Neural networks and betting strategies for tennis - Separate Appendix

June 22, 2020

Appendix: Variables' definition

This appendix is devoted to the explanation of all the variables used in the work. All these variables are setted in the training and validation samples considering the information set up to the end of each match (see Def. (1)). All the variables employed in the testing sample instead use the information set up to the begin of each match (see Def. (2)). More details on this aspect are available upon request. The variables refer to the favourite (labelled as usual as f) and the underdog, (labelled as u), according to Def. (4). As described in the text, different models may consider different favourite's identification. For instance, the variables included in the ANN model specify as favourite the player with the shortest odd. The appendix will describe how each variable is configured and, more importantly, how each variable feeds the input nodes in the ANN. The variables used in the competing models are setted according to the original configuration (that is, according to the work of Lisi and Zanella (2017) for the LZR, the work of Klaassen and Magnus (2003) for the KMR, and so forth).

Winning frequency on the first serve (X_1)

Let $T1SP_{in,i,j}$ be the "total first service points in" by player i in match j. Moreover, let $T1SP_{win,i,j}$ be the number of won points on first serve by player i on match j. Hence, the winning percentage on the first serve for player i, on the match j, is obtained as:

$$X_{1,i,j} = \frac{T1SP_{win,i,j}}{T1SP_{in,i,j}}. \tag{A.1}$$

The variable included in the ANN is given by the following difference:

$$X_{1,j} = X_{1,f,j} - X_{1,u,j}. \tag{A.2}$$

Winning frequency on the second serve (X_2)

Let $T2SP_{in,i,j}$ be the "total second service points in" by player i in match j. Moreover, let $T2SP_{win,i,j}$ be the number of won points on second serve by player i on match j. Hence, the winning percentage on the second serve for

player i, on the match j, is obtained as:

$$X_{2,i,j} = \frac{T2SP_{win,i,j}}{T2SP_{in,i,j}}. \tag{A.3}$$

Therefore, the variable included in the ANN is given by:

$$X_{2,j} = X_{2,f,j} - X_{2,u,j}. \tag{A.4}$$

Won point return frequency (X_3)

Let $TSP_{i,j}$ be the total service points played by player i, on match j. The won point return percentage for the favourite player on the match j, is obtained as:

$$X_{3,f,j} = \frac{TSP_{u,j} - T1SP_{win,u,j} - T2SP_{win,u,j}}{TSP_{u,j}}. \tag{A.5}$$

The won point return frequency for the underdog player on the match j is instead obtained as:

$$X_{3,u,j} = \frac{TSP_{f,j} - T1SP_{win,f,j} - T2SP_{win,f,j}}{TSP_{f,j}}. \tag{A.6}$$

The won point return frequency for the underdog player on the match j, is instead obtained as:

$$X_{3,u,j} = \frac{TSP_{f,j} - T1SP_{win,f,j} - T2SP_{win,f,j}}{TSP_{f,j}}. \tag{A.7}$$

Hence, the variable included in the ANN is given by the following difference:

$$X_{3,j} = X_{3,f,j} - X_{3,u,j}. \tag{A.8}$$

Service points won frequency (X_4)

The service points won frequency for player i on the match j is obtained as:

$$X_{4,i,j} = \frac{T1SP_{win,i,j} + T2SP_{win,i,j}}{TSP_{i,j}}. \tag{A.9}$$

The variable included in the ANN is given by the following difference:

$$X_{4,j} = X_{4,f,j} - X_{4,u,j}. \tag{A.10}$$

Winning frequency on break point (X_5)

Let $BP_{faced,i,j}$ and $BP_{saved,i,j}$ be the number of break points faced and saved, respectively, by player i in the match j. Therefore, the winning percentage on break point for the favourite player on the match j is obtained as:

$$X_{5,f,j} = \frac{BP_{faced,u,j} - BP_{saved,u,j}}{BP_{faced,u,j}}. \tag{A.11}$$

Instead, the winning frequency on break point for the underdog player on the match j is obtained as:

$$X_{5,u,j} = \frac{BP_{faced,f,j} - BP_{saved,f,j}}{BP_{faced,f,j}}. \tag{A.12}$$

If $BP_{faced,i,j} = 0$ or $BP_{faced,i,j} = BP_{saved,i,j}$, $X_{5,i,j} = 0$. Finally, the variable included in the ANN is:

$$X_{5,j} = X_{5,f,j} - X_{5,u,j}. \tag{A.13}$$

First serve success frequency (X_6)

The first serve success frequency for player i on the match j is obtained as:

$$X_{6,i,j} = \frac{T1SP_{in,i,j}}{TSP_{i,j}}. \tag{A.14}$$

As usual, the variable included in the ANN is given by:

$$X_{6,j} = X_{6,f,j} - X_{6,u,j}. \tag{A.15}$$

Completeness (X_7)

The completeness is a proxy of the total strength of player i, in terms of service and return won points. More in detail, the service strength is given by the service points won frequency (X_4) and the return strength is given by

the won point return percentage (X_3). Hence, the completeness for player i in match j is:

$$X_{7,i,j} = X_{4,i,j} \cdot X_{3,i,j}. \tag{A.16}$$

Thus, the variable included in the ANN is:

$$X_{7,j} = X_{7,f,j} - X_{7,u,j}. \tag{A.17}$$

Advantage on serving (X_8)

The advantage on serving is a variable expressing the strength of the player on serve (again, X_4) with respect to the strength of player on return (X_3). Thus, for the favourite and the match j, the advantage on serving is given by:

$$X_{8,f,j} = X_{4,f,j} - X_{3,u,j}, \tag{A.18}$$

while for the underdog, the variable is obtained as:

$$X_{8,u,j} = X_{4,u,j} - X_{3,f,j}. \tag{A.19}$$

The variable included in the ANN is:

$$X_{8,j} = X_{8,f,j} - X_{8,u,j}. \tag{A.20}$$

Average number of aces per game (X_9)

Let $ACE_{i,j}$ and $TNG_{i,j}$ be the total number of aces and games won by player i in match j, respectively. The average number of aces per game is:

$$X_{9,i,j} = \frac{ACE_{i,j}}{TNG_{i,j}}. \tag{A.21}$$

The variable included in the ANN is:

$$X_{9,j} = X_{9,f,j} - X_{9,u,j}. \tag{A.22}$$

Minute-based fatigue (X_{10})

The Minute-based fatigue is a variable expressing how much a player is tired after long matches. It is highly plausible then the performance of

today's match depends on yesterday's match, mainly if this match lasted a lot of time. According to Sipko and Knottenbelt (2015), we consider only the last 3 matches played in consecutive days. This means that we take into account the minutes of match $j-2$ and $j-1$ and, naturally, the minutes of the current match j. As in Sipko and Knottenbelt (2015), we consider a discounting factor δ^{lag}, with $\delta < 1$, in order to weigh less the minutes of the match played two days ago (such that $lag = 2$) with respect to the weight of yesterday's match (if present), for which $lag = 1$. Hence, let $TM_{i,j}$ be the total minutes played in match j. The previous match, namely the match $j-1$, can be played the day before or also many days before. Suppose that the match $j-1$ is the last match of the season (that is, in November). Thus, the following match j is played after months, and not after days. Therefore, let $D(a,b)$ be the function returning the days between the date a and b. Finally, the minute-based fatigue is modelled as follows:

$$\begin{aligned} X_{10,i,j} = \quad & TM_{i,j} \\ & + \begin{cases} TM_{i,j-1} \cdot \delta & \text{if } D(j,j-1) = 1 \\ 0 & \text{otherwise} \end{cases} \\ & + \begin{cases} TM_{i,j-2} \cdot \delta^2 & \text{if } D(j,j-2) = 2 \\ 0 & \text{otherwise} \end{cases} \end{aligned} . \tag{A.23}$$

Finally, the variable included in the ANN is:

$$X_{10,j} = X_{10,f,j} - X_{10,u,j}. \tag{A.24}$$

Games-based fatigue (X_{11})

The Games-based fatigue is another variable expressing how much a player is tired after very struggling matches. Contrary to X_{10}, this variable relies on the fatigue depending on the games played. Even if a match has a short duration, in terms of minutes played, it can be very stressful from a mental point of view, in the case of many games disputed. An example could be matches whose sets are always terminated with tie-breaks. Let TNG_j be the total number of games played in the match j in which player i was a competitor. Hence, in order to take into account the potential fatigue arising from matches with many games, we set up the Games-based fatigue variable, which also considers a discounted factor δ^{lag} and the maximum

lag allowed is 2. This means that only the last two consecutive matches plus the today's match are used to define the variable:

$$\begin{aligned} X_{11,i,j} = \quad & TNG_j \\ & + \begin{cases} TNG_{j-1} \cdot \delta & \text{if } D(j, j-1) = 1 \\ 0 & \text{otherwise} \end{cases} \\ & + \begin{cases} TNG_{j-2} \cdot \delta^2 & \text{if } D(j, j-2) = 2 \\ 0 & \text{otherwise} \end{cases} \end{aligned} \quad . \tag{A.25}$$

Therefore, the variable included in the ANN is:

$$X_{11,j} = X_{11,f,j} - X_{11,u,j}. \tag{A.26}$$

Head-to-head of the favourite over the underdog (X_{12})

The head-to-head of the favourite over the underdog expresses the number of times (if any) that the favourite of the match j has defeated the underdog. Let $TNMP_{f_j,u_j,1:j}$ be the Total Number of Matches Played from the beginning of the dataset until the match j between the favourite of this match (f_j) and the relative underdog (u_j). Moreover, let $TNMWF_{f_j,u_j,1:j}$ be the Total Number of Matches Won by the Favourite of match j against the underdog of the same match. Thus, the head-to-head of the favourite over the underdog is simply calculated as:

$$X_{12,j} = \frac{TNMWF_{f_j,u_j,1:j}}{TNMP_{f_j,u_j,1:j}}. \tag{A.27}$$

Given that $TNMP_{f_j,u_j,1:j}$ includes the today's match, it results that $TNMP_{f_j,u_j,1:j} \geq 1, \forall j$.

Variable describing the ATP Rank (X_{13})

This variable denotes the ranks of both the player at the time of the match j, considering the tranformation suggested by Klaassen and Magnus (2003):

$$X_{13,j} = \log_2(Rank_{u,j}) - \log_2(Rank_{f,j}), \tag{A.28}$$

where $Rank_{i,j}$ represents the ATP rank of player i in the match j.

Variable describing the ATP Points (X_{14})

This variable denotes the points of both the player at the time of the match j. Higher ATP points mean that the players are better positioned in ranking. Let $Points_{i,j}$ be the ATP points of player i in the match j. Hence, the variable included in the ANN model is:

$$X_{14,j} = Points_{f,j} - Points_{u,j}. \tag{A.29}$$

Age (X_{15} and X_{15sq})

In tennis, the age of the player is a proxy of the experience. This does not mean that an older player has more chances to win against a younger player, but there is not a monotonic relationship such that the chances clearly decrease if the age increases. The variable Age included in the ANN model is nothing else that the difference of ages of the players at the time of the match j:

$$X_{15,j} = Age_{f,j} - Age_{u,j}. \tag{A.30}$$

In addition to the previous variable, we also include the squared version of $X_{15,j}$, labelled as $X_{15sq,j}$ and obtained as:

$$X_{15sq,j} = X_{15,j}^2. \tag{A.31}$$

Height (X_{16} and X_{16sq})

The height in tennis also is an important variable. Taller player have advantages on serve but are slower in the movements. Let $Height_{i,j}$ be the height of player i in match j, so that the variable included in the ANN model is:

$$X_{16,j} = Height_{f,j} - Height_{u,j}. \tag{A.32}$$

As for $X_{15,j}$, we also include the squared version of $X_{16,j}$, labelled as $X_{16sq,j}$ and obtained as:

$$X_{16sq,j} = X_{16,j}^2. \tag{A.33}$$

Surface winning frequency (X_{17})

In tennis the performances of player also depend on the surface of the match. Some players have more attitude on some surfaces, some others on other surfaces. The most prominent example is Rafael Nadal, which has, overall, a winning percentage on clay of 91.7%, at the time of this writing. In order to take into account the (updating) winning percentages on the surface of the match j, let $TNMP_{i,s,1:j}$ and $TNMW_{i,s,1:j}$ be the Total Number of Matches Played and the Total Number of Matches Won by player i on the surface s, with $s = \{Clay, Grass, Hard\}$, for the beginning of the dataset until match j. Hence, the surface winning frequency is:

$$X_{17,i,j} = \frac{TNMW_{i,s,1:j}}{TNMP_{i,s,1:j}}. \tag{A.34}$$

Therefore, the variable included in the ANN model is:

$$X_{17,j} = X_{17,f,j} - X_{17,u,j}. \tag{A.35}$$

Overall winning frequency (X_{18})

The previous variable X_{17} can be generalized without considering the surface. Hence, let $TNMP_{i,1:j}$ and $TNMW_{i,1:j}$ be simply the Total Number of Matches Played and the Total Number of Matches Won by player i, irrespective of the surface. The winning frequency of player i up to match j is:

$$X_{18,i,j} = \frac{TNMW_{i,1:j}}{TNMP_{i,1:j}}. \tag{A.36}$$

Hence, the variable included in the ANN is:

$$X_{18,j} = X_{18,f,j} - X_{18,u,j}. \tag{A.37}$$

Shin Implied Probabilities (X_{19})

The information offered by the bookmaker are the most accurate source of probability in sport forecasting. However, the published odds are only a proxy of the probability of winning. In order to derive the implied probability of winning, we consider the Shin normalization procedure. Formally, let $\pi_{i,j}$ be the probability odds for the player i in the match j, obtained

from the reciprocal of the published bookmaker odds $o_{i,j}$. Moreover, let $\Pi_j = \pi_{f,j} + \pi_{u,j}$ be the so-called booksum. Given that $\Pi_j > 1$, $\forall j$, let $m_j = \Pi_j - 1$ be defined as the margin. Finally, let $d_j = \pi_{f,j} - \pi_{u,j}$ be the distance between the probability odds of the favourite and underdog players, respectively. The implied probability according to the Shin method for the match j is labelled as $X_{19,j}$ and is calculated as:

$$X_{19,j} = \frac{\sqrt{z_j^2 + 4(1-z_j)\frac{\pi_{f,j}^2}{\Pi_j}} - z_j}{2(1-z_j)}, \tag{A.38}$$

z_j is assumed to be the proportion of bettors defined as insider traders. Jullien and Salanie (1994) and Cain et al. (2001) demonstrate that in sports with only two outcomes, z_j depends on the margin and the distance between the probability odds:

$$z_j = \frac{m_j(d_j^2 - \Pi_j)}{\Pi_j(d_j^2 - 1)}. \tag{A.39}$$

Current form of the players (X_{20})

Let $\overline{Rank}_{i,(j-6m):(j)}$ be the average rank of the last six months for player i. A player that is in a good form will have a higher current ranking with respect to the past average rank. Thus, the current form of the player i is defined as:

$$X_{20,i,j} = \left(\overline{Rank}_{i,(j-6m):(j)} - Rank_{i,j}\right). \tag{A.40}$$

Finally, the variable included in the ANN is:

$$X_{20,j} = X_{20,f,j} - X_{20,u,j}. \tag{A.41}$$

Bradley-Terry probability (X_{21})

McHale and Morton (2011) proposed to model the probability of winning of player i over the competitor as a function of (past) ability to win a game. Therefore, let $\alpha_{i,j}$ be the ability of player i to win a game in the match j. The probability to win a game for the favourite player over the underdog in

the match j is denoted as $PWG_{f,j}$ and is formally derived as:

$$PWG_{f,j} = \frac{\alpha_{f,j}}{\alpha_{f,j} + \alpha_{u,j}}. \tag{A.42}$$

Similarly, the probability that the underdog wins the game over the favourite is:

$$PWG_{u,j} = \frac{\alpha_{u,j}}{\alpha_{f,j} + \alpha_{u,j}}. \tag{A.43}$$

The abilities $\alpha_{i,j}$ for all the players are obtained by maximizing the likelihood provided in McHale and Morton (2011). Assuming independence among games, it is possible to calculate the probabilities of all the combinations of games letting the favourite to win the set. Let $Bin(TG, G_f)$ be the binomial formula indicating the probability that favourite wins G_f over TG games. Note that for simplicity we have omitted the suffix j indicating the match. Possible outcomes of a set won by the favourite are: (6,0), (6,1), (6,2), (6,3), (6,4), (7,5), (7,6). For instance, the expression (6,0) means that the favourite player wins six games over six, while the underdog wins zero games. Each possible outcome has a precise probability, depending on $PWG_{f,j}$ and $PWG_{u,j}$:

$$\begin{aligned}
Pr(6,0) =& Bin(6,6) = PWG_f^6 \\
Pr(6,1) =& Bin(6,5) PWG_f = 6 PWG_f^6 \cdot PWG_u \\
Pr(6,2) =& Bin(6,5) PWG_f PWG_u + Bin(6,4) \cdot PWG_f^2 = 21 PWG_f^6 PWG_u^2 \\
Pr(6,3) =& \cdots = 56 PWG_f^6 PWG_u^3 \\
Pr(6,4) =& \cdots = 126 PWG_f^6 PWG_u^4 \\
Pr(7,5) =& \cdots = 252 PWG_f^7 PWG_u^5 \\
Pr(7,6) =& \cdots = 504 PWG_f^7 PWG_u^6
\end{aligned}$$

Finally, the probability of winning a set for the favourite player PWS_f is:

$$PWS_f = Pr(6,0) + Pr(6,1) + Pr(6,2) + Pr(6,3) + Pr(6,4) + Pr(7,5) + Pr(7,6). \tag{A.44}$$

Tennis matches conclude when a player wins 2 sets over 3 or 3 over 5 sets. The former are called best-of-three matches, the latter best-of-five.

The variable included in the ANN is the probability that the favourite wins the match, given PWS_f, that is:

$$X_{21,j} = \begin{cases} PWS_{f,j}^2 + 2 \cdot PWS_{f,j}^2 \cdot (1 - PWS_{f,j}) & \text{if } j \text{ is a best-of-three match} \\ PWS_{f,j}^3 + 3 \cdot PWS_{f,j}^3 \cdot (1 - PWS_{f,j}) + & \text{if } j \text{ is a best-of-five match} \\ 6 \cdot PWS_{f,j}^3 \cdot (1 - PWS_{f,j})^2 & \end{cases}. \tag{A.45}$$

ATP ranking intervals (X_{22})

In setting their logit model, Lisi and Zanella (2017) suggest to consider as regressor the ATP rankings expressed in intervals. Therefore, we use the their same estimated intervals as resulting by the hierarchical cluster analysis. These intervals are reported in Table A.1 (second row), while in the third row the values assumed by the variable $X_{22,i,j}$ are reported. Therefore, the variable included in the ANN is given by:

$$X_{22,j} = X_{22,f,j} - X_{22,u,j}. \tag{A.46}$$

Table A.1: Ranking classification

	I Interval	**II Interval**	**III Interval**	**IV Interval**	**V Interval**
	0-560pt	561-920pt	921-1460pt	1461-2000pt	>2000pt
$X_{22,i,j}$	0	1	2	3	4

Notes: The table reports the values of the variable $X_{22,i,j}$ for each player i of the match j, according to the intervals of ATP points suggested by Lisi and Zanella (2017).

Home factor (X_{23})

According to Dixon and Coles (1997), home advantage is an important effect in sports that should be taken into account. Practically speaking, the home advantage for a player/team consists of a performance above the usual level when that player/team competes in matches played in his/their

own country. In tennis, Koning (2011) claims that the home advantage exists but only in male tennis. As done by Lisi and Zanella (2017), we model the home advantage as a dummy variable:

$$X_{23,i,j} = \begin{cases} 1 & \text{if player } i \text{ competes the match } j \text{ in his country} \\ 0 & \text{otherwise} \end{cases}. \tag{A.47}$$

Hence, the variable included in the ANN is:

$$X_{23,j} = X_{23,f,j} - X_{23,u,j}. \tag{A.48}$$

BCA winning probability (X_{24})

According to Barnett and Clarke (2005), it is possible to derive the winning probability of the whole match starting from the service points won (given by X_4). Formally, let p_i synthetically denote the relative frequency of service points won by player i for a match j. Therefore, the probabilities of all the possible situations within a game are:

$$\begin{aligned} Pr(40,0) =& p_i^4; \\ Pr(40,15) =& 4p_i^4(1-p_i); \\ Pr(40,30) =& 10p_i^4(1-p_i)^2; \\ Pr(Adv,40) =& 20p_i^5(1-p_i)^3/(1-2p_i(1-p_i)), \end{aligned}$$

where $Pr(Adv,40)$ is the probability of winning a game when the score is tied at 40-40. Hence, the probability that a player on serve wins the game, denoted as $PWGS_i$, is:

$$PWGS_i = Pr(40,0) + Pr(40,15) + Pr(40,30) + Pr(Adv,40). \tag{A.49}$$

Similarly to what done for the variable X_{21}, the probability of winning

a set for the player i is:

$$\begin{aligned}
Pr(6,0) =& PWGS_i^6 \\
Pr(6,1) =& PWGS_i^6(1-PWGS_i) \\
Pr(6,2) =& 21PWGS_i^6(1-PWGS_i)^2 \\
Pr(6,3) =& 56PWGS_i^6(1-PWGS_i)^3 \\
Pr(6,4) =& 126PWGS_i^6(1-PWGS_i)^4 \\
Pr(7,5) =& 252PWGS_i^7(1-PWGS_i)^5 \\
Pr(7,6) =& 504PWGS_i^7(1-PWGS_i)^6
\end{aligned}$$

The remaining probabilities are calculated as to Eqs. (A.44) and (A.45). Finally, once got the probability of winning a match according to the BC model, denoted as $X_{24,i,j}$, the variable included in the model is:

$$X_{24,j} = X_{24,f,j}/X_{24,u,j}. \tag{A.50}$$

Top-10 former presence (X_{25})

According to Del Corral and Prieto-Rodríguez (2010), the former presence as a top-10 player is a relevant variable that may concur to estimate the probability of winning. Therefore, let $X_{25,i,j}$ be the variable equalling one if in the last five years (preceding match j) player i is or has been a top-10 player. The variable included in the ANN model is:

$$X_{25,j} = X_{25,f,j} - X_{25,u,j}. \tag{A.51}$$

Handedness (X_{26}, X_{27}, X_{28} and X_{29})

In defining the variables of their probit model, Del Corral and Prieto-Rodríguez (2010) suggest to control also for the right- and left- handedness of the players through four different dummies, described in Table A.2.

Table A.2: Handedness

	Right-handed		**Left-handed**	
	Favourite	Underdog	Favourite	Underdog
X_{26}	$\surd$	$\surd$		
X_{27}			$\surd$	$\surd$
X_{28}	$\surd$			$\surd$
X_{29}		$\surd$	$\surd$	

Notes: The table reports the configuration of the dummy variables according to four possibilities. For simplicity, the variables do not have the match label j.

Grand-slam matches (X_{30})

Even though Del Corral and Prieto-Rodríguez (2010) employ a dataset consisting of only Grand Slam matches, they include a dummy variable for each tournament. Given that our dataset considers also matches played at a Master level, we simply set up a dummy variable, $X_{30,j}$, which equals one if the match j is played in a Grand Slam and zero otherwise.

Bookmaker info (X_{31})

Lisi and Zanella (2017) take advantage of the bookmaker odds only when their favourite player (the one with the highest ranking) has an odds greater than two. In this case, X_{31} equals to the value of the bookmaker odds. Otherwise, X_{31} equals zero. Because in the ANN model the information provided by the bookmaker are already included in the model through the variable X_{19}, X_{31} has been used only in the model of Lisi and Zanella (2017).

References

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

Cain, M., D. Law, and D. A. Peel (2001). The incidence of insider trading in betting markets and the Gabriel and Marsden anomaly. *The Manchester School 69*(2), 197–207.

Del Corral, J. and J. Prieto-Rodríguez (2010). Are differences in ranks good predictors for Grand Slam tennis matches? *International Journal of Forecasting 26*(3), 551–563.

Dixon, M. J. and S. G. Coles (1997). Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 46*(2), 265–280.

Jullien, B. and B. Salanie (1994). Measuring the Incidence of Insider Trading: A Comment on Shin. *The Economic Journal 104*(427), 1418–1419.

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

Koning, R. H. (2011). Home advantage in professional tennis. *Journal of Sports Sciences 29*(1), 19–27.

Lisi, F. and G. Zanella (2017). Tennis betting: Can statistics beat bookmakers? *Electronic Journal of Applied Statistical Analysis 10*(3), 790–808.

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

Sipko, M. and W. Knottenbelt (2015). Machine learning for the prediction of professional tennis matches. *MEng computing-final year project, Imperial College London*.