (Renick, 1976) or racquetball (Keller, 1984). Our approach also uses the probabilities of winning points on one's own service as input parameters.

Due to the Markov property, by which the next state depends only on the current state, the models presented previously are based on the assumption of identical and independent point distribution. The term 'independent' means that the probability of winning points at one's service does not depend on the outcome of the point previously played. On the other hand, 'identical distribution' means that every point is considered equal, regardless of whether it is a very important point (e.g., break ball or match ball) or a less important point during the match. Although this assumption greatly facilitates the modeling of tennis matches, it conflicts with intuition about psychological momentum and pressure, known in psychology as the effect of psychological momentum (Iso-Ahola and Mobily, 1980). The existence of psychological momentum is a topic of research in many other sports, such as basketball or baseball, and in everyday speech it is almost impossible to mishear expressions such as "hot teams" or "hot hands" (Gilovich *et al*., 1985; Green and Zwiebel, 2017; Tversky and Gilovich, 1989; Albright, 1993; Attali, 2013; Dadelo *et al*., 2014; Ballı and Korukoğlu, 2014).

The reassessment of the assumption of an independent and identical point distribution in tennis

was started by Klaassen and Magnus (2001). They showed that players are more likely to score points at their service if they have previously won a point on their service (psychological momentum), while the likelihood is lower when serving a very significant point in the match (psychological pressure). This means that the outcomes of a player's several previous services need to be taken into account to more accurately predict whether and with what likelihood the player will win the next point on his service. So, like in many other practical applications, when dependence between trials is present, it is reasonable to assume that the probability of success on the current trial depends on the outcome of the number of last trials. Similar research was conducted by Martin (2006) who presented a recursive method of computing the distribution of the number of successes in a sequence of binary trials that are Markovian of general order is given. Despite the above, the same scientists have proven that these facts are rather weak and that the assumption of identical and independent point distribution is good enough when predicting the outcome of tennis matches (Klaassen and Magnus, 2001). For this reason, we also use the assumption of identical and independent point distribution in this paper. A lot of scientific papers still try to tackle this area of research (Šarčević *et al.*, 2021; Percy, 2015; Carrari *et al.*, 2017; Chang, 2019; Wetzels *et al.*, 2016; Dietl and Nesseler, 2017).

3. Tennis scoring rules

As stated, in a tennis match, scoring is performed through points, games, and sets. By winning enough points the player wins the game. By winning enough games the player wins the set, and ultimately by winning enough sets the player wins the match.

Points in the game are given as 0, 15, 30, 40, and *game*/*Ad*[1], although functionally they are similar to a simpler system using just numbers 0 to 4. The player who first wins 4 points wins the game. The exception is when both players score 3 points each (the *deuce* score, or $40-40$). Then the game continues until one of the players achieves a two-point difference. When the server wins the deuce point, it is called *Ad-In*, but when he loses the deuce point, it is called *Ad-Out*. If the player with the advantage (*Ad-In* or *Ad-Out*) wins another point, he wins the game, or it goes back to deuce. The same player is serving throughout the game.

Depending on the rules of the tournament, there are two types of matches. A match can consist of the best 2 out of 3 sets or the best 3 out of 5 sets. The former case applies to all women tournaments, while the latter is played in certain tournaments for men. Each set consists of at least 6 games and the winner is the player who wins the 6 games first. However, if both players win 5 games each, the winner is the one who wins two games in a row. The score can reach $6-6$. Then the tiebreak game is played. If it is the last set in the match, depending on the rules of the tournament, instead of playing the tiebreak game, the set may continue until one of the players achieves a two-game difference (the so-called advantage set). The Wimbledon and Australian Open tournaments depart from these rules. Specifically, in 2019, the rules of the Wimbledon Tournament changed. If it is not possible to determine the winner of the set by regular scoring, the last set will continue until one of the players wins with a two-game difference, however, only up to a score of $12-12$. Only after the result $12-12$, is the tiebreak game played. In Australia, a super-tiebreak will be played if the deciding set reaches the result $6-6$. In the super-tiebreak, players need to win ten points by a margin of 2. Players are serving alternately in the set.

The tiebreak game is played until one of the players scores 7 points. Similar to the set scoring rules, if both players score 6 points, the game continues until one of the players achieves a two-point difference. The player who started serving in the set also starts serving in the tiebreak game. After this service, the players alternate so that each player serves two points.

4. Combinatorial model

Tennis is an example of a hierarchical sport since the match consists of a sequence of sets, a set of a sequence of games, and a game of a sequence of points. For this reason, each level of a tennis match is modeled separately. First, mathematical expressions for modeling a standard and a tiebreak game are given. Subsequently, similar terms are given for the tiebreaker and advantage set. The proposed mathematical formulae are based on a binomial distribution. To predict the final score at any moment in the match, it is necessary to combine these formulae.

The following is a description of the main notation used in the formulae presented in the rest of the paper (the rest of the notation will be introduced gradually):

- p represents the probability of winning a point on serve for player A. The definition of player A depends on the level of the match being modeled. In the case of standard game modeling, player A is the player who serves in that game. In the case of tiebreak game or set modeling, player A is the player who starts serving in the tiebreak game or set;
- q represents the probability of winning a point on serve for player B;
- x, y represent the current score in the standard game/tiebreak game/set (depending on what level of match is being modeled);

[1]The *game* tag indicates that one of the players has won the game. The *Ad* tag indicates that one of the players has the advantage.

- a, b represent the final score in the standard game/tiebreak game/set (depending on what level of match is being modeled);
- a_s, b_s represent the number of services remaining from the score (x, y) to the score (a, b) on player's (A/B) serve.

Note. All the probabilities calculated below represent the absorption probabilities of a Markov chain. An explicit computation of the final probabilities starting from any current score can be achieved using a simple Markovian approach. For more details, see the work of Barnett and Brown (2012).

4.1. Game-level model.

4.1.1. Standard game. To calculate the likelihood that player A will win a game by a certain score, consider

$$p_g(p,x,y,a,b) = \begin{cases} \binom{a_s-1}{b-y} p^{a_s-(b-y)}(1-p)^{b-y}, \\ a=4, b\le 2, x\le 3, y\le 2, & \text{(1a)} \\ \binom{a_s}{b-y} p^{a_s-(b-y)}(1-p)^{b-y}\dfrac{p^2}{p^2+(1-p)^2}, \\ a=3, b=3, x\le 3, y\le 3. & \text{(1b)} \end{cases}$$

Input values (x, y, a, b) are entered as 0(0), 1(15), 2(30), 3(40), 4(*game*). The variable a_s represents the number of services remaining from the score (x,y) to the score (a,b) and is calculated as $a_s = (a-x)+(b-y)$.

Because of the tennis scoring rules there is a need for two formulae, (1a) and (1b). Formula (1a) is used to calculate the probability of winning the game with the final score $game-0$, $game-15$ or $game-30$. Formula (1b) is required to calculate the probability of winning a game with the final score $game-40$. The product of the first three factors in formula 1b ($\binom{a_s}{b-y}p^{a_s-(b-y)}(1-p)^{b-y}$) represents the probability of reaching the score $40-40$, and the term $p^2/(p^2+(1-p)^2)$ represents the probability of winning the game after the score $40-40$ (see Appendix A for more details).

It is important to note that even though the notation of the combinatorial model and recursive model differs, these are equivalent mathematical models that give identical results in terms of accuracy (see Appendix B for more details). Both approaches are based on counting all the paths that lead to the final score and calculating the probabilities on those paths.

Note. The notation $game-30$ means that player A won the game and player B scored 30 points. The same applies to other results.

Example 1. (*Winning the game with a particular final score*) For easier understanding of the logic behind (1a) and (1b), an example is given. Assuming the current score in a game is $0-15$, the probability that the game will end with the score $game-30$ is calculated as follows: from the current score $0-15$ to the final score $game-30$ the player will serve 5 more times. To reach that particular score, one of those services must be lost (probability $1-p$), while the remaining 4 services must be won (probability p^4), with the restriction that the lost service cannot be the last one. So from the remaining 4 services, 1 is selected which the player can lose, which can be done in $\binom{4}{1}$ ways. Ultimately, the following expression is obtained: $4\times(1-p)\times p^4$. This expression is also obtained by applying (1a). ♦

Example 2. (*Winning the game*) Formulae (1a) and (1b) can also be used to calculate the probability that a player serving in a game will win that game by simply summing up all the possible outcomes from the current score. For example, $game-15$, $game-30$, and $game-40$ are possible outcomes from the score $0-15$. By separately calculating the likelihood of each score using (1a) and (1b) and subsequently summing the results, it is easy to ultimately obtain the probability of winning the game from any score. In the rest of this paper, the probabilities of winning the game from the score $0-0$ will be referred to as $G(p)$ and $G(q)$, and the probability of player A winning the game from the current score $ptsA-ptsB$ will be referred to as $G(p,ptsA,ptsB)$. The function arguments p and q denote the probabilities of winning a point on players' own serve. ♦

The same idea was used in modeling other levels of a tennis match, and the derived formulae will be given in the following paragraphs. Also, similar formulae can be written to calculate the probability of losing different levels of a tennis match by a particular score.

4.1.2. Tiebreak game. The formula for modeling the tiebreak game is slightly more complicated compared with that for modeling the standard game because of the tiebreak serving rules—the player who started serving in the set also starts serving in the tiebreak game, and, after that service, the players alternate so that each serves two points. If one wants to calculate the likelihood that a tiebreak game will end with a certain score, it is necessary to go through all the paths which lead to that score while taking into account these serving rules, therefore the summation appears in (2a), (2b) and (2c) below. The parameter k is a simple iterator used to "walk" through all possible sets of paths leading to the final score and indicates how many points player A/B has lost on their own serves on that set of paths. The number of paths in a set of paths is determined by the binomial coefficient in

formulae (2a), (2b) and (2c) (similar as in formulae (1a) and (1b).

A more detailed explanation is given by the example below.

Example 3. (*Explanation of the parameter k and the binomial coefficient*) Suppose a tiebreak game is played and the current score is $3-1$, and one wants to calculate the likelihood that the tiebreak game will end with the score $7-3$. Four points have already been played, and six more will be played to achieve the score $7-3$. Considering the tiebreak serving rules, and assuming that player A has started serving in the tiebreak game it is easy to see that, of those remaining 6 points, player A will serve the first, the fourth, and the fifth point, and player B will serve the second, the third and the sixth point. There are 3 sets of paths to get the score $7-3$ from the score $3-1$:

- Player A (player who began to serve in a tiebreak game) can score 3 points on their own serves (the first, the fourth, and the fifth point of the remaining 6 points) and 1 point on player B's serve (the last point required, i.e., the sixth point of the remaining 6 points). In this set of paths, the value of the iterator k will be 0 because player A won each point on their own serves. The binomial coefficient will be 1 because this set of paths consists of only 1 path—there is only one way to reach the final score.
- Player A can score 2 points on their own serves (the first and the fourth point, or the first and the fifth point, or the fourth and the fifth point of the remaining 6 points) and 2 points on player B's serves (the last point required, i.e., the second and the sixth point, or the third and the sixth point of the remaining 6 points). In this set of paths, the value of the iterator k will be 1 because player A lost one point on their own serve. The binomial coefficient will be 6 because this set of paths consists of 6 paths—there is 6 ways to reach the final score.
- Player A can score 1 point on their own serve (the first, or the fourth, or the fifth point of the remaining six points) and 3 points on player B's serves (the second, the third, and the sixth point from the remaining six points). In this set of paths, the value of the iterator k will be 2 because player A lost 2 points on their own serves. The binomial coefficient will be 3 because this set of paths consists of 3 paths—there are three ways to reach the final score.

♦

Depending on the likelihood of winning a point on own services for both players (p and q) and the current tiebreak score (x and y), (2a), (2b) and (2c) have been given to calculate the likelihood of player A winning the tiebreak game with a particular score (a and b).

Remark 1. the summation sign in formulae (2a), (2b), (2c), (3a), (3b), (4) and (5) can be interpreted as follows: the iterator k varies from 0 to n ($n \in \mathbb{N}$) provided that $x'(k)$, $y'(k)$, $z'(k)$ and $w'(k)$ are greater than or equal to zero:

$$
\begin{aligned}
&p_s\,(p,q,x,y,a,b)\\
&\quad=\begin{cases}
\displaystyle\sum_{\substack{k=0\\ x'(k),y'(k)\geq 0,\\ z'(k),w'(k)\geq 0}}^{n}\binom{a_{\mathrm{s}}-1}{x'(k)-1}\binom{b_{\mathrm{s}}}{w'(k)}p^{x'(k)}\\
\quad\times(1-p)^{y'(k)}q^{z'(k)}(1-q)^{w'(k)},\\
\quad a=7, b\leq 5, (a+b)\bmod 4\ \in\ [0,1],\\
\quad x<7, y\leq 5 & \text{(2a)}\\[2ex]
\displaystyle\sum_{\substack{k=0\\ x'(k),y'(k)\geq 0,\\ z'(k),w'(k)\geq 0}}^{n}\binom{a_{\mathrm{s}}}{x'(k)}\binom{b_{\mathrm{s}}-1}{w'(k)-1}p^{x'(k)}\\
\quad\times(1-p)^{y'(k)}q^{z'(k)}(1-q)^{w'(k)},\\
\quad a=7, b\leq 5, (a+b)\bmod 4\ \in\ [2,3],\\
\quad x<7, y\leq 5, & \text{(2b)}\\[2ex]
\displaystyle\sum_{\substack{k=0\\ x'(k),y'(k)\geq 0,\\ z'(k),w'(k)\geq 0}}^{n}\binom{a_{\mathrm{s}}}{x'(k)}\binom{b_{\mathrm{s}}}{w'(k)}p^{x'(k)}\\
\quad\times(1-p)^{y'(k)}q^{z'(k)}(1-q)^{w'(k)}\\
\quad\times\dfrac{p(1-q)}{p(1-q)+(1-p)q},\\
\quad a=6, b=6, x\leq 6, y\leq 6, & \text{(2c)}
\end{cases}
\end{aligned}
$$

The functions $x'(k)$, $y'(k)$, $z'(k)$ and $w'(k)$ are calculated as follows:

- in (2a) and (2c):

$$
\begin{aligned}
x'(k) &= \min(a_s, a-x)-k,\\
y'(k) &= b-y-z'(k),\\
z'(k) &= \min(b_s, b-y)-k,\\
w'(k) &= a-x-x'(k),
\end{aligned}
$$

- in (2b):

$$
\begin{aligned}
x'(k) &= \begin{cases} a_s-k, & a_s<(a-x),\\ a-x-k-1, & \text{otherwise,}\end{cases}\\
y'(k) &= b-y-z'(k),\\
z'(k) &= \begin{cases} b-y-k, & (b-y)<b_s, a_s<(a-x),\\ b_s-k-1, & \text{otherwise,}\end{cases}\\
w'(k) &= a-x-x'(k).
\end{aligned}
$$

Each term in the summation represents a set of possible paths from the current to the final score, and the functions $x'(k)$, $y'(k)$, $z'(k)$ and $w'(k)$ determine how many points player A/ B must win/ lose on that set of paths in order for the tiebreak game to end with the given final score (from a given current score). In more details:

- $x'(k)$ represents the number of points that player A must score on their own serves in order to reach the given final score;
- $y'(k)$ represents the number of points that player A must lose on their own serves in order to reach the given final score;
- $z'(k)$ represents the number of points that player B must score on their own serves in order to reach the given final score;
- $w'(k)$ represents the number of points that player B must lose on their own serves in order to reach the given final score.

The derivation of the formulae $x'(k)$, $y'(k)$, $z'(k)$ and $w'(k)$ in the case of using the Eqns. (2a) and (2c) is explained below. The same principle can be used to derive the formulae in the case of using Eqn. (2b).

For player A to win the tiebreak game with the score $a-b$ from the score $x-y$, player A needs to win $a-x$ points and will serve a_s times. If player A needs fewer points to reach the final score $(a-x)$ than the number of remaining services (a_s), player A cannot win all points on their own serves. The maximum number of points that player A can win on their own serves in this case is $a-x$. Otherwise, player A can win every point on their own serves (a_s). Therefore, in calculating the value of $x'(k)$ it is necessary to take the minimum between the variables a_s and $a-x$. The iterator k is used to go through other possible sets of paths leading to the final score from the given current score. We start with $k=0$. If player A has not achieved the required number of points by winning points on their own serves, player A can acquire those points if the opposing player loses points on their own serves. The number of points that player B must lose on their own serves in order for player A to achieve the predetermined final number of points is determined by calculating the function $w'(k)$ as the difference between the required points $(a-x)$ and the points player A has won/will win on their own serves $(x'(k))$. The formulae for calculating $y'(k)$ and $z'(k)$ are derived in a similar way.

Again, because of the rules in the tiebreak game, the formula for calculating the likelihood of a particular score in a tiebreak game needs to be written in three terms: (2a), (2b) and (2c). Formulae (2a) and (2b) can be used to calculate the likelihood of winning a tiebreak game with scores $7-0, 7-1, 7-2, 7-3, 7-4$, and $7-5$, depending on which player serves the last point in the tiebreak game. The case when the tiebreak game does not end with one of the above scores, but the tiebreak game continues until player A wins the tiebreak game with a 2 point difference, is covered by (2c). Similarly to (1b), the first 6 product factors in (2c) represent the probability of reaching the score $6-6$, and the term $p(1-q)/(p(1-q)+(1-p)q)$ is the probability of player A winning the tiebreak game after the score $6-6$ (see Appendix (A) for more details, although Appendix A describes the derivative of formula (1b), a similar approach is used to derive formula (2c)).

4.2. Set-level model. The formulae for modeling the set are similar to those for modeling the tiebreak game. However, when modeling a set, there are two possible scoring rules to consider. If the set does not finish with the scores $6-0$, $6-1$, $6-2$, $6-3$ or $6-4$, the set may continue until one of the players wins the set by two games difference (the so-called advantage set) or the so-called tiebreak game is played to decide the winner of the set. Accordingly, the set modeling formula consists of a combination of three formulae: (3a), (3b) and (4) in the case of playing an advantage set, or (3a), (3b) and (5) when a tiebreak game is played to decide the winner of the set. The summation and the iterator k also appear in formulae (3a), (3b), (4) and (5). The reason is again to pass all possible paths from the current to the final score:

$$p_s\,(p,q,x,y,\ \mathrm{a},b) = \begin{cases} \displaystyle\sum_{\substack{k=0\\ x'(k),y'(k)\geq 0,\\ z'(k),w'(k)\geq 0}}^{n} \binom{a_s-1}{x'(k)-1}\binom{b_s}{w'(k)} \\ \quad \times G(p)^{x'(k)}(1-G(p))^{y'} \\ \quad \times G(q)^{z'(k)}(1-G(q))^{w'(k)}, \\ \qquad a=6, b\leq 4, (a+b)\bmod 2\neq 0, \\ \qquad x<6, y\leq 4, \qquad (3a) \\ \displaystyle\sum_{\substack{k=0\\ x'(k),y'(k)\geq 0,\\ z'(k),w'(k)\geq 0}}^{n} \binom{a_s}{x'(k)}\binom{b_s-1}{w'(k)-1} \\ \quad \times G(p)^{x'(k)}(1-G(p))^{y'(k)} \\ \quad \times G(q)^{z'(k)}(1-G(q))^{w'(k)}, \\ \qquad a=6, b\leq 4, (a+b)\bmod 2=0, x<6, \\ \qquad y\leq 4. \qquad (3b) \end{cases}$$

If we change the conditional expressions in (2a) and (2b), and replace the probabilities of winning the point on service (p and q) with the probabilities of winning a game on service ($G(p)$ and $G(q)$—the calculation of these values is explained in Example 2), we get (3a) and (3b). Values $x'(k)$, $y'(k)$, $z'(k)$ and $w'(k)$ in (3a) are calculated

according to the same rules as in (2a). In (3b) values $x'(k)$, $y'(k)$, $z'(k)$ and $w'(k)$ are calculated according the same rules as in (2b). Formulae (3a) and (3b) are used to calculate the probability of winning a set with scores $6-0$, $6-1$, $6-2$, $6-3$, $6-4$.

4.3. Advantage set. If the advantage set is played, in addition to (3a) and (3b), (4) is used to calculate the probability of winning the set—an analogy with (2c) can be observed:

$$\begin{aligned} &p_{\mathrm{sA}}\,(p,q,x,y,a,b)\\ &\quad=\sum_{\substack{k=0\\ x'(k),y'(k)\geq 0,\\ z'(k),w'(k)\geq 0}}^{n}\binom{a_s}{x'(k)}\binom{b_s}{w'(k)}G(p)^{x'(k)}\\ &\quad\times(1-G(p))^{y'(k)}G(q)^{z'(k)}(1-G(q))^{w'(k)}\\ &\quad\times\frac{G(p)(1-G(q))}{G(p)(1-G(q))+(1-G(p))G(q)},\\ &\qquad a=5,b=5,x\leq 5,y\leq 5. \end{aligned} \tag{4}$$

The fraction in formula (4) represents the probability that player A will win the set if the score is $5-5$. The rest of formula (4) represents the probability of reaching the score $5-5$ from any score.

4.4. Tiebreak set. To complete the set modeling formula, it is still necessary to cover the case of the tiebreak set with

$$\begin{aligned} &p_{\mathrm{sT}}\,(p,q,x,y,a,b)\\ &\quad=\sum_{\substack{k=0\\ x'(k),y'(k)\geq 0,\\ z'(k),w'(k)\geq 0}}^{n}\binom{a_{\mathrm{s}}}{x'(k)}\binom{b_{\mathrm{s}}}{w'(k)}G(p)^{x'(k)}\\ &\quad\times(1-G(p))^{y'(k)}G(q)^{z'(k)}\\ &\quad\times(1-G(q))^{w'(k)}(G(p)G(q)\\ &\quad+(1-G(p))(1-G(q)))p_t,\\ &\qquad a=5,b=5,x\leq 5,y\leq 5. \end{aligned} \tag{5}$$

The first six product factors in (5) represent the probability of reaching the score $5-5$ from any given score. The term $G(p)G(q)+(1-G(p))(1-G(q))$ in (5) represents the probability of reaching the score $6-6$ from the score $5-5$. Finally, p_t in (5) represents the probability of player A winning the tiebreak game (2a), (2b) and (2c).

By combining formulae for winning a game (1a), (1b), (2a), (2b) and (2c) and winning a set (3a), (3b), (4) and (5), it is possible to take into account the current score in a game when calculating the probability of a particular outcome of a set. This can, at any point in the match, calculate the probability of a particular outcome in the set.

4.5. Winning the match. Match modeling formulae vary depending on the tournament being played (see Section 3). This section gives an example for calculating the probability of winning a best-of-3 all tiebreak set match. First, we repeat/introduce some additional notation:

- $G(p,ptsA,ptsB)$ represents the probability of player A (player who serves in the game) winning the game from the current score $ptsA-ptsB$ (see Example 2).
- $S(p,q,gA,gB)$ represents the probability of player A (player who started serving in the set) winning the tiebreak set from the current score $gA-gB$. It is calculated similarly as the probability of winning the game from a given score (see Example 2).

Assume the current score is $ptsA*-ptsB\ gA-gB\ sA-sB$ = $15-0\ 3-1\ 1-0$, where $ptsA*-ptsB$ denotes the score in the currently played game (the star sign denotes the player on serve), $gA-gB$ denotes the score in the set being played, and $sA-sB$ denotes the number of sets won by both players. There are four possible ways for player A to win the match:

- Player A can win the currently played game with the probability $G(p,15,0)$, and then win the currently played set with the probability $S(p,q,4,1)$. The total probability in this case is $G(p,15,0)\times S(p,q,4,1)$.
- Player A can lose the currently played game with the probability $1-G(p,15,0)$, and then win the currently played set with the probability $S(p,q,3,2)$. The total probability in this case is $(1-G(p,15,0))\times S(p,q,3,2)$.
- Player A can win the currently played game with the probability $G(p,15,0)$ and then lose the currently played set with the probability $1-S(p,q,4,1)$, and after that player A can win the next set with the probability $S(p,q,0,0)$. The total probability in this case is $G(p,15,0)\times(1-S(p,q,4,1))\times S(p,q,0,0)$.
- Player A can lose the currently played game with the probability $1-G(p,15,0)$ and then lose the currently played set with the probability $1-S(p,q,3,2)$, and after that player A can win the next set with the probability $S(p,q,0,0)$. The total probability in this case is $(1-G(p,15,0))\times(1-S(p,q,3,2))\times S(p,q,0,0)$.

The total probability that player A will win the match can be obtained by summing the probabilities of all four cases. Note that in the third and the fourth case we used $S(p,q,0,0)$ to denote the probability of player A winning the set from the score $0-0$. It is unknown which player

will start serving in the next set. However, the outcome of the set is not influenced by the choice of the first serving player. That is why we assumed player A will start serving in the next set.

5. Evaluation

In this section, we demonstrate the execution times of both approaches and prove the existence of a relative difference between them. The combinatorial approach is first compared with forward recursion to estimate the final score of the match. Then, it is compared with backward recursion to predict the winner of the match.

Formulae for predicting the outcome of tennis matches were implemented in the programming language C++, using the Visual Studio Code development environment. The computer features for conducting the experiment are

- operative system: *Windows 10*,
- processor: *Intel(R) Core(TM) i7-6700HQ CPU @ 2.60 GHz 2.59 GHz*,
- RAM: *16 GB*,
- disk: *256 GB SSD*.

It is important to note that each execution time presented in this section represents the time required for 100000 random predictions. The execution time measurements were performed 100 times and the paper presents the minimum and maximum execution time, median, and arithmetic mean. Times are expressed in seconds.

Formulae for each level of a tennis match (game, set) were tested separately. First, times required to predict the final score and the winner of the game were compared. Then we analyzed the same at the set level.

Except measuring the execution times of both approaches, it is easy to see that the complexity of recursive approaches is exponential regardless of the match level being modeled (see the formulae in the work of Barnett and Brown (2012)). The complexity of the combinatorial approach is linear at the game level and quadratic at the set level (see formulae (1a), (1b), (2a), (2b), (2c), (3a), (3b), (4) and (5).

5.1. Predicting the final score of the match.

5.1.1. Game level. The execution time presented in Table 1 represents the time required to predict 100000 different probabilities that a standard game will end with the score $game - 30$ from randomly selected scores (execution time test of (1a)), and the execution time required to predict 100000 different probabilities that the standard game will finish after deuce from randomly selected scores is shown in Table 2 (execution time test of (1b)).

Table 1. Execution time (in [s]) required to predict 100000 different probabilities that the game will end with the score $game - 30$.

Approach	Min	Max	Median	Mean
Recursion	0.02094	0.03991	0.02293	0.02293
Combinatorics	0.00896	0.03889	0.00997	0.01058

Table 2. Execution time (in [s]) required to predict 100000 different probabilities that the game will end with the score $game - 40$.

Approach	Min	Max	Median	Mean
Recursion	0.03354	0.07979	0.03890	0.03953
Combinatorics	0.01496	0.03590	0.01695	0.01733

Table 3. Execution time (in [s]) required to predict 100000 different probabilities that the tiebreak game will end with the score $7 - 4$.

Approach	Min	Max	Median	Mean
Recursion	1.243	1.396	1.267	1.281
Combinatorics	0.02893	0.05783	0.03193	0.03266

The combinatorial approach is on the average two times faster than the recursive one in both cases.

Tables 3 and 4 show the execution time required to predict the final score in a tiebreak game from 100000 randomly selected current scores. Table 3 shows the execution time required to predict 100000 different probabilities that a tiebreak game will end with the score $7-4$ from randomly generated scores, while the execution time required to predict 100000 different probabilities that a tiebreak game will end with the score $8 - 6$ from randomly generated scores is shown in Table 4.

A much shorter execution time compared with the recursive approach is now noticeable. The combinatorial approach gives on the average almost 39 times shorter execution time than the recursive one when it comes to predicting the final score in a tiebreak game. An even bigger time difference is evident when the tiebreak is predicted not to end within 14 points played. Then the combinatorial approach is on the average almost 163 times faster than the recursive one. Greater acceleration of the combinatorial model in comparison with the recursive model in the tiebreak game prediction was expected because the tiebreak game formula must cover more cases than the standard game formula.

5.1.2. Set level. The execution times required to predict the final score in a set are shown in Tables 5, 6 and 7. The execution time required to predict 100000 different probabilities that a set will end with the score $6 - 2$ from randomly generated scores is shown in Table

5, which presents the execution time measured using the tiebreak set recursion formula (*Recursion T*), the advantage set recursion formula (*Recursion A*), as well as the combinatorics formula (*Combinatorics*).

When modeling a set, recursive expressions can be found in the literature that, instead of the probabilities of winning points at their own service (p and q), receive the probabilities of winning games at their own service ($G(p)$ and $G(q)$) as input parameters. The execution times required to generate the probability of reaching a particular score in a set with such modification are also shown in Table 5 (*Recursion' T*, *Recursion' A*). With such a modification, the prediction times using the modified recursive approaches become faster than the execution times obtained using the combinatorial technique. However, if the proposed combinatorial formula is adjusted in the same way (*Combinatorics'*), a faster execution time is obtained when compared to the modified recursive approaches. The modified combinatorial technique is on average almost 1.5 times faster than the modified recursive approach when using the tiebreak set recursion formula, and almost on average 5 times faster than the modified recursive technique when using the advantage set recursion formula.

Table 6 presents the time in which the set ends in a tiebreak game from 100000 randomly selected current scores. In addition to pre-calculating the probabilities of winning a service game for both players, the recursive function is further modified and uses a significantly faster backward recursion to calculate the likelihood of winning the tiebreak game (needed to generate the final probability). Despite the shortening of the prediction time relative to the initially defined recursive approach, the modified recursive technique (*Recursion'*) still did not reach the time of the initial combinatorial approach. If the combinatorial approach is modified as previously described, the time of the proposed combinatorial technique is further reduced (*Combinatorics'*). Even with the acceleration of the recursive technique. It is still on the average 3 times slower than the combinatorial approach without the proposed modification.

Finally, the time required to generate 100000 different probabilities that the advantage set will end with the score $8-6$, from randomly selected current scores, is given in Table 7. The combinatorial and recursive approach modifications were performed as described above, and the combinatorial technique without and with the modification is faster than both recursive approaches.

5.2. Predicting the winner of the match. In the previous subsection was proved that the proposed combinatorial approach is significantly faster than forward recursion if one wants to predict the final score of each level of a tennis match. Below, the proposed combinatorial approach is compared with backward recursion to predict the winner of each level of a tennis match.

Table 4. Execution time (in [s]) required to predict 100000 different probabilities that the tiebreak game will end with the score $8-6$.

Approach	Min	Max	Median	Mean
Recursion	5.894	6.666	6.021	6.045
Combinatorics	0.03291	0.05887	0.03686	0.03698

Table 5. Execution time (in [s]) required to predict 100000 different probabilities that the set will end with the score $6-2$.

Approach	Min	Max	Median	Mean
Recursion T	2.482	2.657	2.535	2.544
Recursion' T	0.02394	0.03493	0.02694	0.02721
Recursion A	7.767	8.449	7.853	7.899
Recursion' A	0.09972	0.14860	0.10372	0.10503
Combinatorics	0.2005	0.2475	0.2045	0.2075
Combinatorics'	0.01895	0.04388	0.02194	0.02206

Table 6. Execution time (in [s]) required to predict 100000 different probabilities that the set will end with the score $7-6$.

Approach	Min	Max	Median	Mean
Recursion	38.77	39.97	39.26	39.28
Recursion'	2.490	2.902	2.520	2.535
Combinatorics	0.7460	0.8228	0.7631	0.7704
Combinatorics'	0.2947	0.3487	0.3141	0.3161

Table 7. Execution time (in [s]) required to predict 100000 different probabilities that the set will end with the score $8-6$.

Approach	Min	Max	Median	Mean
Recursion	86.97	88.94	87.72	87.68
Recursion'	1.234	1.414	1.257	1.260
Combinatorics	0.4695	0.5349	0.4855	0.4871
Combinatorics'	0.01667	0.04357	0.02995	0.03057

The execution time required to predict the winner of a standard game from 100000 randomly generated current scores is shown in Table 8. The combinatorial approach in the case of predicting the outcome of the standard game is almost on the average 2 times faster when compared with the backward recursion one.

The execution time in the case of predicting the winner of a tiebreak game from 100000 randomly selected current scores is shown in Table 9. When predicting the outcome of a tiebreak game, the combinatorial approach is almost on average 1.5 times faster than the backward recursive one.

If the recursive or the combinatorial approach is used to predict the winner of the tiebreak or the advantage set from 100000 randomly generated current scores, this will require times shown in Tables 10 and 11. The combinatorial approach is on average 93 times faster

Table 8. Execution time (in [s]) required to predict the winner of a standard game 100000 times.

Approach	Min	Max	Median	Mean
Recursion	0.03787	0.05286	0.03991	0.04043
Combinatorics	0.02197	0.03989	0.02396	0.02440

Table 9. Execution time (in [s]) required to predict the winner of a tiebreak game 100000 times.

Approach	Min	Max	Median	Mean
Recursion	0.1661	0.2034	0.1815	0.1808
Combinatorics	0.1036	0.1598	0.1194	0.1198

Table 10. Execution time (in [s]) required to predict the winner of a tiebreak set 100000 times.

Approach	Min	Max	Median	Mean
Recursion	154.1	164.2	157.6	157.9
Combinatorics	1.660	1.843	1.686	1.699

Table 11. Execution time (in [s]) required to predict the winner of a advantage set 100000 times.

Approach	Min	Max	Median	Mean
Recursion	39.32	41.95	40.06	40.17
Combinatorics	1.070	1.254	1.088	1.095

than the recursive one when predicting the winner of the tiebreak set.

In the case of predicting the winner of the advantage set, the combinatorial approach is on average 37 times faster than the recursive one.

In conclusion, if one wants to predict the winner of each level of a live tennis match or the score by which each level will end, it is not necessary to use two different recursive approaches; backward and forward recursion. With one combinatorial approach, it is possible to predict all of the above with the same accuracy and significantly less execution time.

6. Conclusion

Building predictive models that try to forecast the outcome of sporting events have become extremely popular in recent years. The nature of the chosen sport plays an important role in the selection of sports event modeling methods. Tennis is an example of a sport with a strongly defined structure and a rigid scoring system, making it relatively easy to model its matches in the form of discrete stochastic processes, such as Markovian. In the scientific literature, several approaches to modeling tennis matches are presented, with hierarchical recursive Markov models being amongst the more popular ones. Using the combination of two recursive approaches (backward and forward recursion) it is possible to determine the winner of the match and to predict the final score in the match. This paper has shown that, by leveraging the provided combinatorial model to estimate these probabilities, identical accuracy can be gained, but with a significant improvement in execution time, making it a superior alternative in environments that require real-time adjustment of betting odds.

This paper offers an overview of the theoretical foundation for this approach in the form of mathematical formulae which are supported by concrete examples. A detailed evaluation of the proposed combinatorial model was performed. The paper shows that the combinatorial approach gains a significant advantage when it comes to execution time—in some cases time is decreased by two orders of magnitude. Furthermore, the combinatorial approach offers a noticeable decrease in general complexity, since recursive models require implementing two different recursive technique, which can both be efficiently replaced by using the proposed combinatorial formulae. The future work will be focused on building more refined models that will loosen the i.i.d. (identical and independent distribution) assumption on the point spread and integrate the concept of sport momentum.

While this paper focuses mainly on tennis, the approach can easily be transferred to similarly structured sports, such as volleyball, badminton, table tennis, etc. By leveraging the combinatorial approach, real-time betting systems can achieve better efficiency and faster reaction times, without the need for scaling up the underlying hardware and software architecture.

Acknowledgment

This research has been supported by the European Regional Development Fund under the grant KK.01.1.1.01.0009 (DATACROSS).

References

Albright, S.C. (1993). A statistical analysis of hitting streaks in baseball, *Journal of the American Statistical Association* **88**(424): 1175–1183.

Attali, Y. (2013). Perceived hotness affects behavior of basketball players and coaches, *Psychological Science* **24**(7): 1151–1156.

Baker, R.D. and McHale, I.G. (2014). A dynamic paired comparisons model: Who is the greatest tennis player?, *European Journal of Operational Research* **236**(2): 677–684.

Baker, R.D. and McHale, I.G. (2017). An empirical Bayes model for time-varying paired comparisons ratings: Who is the greatest women's tennis player?, *European Journal of Operational Research* **258**(1): 328–333.

Ballı, S. and Korukoğlu, S. (2014). Development of a fuzzy decision support framework for complex multi-attribute

decision problems: A case study for the selection of skilful basketball players, *Expert Systems* **31**(1): 56–69.

Barnett, T. and Brown, A. (2012). *The Mathematics of Tennis*, http://www.strategicgames.com.au/.

Barnett, T., Brown, A. and Clarke, S. (2006). Developing a model that reflects outcomes of tennis matches, *Proceedings of the 8th Australasian Conference on Mathematics and Computers in Sport, Coolangatta, Australia*, pp. 3–5.

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. and Clarke, S.R. (2002). Using Microsoft Excel to model a tennis match, *6th Conference on Mathematics and Computers in Sport, Queensland, Australia*, pp. 63–68.

Boulier, B.L. and Stekler, H.O. (1999). Are sports seedings good predictors? An evaluation, *International Journal of Forecasting* **15**(1): 83–91.

Boulier, B.L. and Stekler, H.O. (2003). Predicting the outcomes of national football league games, *International Journal of Forecasting* **19**(2): 257–270.

Bradley, R.A. and Terry, M.E. (1952). Rank analysis of incomplete block designs. I: The method of paired comparisons, *Biometrika* **39**(3/4): 324–345.

Carbone, J., Corke, T. and Moisiadis, F. (2016). The rugby league prediction model: Using an Elo-based approach to predict the outcome of National Rugby League (NRL) matches, *International Educational Scientific Research Journal* **2**(5): 26–30.

Carrari, A., Ferrante, M. and Fonseca, G. (2017). A new Markovian model for tennis matches, *Electronic Journal of Applied Statistical Analysis* **10**(3): 693–711.

Chang, J.C. (2019). Predictive Bayesian selection of multistep Markov chains, applied to the detection of the hot hand and other statistical dependencies in free throws, *Royal Society Open Science* **6**(3): 182174.

Clarke, S.R. and Dyte, D. (2000). Using official ratings to simulate major tennis tournaments, *International Transactions in Operational Research* **7**(6): 585–594.

Croucher, J.S. (1986). The conditional probability of winning games of tennis, *Research Quarterly for Exercise and Sport* **57**(1): 23–26.

Dadelo, S., Turskis, Z., Zavadskas, E.K. and Dadeliene, R. (2014). Multi-criteria assessment and ranking system of sport team formation based on objective-measured values of criteria set, *Expert Systems with Applications* **41**(14): 6106–6113.

Dangauthier, P., Herbrich, R., Minka, T. and Graepel, T. (2007). Trueskill through time: Revisiting the history of chess, *Advances in Neural Information Processing Systems* **20**: 337–344.

Dietl, H. and Nesseler, C. (2017). Momentum in tennis: Controlling the match, *UZH Business Working Paper Series*, University of Zurich, Zurich.

EGBA (2020). EU Market: Gambling is becoming more and more an online activity, https://www.egba.eu/eu-market/.

Elo, A.E. (1978). *The Rating of Chessplayers, Past and Present*, Arco Pub., New York.

Gilovich, T., Vallone, R. and Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences, *Cognitive Psychology* **17**(3): 295–314.

Glickman, M.E. (1999). Parameter estimation in large dynamic paired comparison experiments, *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **48**(3): 377–394.

Glickman, M.E. (2001). Dynamic paired comparison models with stochastic variances, *Journal of Applied Statistics* **28**(6): 673–689.

Green, B. and Zwiebel, J. (2017). The hot-hand fallacy: Cognitive mistakes or equilibrium adjustments? Evidence from major league baseball, *Management Science* **64**(11): 5315–5348.

Hvattum, L.M. and Arntzen, H. (2010). Using Elo ratings for match result prediction in association football, *International Journal of Forecasting* **26**(3): 460–470.

Iso-Ahola, S.E. and Mobily, K. (1980). Psychological momentum: A phenomenon and an empirical (unobtrusive) validation of its influence in a competitive sport tournament, *Psychological Reports* **46**(2): 391–401.

Keller, J.B. (1984). Probability of a shutout in racquetball, *SIAM Review* **26**(2): 267–268.

Klaassen, F.J. 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.

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

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

Kovalchik, S. and Reid, M. (2019). A calibration method with dynamic updates for within-match forecasting of wins in tennis, *International Journal of Forecasting* **35**(2): 756–766.

Lebovic, J.H. and Sigelman, L. (2001). The forecasting accuracy and determinants of football rankings, *International Journal of Forecasting* **17**(1): 105–120.

Leitner, C., Zeileis, A. and Hornik, K. (2010). Forecasting sports tournaments by ratings of (prob) abilities: A comparison for the Euro 2008, *International Journal of Forecasting* **26**(3): 471–481.

Liu, Y. (2001). Random walks in tennis, *Missouri Journal of Mathematical Sciences* **13**(3): 1–9.

Martin, D.E. (2006). A recursive algorithm for computing the distribution of the number of successes in higher-order Markovian trials, *Computational Statistics & Data Analysis* **50**(3): 604–610.

McHale, I. and Morton, A. (2011). A Bradley–Terry type model for forecasting tennis match results, *International Journal of Forecasting* **27**(2): 619–630.

Morris, B., Bialik, C. and Boice, J. (2016). How we're forecasting the 2016 US Open, `https://fivethirtyeight.com/features/how-were-forecasting-the-2016-us-open/`.

Newton, P.K. and Aslam, K. (2009). Monte Carlo tennis: A stochastic Markov chain model, *Journal of Quantitative Analysis in Sports* **5**(3): 1–44.

Newton, P.K. and Keller, J.B. (2005). Probability of winning at tennis. I: Theory and data, *Studies in Applied Mathematics* **114**(3): 241–269.

O'Malley, A.J. (2008). Probability formulas and statistical analysis in tennis, *Journal of Quantitative Analysis in Sports* **4**(2): 1–23.

Percy, D.F. (2015). Strategy selection and outcome prediction in sport using dynamic learning for stochastic processes, *Journal of the Operational Research Society* **66**(11): 1840–1849.

Pollard, G. (1983). An analysis of classical and tie-breaker tennis, *Australian Journal of Statistics* **25**(3): 496–505.

Radicchi, F. (2011). Who is the best player ever? A complex network analysis of the history of professional tennis, *PloS ONE* **6**(2): e17249.

Renick, J. (1976). Optimal strategy at decision points in singles squash, *Research Quarterly. American Alliance for Health, Physical Education and Recreation* **47**(3): 562–568.

Ryall, R. and Bedford, A. (2010). An optimized ratings-based model for forecasting Australian rules football, *International Journal of Forecasting* **26**(3): 511–517.

Šarčević, A., Pintar, D., Vranić, M. and Gojsalić, A. (2021). Modeling in-match sports dynamics using the evolving probability method, *Applied Sciences* **11**(10): 4429.

Schutz, R.W. (1970). A mathematical model for evaluating scoring systems with specific reference to tennis, *Research Quarterly: American Association for Health, Physical Education and Recreation* **41**(4): 552–561.

Silver, N. and Fischer-Baum, R. (2015). How we calculate NBA Elo ratings, `https://fivethirtyeight.com/features/how-we-calculate-nba-elo-ratings/`.

Spanias, D. and Knottenbelt, W. J. (2013). Predicting the outcomes of tennis matches using a low-level point model, *IMA Journal of Management Mathematics* **24**(3): 311–320.

Tversky, A. and Gilovich, T. (1989). The cold facts about the "hot hand" in basketball, *Chance* **2**(1): 16–21.

Wetzels, R., Tutschkow, D., Dolan, C., Van der Sluis, S., Dutilh, G. and Wagenmakers, E.-J. (2016). A Bayesian test for the hot hand phenomenon, *Journal of Mathematical Psychology* **72**: 200–209.

Wozniak, J. (2011). Inferring tennis match progress from in-play betting odds, *Project report*, Imperial College London, London.

Ana Šarčević is a researcher at the University of Zagreb, Faculty of Electrical Engineering and Computing. She is currently pursuing her PhD degree in electrical engineering and computing, also at the University of Zagreb. The field of her research is predictive analytics, which encompasses various statistical techniques from predictive modeling, machine learning and data mining.

Mihaela Vranić (PhD) is an assistant professor at the University of Zagreb, Faculty of Electrical Engineering and Computing. Her interests include data management, data analytics, educational data mining and machine learning. She has been involved in many industry projects both as a researcher and a project manager.

Damir Pintar (PhD) is an associate professor at the University of Zagreb, Faculty of Electrical Engineering and Computing. His interests include data management, machine learning, data mining and big data technologies, with a particular focus on predictive modeling and practical implementations of data science principles in industry and academia. He is actively involved in a number of data science-related projects both as a researcher and the project leader.

Appendix A

Let us assume the probability of winning a point on service of a chosen player A is equal to p. The probability that this player will lose a point on his service is therefore $1-p$. Player A will serve exactly a_s times to get from the current score x, y to the final score a, b. Out of these a_s services, player A must lose exactly $b-y$ points. The probability of this event is $(1-p)^{b-y}$. Player A must win all the remaining points on his serve. The probability of this event is $(p)^{a_s-(b-y)}$. The binomial coefficient counts all possible paths from the current score x, y to the final score a, b. From a_s-1 services (note that player A can not lose the last serve) it is necessary to choose $b-y$ services that player A will lose. This can be done in $\binom{a_s-1}{b-y}$ ways. When all this is multiplied, formula (1a) is obtained. Formula (1b) can be interpreted in the same way. The product of the first three factors in formula (1b) ($\binom{a_s}{b-y}p^{a_s-(b-y)}(1-p)^{b-y}$) represents the probability of reaching the score $40-40$, and the term $\frac{p^2}{p^2+(1-p)^2}$ represents the probability of winning the game after the score $40-40$. This mathematical expression can be derived by taking into account Fig. A1, which shows all possible changes of scores after the score $40-40$ (deuce).

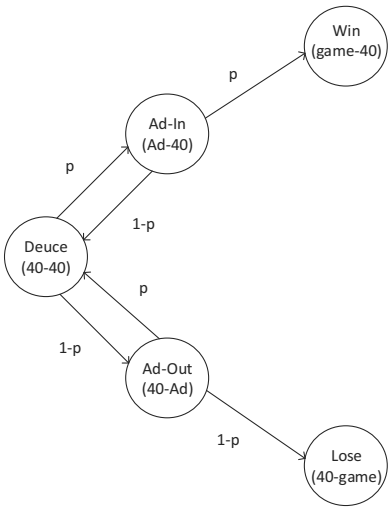

Fig. A1. Markov chain with possible changes of scores after deuce.

Let d denote the probability that player A will win the game if the current score is $40-40$. Looking at Fig. A1, it is possible to spot 3 ways to get from the *Deuce* state to the *Win* state. The first way is a direct path from the *Deuce* state to the *Win* state with a probability of $p \times p$. The second way is that player A first wins one point and goes into the *Ad-In* state with probability p. After that, player A loses a point and returns to the *Deuce* state with probability $1-p$. From the *Deuce* state, player A will win the game with probability d. The overall probability for this path is $p \times (1-p) \times d$. The last option is for player A to go from the state *Deuce* to the state *Ad-Out* by losing a point on his own serve with the probability $1-p$. After that, player A wins a point on his own serve with a probability p and returns to the *Deuce* state, and finally wins the game with the probability d from deuce. The overall probability for this described case is $(1-p) \times p \times d$. Summarizing all 3 cases gives the equation $d = p \times p + p \times (1-p) \times d + (1-p) \times p \times d$. By drawing the variable d from the given equation, the expression is obtained that player A will win the game if the current score is $40-40$.

Appendix B

This appendix compares the accuracies of the combinatorial approach and the recursive techniques. The comparison is made at the standard game level, and the same can be done for other levels of a tennis match. The combinatorial approach is first compared with forward recursion to estimate the final score of the standard game, and then with backward recursion to predict the winner of the standard game.

Table B1 shows the likelihood that a standard game will end with the score $game-30$ from randomly selected scores (accuracy test of (1a)). The probabilities that the standard game will finish after deuce (the score $40-40$) from randomly selected scores are shown in Table B2 (accuracy test of (1b)). As shown, the accuracy of both approaches (recursive and combinatorial) is identical on the game level. The same can be proven for the tiebreak game, the set and the match.

Table B1. Score $game-30$ from random scores.

Random score	Recursion (forward)	Combinatorics
0-15	0.207	0.207
30-15	0.288	0.288
40-15	0.240	0.240
0-0	0.207	0.207
15-30	0.216	0.216
0-30	0.130	0.130
40-0	0.096	0.096

Table B2. Score $game-40$ from random scores.

Random score	Recursion (forward)	Combinatorics
15-15	0.239	0.239
0-40	0.150	0.150
0-30	0.239	0.239
30-40	0.415	0.415
0-15	0.239	0.239
0-0	0.191	0.191
40-0	0.044	0.044

Table B3. Standard game.

Random score	Recursion (back)	Combinatorics
0-15	0.576	0.576
30-15	0.847	0.847
40-15	0.951	0.951
0-0	0.736	0.736
15-30	0.515	0.515
0-30	0.369	0.369
40-0	0.980	0.980

The accuracy of the results when predicting the match winner can be directly compared with the results of Barnett and Brown (2012). In this book, the evaluation of the standard game was made with the input parameter $p = 0.6$, therefore the same values were chosen in this paper.

When comparing the combinatorial and backward recursive approaches, the accuracy is identical at the standard game level. The same can be proven for a tiebreak game, set and match.

Received: 1 February 2021
Revised: 19 June 2021
Re-revised: 15 July 2021
Accepted: 25 July 2021