The temporalized Massey's method

Abstract

We propose and throughly investigate a temporalized version of the popular Massey's technique for rating actors in sport competitions. The method can be described as a dynamic temporal process in which team ratings are updated at every match according to their performance during the match and the strength of the opponent team. Using the Italian soccer dataset, we empirically show that the method has a good foresight prediction accuracy.

1 Introduction

Rating and ranking in sport have a flourishing tradition. Each sport competition has its own official rating, from which a ranking of players and teams can be compiled. The challenge of many sports' fans and bettors is to beat the official rating method: to develop an alternative rating algorithm that is better than the official one in the task of predicting future results. As a consequence, many sport rating methods have been developed. Amy N. Langville and Carl D. Meyer even wrote a (compelling) book about (general) rating and ranking methods entitled *Who's #1?* (Langville and Meyer, 2012).

In 1997, Kenneth Massey, then an undergraduate, created a method for ranking college football teams. He wrote about this method, which uses the mathematical theory of least squares, as his honors thesis (Massey, 1997). Informally, at any given time t, Massey's method rates a team i according to the following two factors: (a) the difference between points for and points against i, or point spread of i, up to time t, and (b) the ratings of the teams that i matched up to time t. Hence, highly rated teams have a large point differential and matched strong teams so far. Below in the ranking are teams that did well but had an easy schedule as well as teams that did not so well but had a tough schedule.

In this paper we propose a temporalized version of the original Massey's method. The idea is the following. For a given team i and time t, the original Massey rates i according to the point spread of i up to time t and the ratings of the teams that i matched up to time t. Notice, however, that the rating of a matched team j is computed with respect to time t, and not, as we argue it should be more reasonable, with respect to the (possibly previous) time when i and j matched. Suppose, for instance, that i and j matched at time 7, when team j was strong (high in the ranking), and now, at time 19, team j lost positions in the ranking and is thus weaker. The original Massey's method adds up to the rating of i the *current* low rating of j computed at time 19, and not the *past* high rating of j computed at time 7. For example, in college football, say Notre Dame is highly touted early in the season, and Miami beats #1 ranked Notre Dame in September.

Then Notre Dame suffers some injurys (physical or psyhcological) and loses three more games. Traditional methods will miss the fact that Miami beat Notre Dame when they were at full strength and confident. The temporalized Massey's method we propose solves this issue. At any given time t of the season, the temporalized Massey's method rates a team i according to (a) the point spread of i up to time t, and (b) the ratings of the teams that i matched up to time t computed with respect to the *time they matched*.

Various authors addressed dynamic modelling of sports tournaments. A recent account of can be found for example in Cattelan, Varin, and Firth (2013). In the paper, only the outcomes (win-draw-loss) of the matches, and not point spreads, are considered. The abilities of the home and visiting teams are assumed to evolve separately in time following an exponentially weighted moving average process ruled by a constant coefficients linear recurrence. In our approach the two abilities are twisted together and the evolution is described by a variable coefficients recurrence.

A good survey of dynamic models for teams strengths in NFL can be found in Glickman and Stern (2017). Generally teams' abilities are assumed to evolve through a first order autoregressive process. For example in Harville (1980) this strategy is used to model season to season changes of team' abilities while in Glickman and Stern (1998) week to week changes. As we will see, our approach gives, as season proceeds, a greater importance to the history of the results compared with the one given by an autoregressive model.

Chartier, Kreutzer, Langville, and Pedings (2011) propose nonuniform weighting for sports rankings. Their technique allows to weight differently late season play but also, for example, home court advantage or high-pressure games. The authors discuss and experiment various strategies for choosing the weights: in the simplest one the weights linearly increase from the first day of the season to the last day. They apply their strategy also to Colley's method a close sibling of Massey's method. We remark that the temporalization technique that we develop for Massey's method can easily be extended to Colley's method.

A popular time-varying rating system used is sport competitions is Elo's method (Elo, 1978, Langville and Meyer, 2012). There is an intriguing similarity between Elo's method and temporalized Massey's method. Both methods update the old rating of a team in terms of the same ingredients: the current performance of the team and the rating of the opponent team. However, the two methods mix these ingredients in different ways, and hence the resulting recipe differs. While Elo uses a logistic (exponential) function to mix performance and opponent rating, Massey linearly combines the two.

The paper is organized as follows. Section 2 reviews the original Massey's method. We propose the temporalized interpretation of the Massey's method in Section 3. In Section 3.1 we investigate the algebra of the proposed method while in

Section 3.2 we describe a bootstrap-based procedure for quantifying the uncertainty of the rating estimates. We apply the Massey's method to the Italian Serie A soccer league (season 2015-2016) in Section 4. Finally, we conclude in Section 5.

2 The Massey's method for sports ranking

In this section we offer a brief introduction to the original Massey's method. A more general introduction can be found in Glickman and Stern (2017). The main idea of Massey's method, as proposed in Massey (1997), is enclosed in the following equation:

$$r_i - r_j = y_k$$

where r_i and r_j are the ratings of teams i and j and y_k is the margin of victory for game k of team i. If there are n teams who played m games, we have a linear system:

$$Xr = y \tag{1}$$

where X is a $m \times n$ matrix such the k-th row of X contains all 0s with the exception of a 1 in location i and a -1 in location j, meaning that team i beat team j in match k (if match k ends with a draw, either i or j location can be assigned 1, and the other -1). Observe that, if e denotes the vector of all 1's, then $Xe = 0$. Let $M = X^T X$ and $p = X^T y$. Notice that

$$M_{i,j} = \begin{cases} \text{the negation of the \# of matches between } i \text{ and } j & \text{if } i \neq j, \\ \text{\# of games played by } i & \text{if } i = j. \end{cases}$$

and p_i is the signed sum of point spreads of every game played by i. Clearly the entries of p sum to 0, in fact $e^T p = e^T X^T y = (Xe)^T y = 0$. The Massey's method is then defined by the following linear system:

$$Mr = p \tag{2}$$

which corresponds to the least squares solution of system (1).

We observe how the Massey's team ratings are in fact interdependent. Indeed, Massey's matrix M can be decomposed as

$$M = D - A,$$

where D is a diagonal matrix with $D_{i,i}$ equal to the number of games played by team i, and A is a matrix with $A_{i,j}$ equal to the number of matches played by team i against team j. Hence, linear system (2) is equivalent to

$$Dr - Ar = p, \tag{3}$$

or, equivalently

$$r = D^{-1}(Ar + p) = D^{-1}Ar + D^{-1}p.$$

That is, for any team i

$$r_i = \frac{1}{D_{i,i}} \sum_j A_{i,j} r_j + \frac{p_i}{D_{i,i}}. \tag{4}$$

This means, and the same observation can be found in Glickman and Stern (2017), that the rating r_i of team i is the sum $r_i^{(1)} + r_i^{(2)}$ of two meaningful components:

1. the mean rating of teams that i has matched

$$r_i^{(1)} = \frac{1}{D_{i,i}} \sum_j A_{i,j} r_j;$$

2. the mean point spread of team i

$$r_i^{(2)} = \frac{p_i}{D_{i,i}}.$$

It is worth pointing out that the ratings computed by Massey's method correspond to averages. Hence, it could happen that a team that plays with good performances a limited number of matches against strong teams obtains an extremely high and not justified rating. Actually this effect has been clearly discussed in Chartier, Harris, Hutson, Langville, Martin, and Wessel (2014). To overcome this problem the authors propose to introduce a dummy team that defeats all the teams that played a number of matches below a suitable cutoff.

In order to better understand the behaviour of the method, it is interesting to analyse what happens to Massey's system at the end of the season, assuming a round-robin competition in which all n teams matched all other teams exactly once. In this case, the opponents rating component

$$r_i^{(1)} = -\frac{r_i}{n-1},$$

where we have used the fact that $\sum_i r_i = 0$, and the point spread component

$$r_i^{(2)} = \frac{p_i}{n-1},$$

hence

$$r_i = r_i^{(1)} + r_i^{(2)} = -\frac{r_i}{n-1} + \frac{p_i}{n-1},$$

and thus

$$r_i = \frac{p_i}{n}.$$

Hence, the final rating of a team is simply the mean point spread of the team. It is possible to be a bit more precise about this property of Massey's method by exploiting the properties of the set of eigenvalues, or spectrum, of the Laplacian matrix $M = D - A$. The spectrum reflects various aspects of the structure of the graph G_A associated with A, in particular those related to connectedness. It is well known that the Laplacian is singular and positive semidefinite (recall that $M = X^T X$ and $Xe = 0$) so that its eigenvalues are nonnegative and can be ordered as follows:

$$\lambda_1 = 0 \leq \lambda_2 \leq \lambda_3 \leq \ldots \leq \lambda_n.$$

It can be shown that $\lambda_n \leq n$, see for example Brouwer and Haemers (2012). The multiplicity of $\lambda_1 = 0$ as an eigenvalue of the Laplacian can be shown to be equal to the number of the connected components of the graph, see again Brouwer and Haemers (2012). If the graph of the matches is connected or, equivalently, M is irreducible, as we assume in the following, $\lambda_2 \neq 0$ is known as *algebraic connectivity* of the graph and is an indicator of the effort to be employed in order to disconnect the graph.

We can write the spectral decomposition of M as $M = UDU^T$ where U is orthogonal and its first column is equal to $e/\sqrt{n}$, and $D = \text{diag}(0, \lambda_2, \ldots, \lambda_n)$. From $Mr = p$ we obtain $r = UD^+U^T p$ where $D^+ = \text{diag}(0, \frac{1}{\lambda_2}, \ldots, \frac{1}{\lambda_n})$. Now

$$r - \frac{p}{n} = UD^+U^T p - \frac{p}{n} = U\left[D^+ - \frac{I}{n}\right]U^T p,$$

where I is the identity matrix. Observe that the first component of the vector $U^T p$ is equal to zero so that

$$r - \frac{p}{n} = U\left[D^+ - \frac{I}{n}\right]U^T p = U\left[D^+ - \frac{\tilde{I}}{n}\right]U^T p,$$

where $\tilde{I} = \text{diag}(0, 1, \ldots, 1)$. If we denote with $\|\cdot\|$ the Euclidean norm we obtain

$$\|r - \frac{p}{n}\| = \|U\left[D^+ - \frac{\tilde{I}}{n}\right]U^T p\| \leq \|p\| \max_{k=2,\ldots,n}\left|\frac{1}{\lambda_k} - \frac{1}{n}\right| \leq \|p\|\frac{n-\lambda_2}{n\lambda_2},$$

where we used the fact that the Euclidean norm of an orthogonal matrix is equal to one. Hence, in the case of a round-robin competition, as the algebraic connectivity λ_2, as well as the other eigenvalues, approach n, that is, as more and more matches are played, the vector r approaches p/n and the equality is reached when the graph of the matches becomes complete.

3 Temporalized Massey's method

We propose a temporalized variant of the original Massey's method. The main idea of the new proposal is to compute the rating of a matched team with respect to the time when the match was played, and not with respect to the current time, as Massey does.

We consider a temporal process of matches between pairs of teams that occur at a given time. Each element of the process is a tern (i, j, t) where i and j are the teams that matched and t is the time of the match. Time is discrete and is represented with natural numbers $0, 1, \ldots$. We assume that each team plays at most one match at any given time. Matches (of different teams) that occur at the same time are considered to happen simultaneously.

Let $s_i(t)$ be the difference of the points for team i and the points against team i in the match of time t, where we assume $s_i(t) = 0$ if i does not play at time t. Let $m_{i,t}$ be the number of games that team i played until time t. Let $j_1, \ldots, j_{m_{i,t}}$ be the teams matched by i until time t and $t_1, \ldots, t_{m_{i,t}}$ be the timestamps of these matches. Then the rating of team i at time t is defined as follows. We set $r_i(0) = 0$ for all teams i. Hence all teams are initially equally ranked. For any team i, if i did not play so far, that is $m_{i,t} = 0$, then its rating is still null. Otherwise, if $m_{i,t} > 0$, we have that, for every $t \geq 1$:

$$r_i(t) = \frac{1}{m_{i,t}} \sum_{k=1}^{m_{i,t}} (r_{j_k}(t_k - 1) + s_i(t_k)). \tag{5}$$

This means that the rating $r_i(t)$ of team i at time t is the sum $r_i^{(1)}(t) + r_i^{(2)}(t)$ of two meaningful components:

- the mean *historical* rating of teams that i has matched:

$$r_i^{(1)}(t) = \frac{1}{m_{i,t}} \sum_{k=1}^{m_{i,t}} r_{j_k}(t_k - 1);$$

- the mean point spread of team i at time t:

$$r_i^{(2)}(t) = \frac{1}{m_{i,t}} \sum_{k=1}^{m_{i,t}} s_i(t_k).$$

Notice that we set $r_i(0) = 0$ for all teams, meaning that at the start of the competition all teams are considered equal. This might be not always realistic: we sometimes know that some teams are potentially stronger than others. Hence, an alternative

solution is to set $r_i(0) = \rho_i$, where ρ_i is the exogenous strength of i before the competition starts. For instance, we can set the exogenous strength to be proportional to the rating of the team at the end of the previous season.

We illustrate the proposed method with the following simple example (a complete application is discussed in Section 4). The table below shows the results of 6 matches (numbered from 1 to 6), divided in 3 days representing a different time (numbered from 1 to 3), involving 4 fictitious teams (labelled A, B, C, D):

match	day	team 1	team 2	score 1	score 2
1	1	A	C	2	1
2	1	B	D	2	1
3	2	A	D	3	0
4	2	B	C	1	1
5	3	A	B	1	0
6	3	C	D	1	0

While there is no doubt that A is the leader of the ranking (it won all matches) and D is the weakest team (it lost all matches), the challenge between B and C is more controversial: each has won one match, lost another match and drew when they matched together.

The following spread matrix contains the cumulative spread of each team at each day. Initially B has a small advantage over C, which is maintained in the second day, and lost in the last day, when they finish with the same spread. Notice that the spread of the last day corresponds, up to a multiplicative constant, to the original Massey rating (see Section 2). Hence, according to the spread or to orignal Massey's method, there is no difference between B and C at the end of the season.

	1	2	3
A	1	4	5
B	1	1	0
C	-1	-1	0
D	-1	-4	-5

However, the temporalized Massey's method tells us a different story. The following matrix contains the temporalized Massey rating for each day and each team:

	1	2	3
A	1	1.5	1.33
B	1	0	0.17
C	-1	0	-0.17
D	-1	-1.5	-1.33

The first day the rating is exactly the spread, hence B has an little advantage over C. Interestingly, this advantage is lost at day 2, while the spread is still in favor of B. The reason is that at day 2, teams B and C matched together and they drew. However, before of the match (at day 1), B was stronger than C, hence C drew against a stronger team with respect to B. Finally, at day 3, B is over C in the ranking (while the spread is equal). In fact, at day 3, B lost, but against the strongest team of the competition (A), and C won, but against the weakest team of the competition (D). In summary, B and C drew the match together (but when B was stronger), and then they both lost against A and won against D. But the subtle difference, which is captured only by the temporalized version of Massey, is that B lost against A at day 3, when A was the strongest team, while C lost against A at day 1, when A was as strong as all other teams. Similarly, B won against D at day 1, when D was as strong as all other teams, while C won against D at day 3, when D was the weakest team. This determines the difference in the final ranking of the temporalized Massey's method.

3.1 A closer look to temporalized Massey's method

Let us consider more closely the temporalized Massey's equation (5). Clearly, if at time t team i does not play then $r_i(t) = r_i(t-1)$. On the contrary, suppose that at time t team i matches with team j (in other words $t = t_k$ for some k). Then the rating of i at time t can be defined in terms of the ratings at $t-1$ of teams i and j as well as the point spread of team i at the current time t:

$$r_i(t) = \frac{m_{i,t}-1}{m_{i,t}} r_i(t-1) + \frac{s_i(t) + r_j(t-1)}{m_{i,t}}. \tag{6}$$

Similarly, the rating of j at time t is:

$$r_j(t) = \frac{m_{j,t}-1}{m_{j,t}} r_j(t-1) + \frac{s_j(t) + r_i(t-1)}{m_{j,t}}. \tag{7}$$

Notice that losing against a strong team can still make the day for the loser, but winning against a weak team can result is a drop of the rating of the winner. We can rewrite Equation 6 as follows:

$$r_i(t) = \alpha_{i,t}\, r_i(t-1) + \beta_{i,t}\, r_j(t-1) + \beta_{i,t_k}\, s_i(t), \tag{8}$$

where $\alpha_{i,t} = (m_{i,t}-1)/m_{i,t}$ and $\beta_{i,t} = 1/m_{i,t}$. Notice that $\alpha_{i,t} + \beta_{i,t} = 1$. Hence, the rating of team i at time t is a convex combination of the ratings at time $t-1$ of teams i and of the matched team j plus a fraction of the spread of i at time t. Of course, by expanding recurrence (8) one obtains back equation (5).

We would like to attract the attention of the reader to the fact that coefficients $\alpha_{i,t}$ and $\beta_{i,t}$ vary in time. More precisely, as the number of games $m_{i,t}$ of team i grows, the component $\alpha_{i,t}$ approaches 1 and $\beta_{i,t}$ vanishes to 0. This means that, if i played few matches and hence $m_{i,t}$ is small, then the latest performance of i can make a significant difference in the ranking position of team i. On the other hand, as $m_{i,t}$ grows, new results can only slightly move the ranking position of the team. This is coherent with the general idea that an established reputation is difficult to shake.

Interestingly, if teams i and j played the same number of matches at time t, that is $m_{i,t} = m_{j,t}$, it is easy to realize that, after a match between i and j, we have that $r_i(t) + r_j(t) = r_i(t-1) + r_j(t-1)$. This means that what one team gains is lost by the other, and the cumulative rating of the system is the same before and after the match. In particular, in a round-robin competition in which at each day in the competition each team matches another team not matched before, it happens that, if initially all teams have rating equal to 0, at any day the cumulative rating of all teams in the competition is 0. It is worth noticing that this property holds also for the original Massey's method but is lost if teams play a different number of games.

From (6) it follows that every rating $r_i(t)$ is a linear combination of spreads whose nonnegative coefficients can be placed in a matrix $C^{(i,t)}$ such that

$$r_i(t) = \sum_{k=1}^{n} \sum_{l=1}^{t} C_{k,l}^{(i,t)} s_k(l).$$

From (6) it is possible to obtain an equivalent relation for these matrices in the case where i matches with j at time t

$$C^{(i,t)} = \frac{m_{i,t}-1}{m_{i,t}} C^{(i,t-1)} + \frac{1}{m_{i,t}} E^{(i,t)} + \frac{1}{m_{i,t}} C^{(j,t-1)}, \tag{9}$$

where $E_{k,l}^{(i,t)} = 1$ if $(i,t) = (k,l)$ and $E_{k,l}^{(i,t)} = 0$ otherwise. Clearly only the first t columns of $C^{(i,t)}$ contain entries different from zero.

As an example let us consider again the 4 fictitious teams A, B, C and D of the previous example that now is convenient to denote with the integers from 1 to

4. In this simple example every team plays at each time hence $m_{i,t} = t$. Therefore Equation (9) becomes

$$C^{(i,t)} = \frac{t-1}{t} C^{(i,t-1)} + \frac{1}{t} E^{(i,t)} + \frac{1}{t} C^{(j,t-1)}, \qquad t = 1,2,3 \tag{10}$$

and this yields

$$C^{(1,1)} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad C^{(1,2)} = \begin{bmatrix} 1/2 & 1/2 \\ 0 & 0 \\ 0 & 0 \\ 1/2 & 0 \end{bmatrix}, \quad C^{(1,3)} = \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/6 & 1/6 & 0 \\ 1/6 & 0 & 0 \\ 1/3 & 0 & 0 \end{bmatrix} \tag{11}$$

where only the nontrivial columns of the matrices are shown. Of course if the 4 teams are involved in a round robin competition then in the 4th day A and C match together again and

$$C^{(1,4)} = \begin{bmatrix} 7/24 & 1/4 & 1/4 & 1/4 \\ 5/24 & 1/8 & 0 & 0 \\ 5/24 & 1/12 & 1/12 & 0 \\ 7/24 & 1/24 & 0 & 0 \end{bmatrix},$$

where, as before, only the nontrivial columns of the matrix are shown. It is possible to verify that $C^{(i,t)}$ for $i = 2,3,4$ are just row permutations of $C^{(1,t)}$.

Notice that the sum of the coefficients in the columns of the matrices $C^{(i,t)}$ in our example has a quite regular behaviour. Let us denote with $C^{(i,t)}_{:,l}$ the l-th column of $C^{(i,t)}$. By using (10), for $l = t$ we obtain

$$e^T C^{(i,t)}_{:,t} = \frac{1}{t} e^T E^{(i,t)}_{:,t} = \frac{1}{t},$$

that is true in particular for $l = t = 1$. Making use of induction we obtain for $l \leq t-1$

$$e^T C^{(i,t)}_{:,l} = \frac{t-1}{t} e^T C^{(i,t-1)}_{:,l} + \frac{1}{t} e^T C^{(j,t-1)}_{:,l} = \frac{t-1}{t}\frac{1}{l} + \frac{1}{t}\frac{1}{l} = \frac{1}{l}.$$

As a consequence, the sum of the entries of $C^{(i,t)}$ is equal to $H_t = \sum_{l=1}^{t} \frac{1}{l}$ for each team i. The number H_t is known as the t-th *harmonic number*. It holds that

$$H_t \min_{\substack{1 \leq k \leq n \\ 1 \leq l \leq t}} s_k(l) \leq r_i(t) \leq H_t \max_{\substack{1 \leq k \leq n \\ 1 \leq l \leq t}} s_k(l).$$

It is well known that $\lim_{t \to \infty} H_t - \ln t = \gamma$ where $\gamma \approx 0.577$ is known as Euler-Mascheroni constant. This implies that the range of the ratings of temporalized

Massey's method increase very slowly in t. For example $H_{38} \approx 4.2$. Moreover, the above inequality tells us that ratings and spreads, which are added up in the temporalized Massey's equation, are of the same order of magnitude.

It is worth noticing that the temporalized Massey's rating of team i at time t is a linear combination of past spreads (performances) of all teams, not just of team i, with multiplicative coefficients described by matrix $C^{(i,t)}$. This contrasts with the original Massey's rating for team i. Indeed, as shown in Section 2, as time goes on, the original Massey's rating for i approaches a linear combination of past performances of i, without considering the performances of other teams.

It is interesting to observe that, if the teams have exogenous initial strengths, then the linear combination of spreads has to be complemented with a linear combination of them. For example, in order to compute $r_1(4)$, one has to add to the combination of spreads whose coefficient appear in $C^{(1,4)}$, the value obtained from

$$\frac{7}{24}r_3(0) + \frac{5}{24}r_4(0) + \frac{5}{24}r_1(0) + \frac{7}{24}r_2(0),$$

since the first match of A is against C and the first match of B is against D.

Finally, it is useful to compare recurrence (8) with its constant coefficient equivalent, namely:

$$r_i(t) = \alpha\, r_i(t-1) + \beta\, r_j(t-1) + \beta\, s_i(t), \tag{12}$$

where now $\alpha, \beta > 0$ are constant with $\alpha + \beta = 1$, and again t is the timestamp of the match of i with j. By expanding this recurrence we obtain

$$r_i(t) = \alpha^{m_{i,t}} r_i(0) + \beta \sum_{k=1}^{m_{i,t}} \alpha^{m_{i,t}-k} \Big(r_{j_k}(t_k - 1) + s_i(t_k) \Big), \tag{13}$$

where $m_{i,t}$ is the number of games that team i played until time t, while $j_1, \ldots, j_{m_{i,t}}$ are the teams matched by i until time t, and $t_1, \ldots, t_{m_{i,t}}$ are the timestamps of these matches. Comparing Equations 5 and 13, we capture the difference between the varying and constant coefficient recurrences. In Equations 5, past performances of a team are treated homogeneously, while with Equations 13 the past is progressively forgotten, giving more importance to recent performances, and this forgetfulness is quicker if α is small (close to 0).

To obtain an alternative intuition of this difference we study the matrices $C^{(i,t)}$ for our simple round robin example. It is not difficult to obtain

$$C^{(1,1)} = \beta \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad C^{(1,2)} = \beta \begin{bmatrix} \alpha & 1 \\ 0 & 0 \\ 0 & 0 \\ \beta & 0 \end{bmatrix}, \quad C^{(1,3)} = \beta \begin{bmatrix} \alpha^2 & \alpha & 1 \\ \alpha\beta & \beta & 0 \\ \beta^2 & 0 & 0 \\ \alpha\beta & 0 & 0 \end{bmatrix},$$

where only the nontrivial columns of the matrices are shown. In addition

$$C^{(1,4)} = \beta \begin{bmatrix} \alpha^3+\beta^3 & \alpha^2 & \alpha & 1 \\ \alpha^2\beta+\alpha\beta^2 & \alpha\beta & 0 & 0 \\ \alpha\beta^2+\alpha^2\beta & \alpha\beta & \beta & 0 \\ \alpha^2\beta+\alpha\beta^2 & \beta^2 & 0 & 0 \end{bmatrix},$$

where again only the nontrivial columns are shown. Notice that, not taking into account the factor β, the entries of each column of these matrices sum up to a power of the binomial $\alpha+\beta$. Since we assumed $\alpha+\beta=1$, we have that, for $l=1,\ldots,t$,

$$e^T C^{(i,t)}_{:,l} = \beta.$$

This result highlights the difference between the varying-coefficient and the constant-coefficient techniques: the latter gives progressively more and more importance to the recent matches with respect to the former.

Again, if exogenous initial strengths are present then the linear combination of spreads has to be complemented with a combination of initial strengths. For example in order to compute $r_1(4)$ to the combination of spreads one has to add

$$\begin{aligned} &\alpha^4 r_1(0)+(\alpha^3\beta+\beta^4)r_3(0)+(\alpha^2\beta^2+\alpha\beta^3)r_4(0) \\ &\quad +(\alpha\beta^3+\alpha^2\beta^2)r_1(0)+(\alpha^2\beta^2+\alpha\beta^3)r_2(0). \end{aligned}$$

3.2 Uncertainty evaluation

The temporalized Massey's method is a deterministic procedure which provides point evaluation of the ratings, taking into account the actual time evolution of team abilities. However, since these findings are based on sport competition data characterized by sampling variability, the acknowledgement of the consequent sampling variability of the evaluating procedure is a crucial, focal point. A proper quantification of the uncertainty of these estimates provides an effective mean for assessing whether the ratings of two teams are significantly different. Furthermore, a time-dependent forecasting distribution could be readily specified for predicting the match results of the forthcoming day.

In order to address this issue, we consider a simple statistical model for describing the match outcomes, following the approach proposed by Massey (1997) for measuring team's ability, and generalized in many subsequent research papers aiming at improving sport rating methods (see, for example, Glickman and Stern, 2017, and references therein). More precisely, we define a basic linear regression model, where the margin of victory for a particular game between two teams is

specified as a linear function of the difference in team strength, with an additional random error term. In this framework, the evaluation of the uncertainty in the estimated ratings can be performed through a parametric or a non-parametric bootstrap analysis (see, for example, Davison and Hinkley, 1997).

We assume that $y_{ij}(t)$, namely the score difference in the match of time t involving team i and team j, for every $t \geq 1$ and $i \neq j$, is defined as

$$y_{ij}(t) = r_i(t-1) - r_j(t-1) + \varepsilon_{ij}(t), \tag{14}$$

where $\varepsilon_{ij}(t)$ is a sequence of uncorrelated random error terms with mean 0 and variance $\sigma^2(t)$, for every i, j, which may vary according to the time t of the match. Moreover, we set the initial rating $r_i(0) = 0$ for all teams. Thus, in this basic model, the score differences are interpreted as random variables with mean value given by the differences in strength before the match, as estimated by the temporalized Massey's method, and a time dependent unknown variance parameter. It is common to complete the model specification by assuming, if supported by a model diagnostic procedure, that the random residuals $\varepsilon_{ij}(t)$, and hence the differences in score $y_{ij}(t)$, follow a normal distribution or another continuous real-valued distribution. Although score differences are integer-valued, a suitable continuous distribution (and in particular the normal distribution) may represent a convenient, easy-to-use approximating model (Stern, 1991, Harville, 2003). Clearly, the values for the score differences, thus obtained for simulation or prediction purposes, have to be rounded to the nearest integer.

If the model (14), with a suitable assumption on the distribution of the error term, gives an adequate description for the score differences and, according to the observed competition data, we get the estimates $\widehat{r}_i(t)$ and $\widehat{\sigma}^2(t)$ for the team ratings $r_i(t)$ and the residual variance $\sigma^2(t)$, we may consider a simple bootstrap parametric procedure for estimating the bias $b_i(t) = E\{\widehat{r}_i(t)\} - r_i(t)$ and the variance $v_i(t) = V\{\widehat{r}_i(t)\}$ of the temporalized Massey's ratings. Notice that we adopt the same notation for the estimated ratings and for the associated sample statistics, since the distinction will be easily inferred by the context. Then, if $\{y_{ij}^b(t), t \geq 1, i \neq j\}$, $b = 1, \ldots, B$, are parametric bootstrap samples simulated from the estimated model (14) and $\widehat{r}_i^b(t)$, $b = 1, \ldots, B$, are the corresponding estimates for the team ratings, the parametric bootstrap estimate for the bias and the variance are, respectively,

$$\begin{aligned} b_i^{boot}(t) &= \frac{1}{B}\sum_{b=1}^{B} \widehat{r}_i^b(t) - \widehat{r}_i(t) = \bar{r}_i^{boot}(t) - \widehat{r}_i(t), \\ v_i^{boot}(t) &= \frac{1}{B}\sum_{b=1}^{B-1} \left\{ \widehat{r}_i^b(t) - \bar{r}_i^{boot}(t) \right\}^2. \end{aligned}$$

Using these resampling estimates for the bias and the variance, it is immediate to obtain an estimate for the standard error associated to $\widehat{r}_i(t)$ and to specify the corresponding $1-2\alpha$ equi-tailed confidence interval

$$\left[\widehat{r}_i(t) - b_i^{boot}(t) - z_{1-\alpha}\sqrt{v_i^{boot}(t)},\ \widehat{r}_i(t) - b_i^{boot}(t) + z_{1-\alpha}\sqrt{v_i^{boot}(t)}\right], \qquad (15)$$

where $z_{1-\alpha}$ is the $1-\alpha$-quantile of the standard normal distribution. This basic confidence interval relies on the assumption that the sample statistic $\widehat{r}_i(t)$ follows, at least approximatively, a normal distribution. This can be assessed by considering a suitable diagnostic analysis on the simulated estimates $\widehat{r}_i^b(t)$, $b = 1,\ldots,B$. If the normal approximation turns out to be poor, alternative bootstrap-based confidence intervals can be defined (Davison and Hinkley, 1997, chapter 5).

Whenever there is no plausible statistical model for describing the random error terms $\varepsilon_{ij}(t)$, the bootstrap analysis can be carried out in a non-parametric fashion. In this case, the bootstrap samples are obtained by repeated sampling from the set of the observed residuals $\widehat{\varepsilon}_{ij}(t) = y_{ij}(t) - \{\widehat{r}_i(t-1) - \widehat{r}_j(t-1)\}$, $t \geq 1$, $i \neq j$. In order to account for the potential modification of the probability distribution of the residuals over time, we may consider repeated sampling from moving, overlapping blocks of observed residuals within a fixed temporal width. If $\{\varepsilon_{ij}^b(t), t \geq 1, i \neq j\}$, $b = 1,\ldots,B$, are the bootstrap samples for the residuals, the bootstrap data will be defined as $y_{ij}^b(t) = \widehat{r}_i(t-1) - \widehat{r}_j(t-1) + \varepsilon_{ij}^b(t)$. The computation of the bootstrap estimate for the bias and the variance of $\widehat{r}_i(t)$ and the specification of the associated confidence intervals are the same as in the parametric case.

4 Application to Italian soccer league

In this section, we analyse the Italian Serie A soccer league of season 2015-2016, which is a round-robin competition with 20 teams and 38 days (each pair of teams matches twice).

In Figure 1 we depict the Kendall correlation between pairs of ranking methods among temporalized Massey (T-M), original Massey (M), and official ranking (O). As days pass, we accrue more and more information about the real strength of teams, and all correlations increase. In particular at day 38, end of the season, we have complete information, and correlations coefficients are close to 1 (0.98 for T-M vs M, 0.93 for M vs O, and 0.91 for T-M vs O), although there are differences in the rankings, in particular when the official compilation is involved. Nevertheless, during the season, when information is partial, the corresponding rankings diverge significantly, and correlation coefficients are far from 1, in particular with respect

to the official ranking. For instance the coefficients at day 10 are: 0.80 for T-M vs M, 0.73 for M vs O, and 0.62 for T-M vs O. Moreover, over all days, the association between Massey and official rankings is higher than the association between temporalized Massey and official rankings.

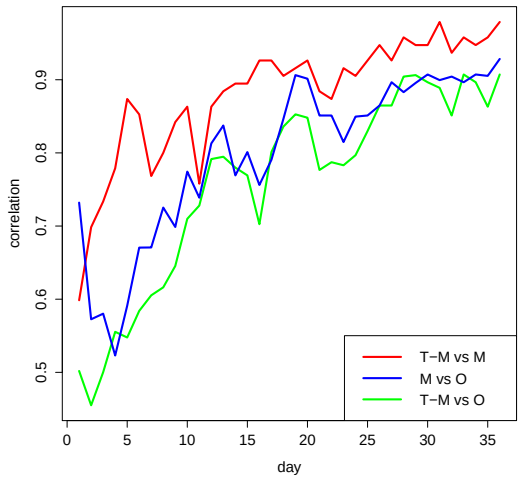

Figure 1: The Kendall correlation coefficients among temporalized Massey (T-M), original Massey (M), and the official ranking (O) as days go by from 3 to 38.

A rigorous test for a rating system is *foresight prediction accuracy* (Langville and Meyer, 2012): how well the vector $r(t)$ of ratings computed at day t can predict the winners at day $t+1$? More precisely, the foresight prediction accuracy of a method is the number of victories that the method corrected foresaw divided by the total number of victories of that competition (we ruled out the ties). Hence, accuracy of 0 means no predictions were correct, while accuracy of 1 means that all predictions were correct. We also computed accuracy introducing a home-field advantage, which was empirically determined for each method and added to the rating of the team playing at home. A home-field advantage matters for foresight prediction in time-varying methods: since initially all teams are rated equal, then in the beginning, before there is enough competition to significantly distinguish the teams' ratings, home-field consideration is the only criterion that the method can use to draw a distinction between two teams. We compared three time-varying rating methods with and without home-field advantage (see Table 1): official rating of the Italian soccer league, temporalized Massey's method, and Elo's method. Temporalized Massey is slightly more predictive than Elo and significantly better than

the official rating. Moreover, for all methods, introducing the home-field advantage has a significant impact in the prediction accuracy. We also computed, for the temporalized Massey's method, the foresight prediction accuracies at each day of the competition (with home-field advantage). The histogram of accuracies is depicted in Figure 2. Only 2 predictions are below the threshold of 50% of accuracy corresponding to randomness (notice that the 3 predictions in the 40%-50% histogram bar are in fact equal to 50%). On the other hand, most of predictions (78%) are above 60% of accuracy, with 12 predictions (32%) above 80% of accuracy and 3 predictions (8%) with 100% of accuracy.

Method	**Without HFA**	**With HFA**
Temporalized Massey	0.611	0.702
Elo	0.611	0.695
Official	0.589	0.674

Table 1: Foresight prediction accuracies with and without home-field advantage (HFA).

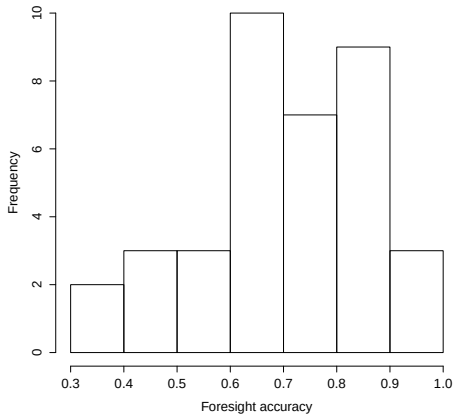

Figure 2: Histogram of foresight prediction accuracies at each day of the competition (with home-field advantage) for temporalized Massey's method.

Related to prediction accuracy, consider the following story. Teams Inter and Juventus had a peculiar season in 2015-2016. Inter immediately won the first matches, but with low spread of points. On the other hand, the start of Juventus was

disastrous. This led Inter well above Juventus in the official ranking, with a maximum distance of 10 points at days 5 and 6. From day 10, however, Juventus started an incredible row of wins, culminating at day 19 when the two teams were pair in official standings. Finally, at day 38, Juventus powerfully won the championship with 24 points above Inter. In Figure 3 we depict the temporal dynamics of the official, original Massey, and temporalized Massey rankings during the first round of the championship. The superiority of Juventus with respect to Inter is not witnessed by the official ranking until the end of the round. On the other hand, Massey and in particular its temporalized version predicted this supremacy well before the end of the round.

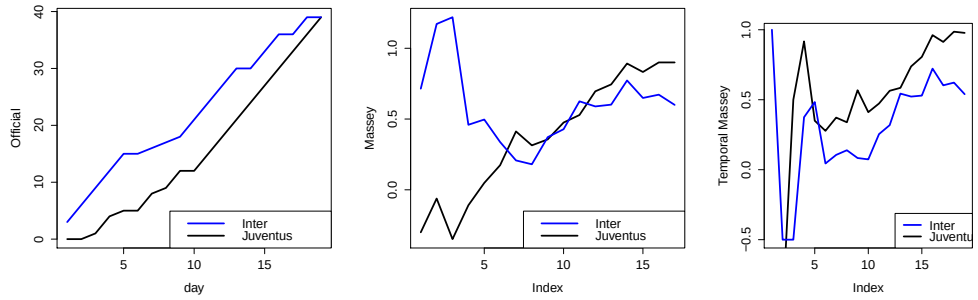

Figure 3: The temporal dynamics of the ratings of Juventus and Inter in the first round (19 days).

Furthermore, with the simple bootstrap procedure outlined in Section 3.2, it is possible to achieve a preliminary evaluation of the uncertainty of the estimated ratings given by the temporalized Massey's method. Then, it will be possible to assess whether the difference in the temporal dynamics of the ratings of two teams, such as Juventus and Inter as represented in Figure 3, can be considered as reasonably significant.

A preliminary graphical analysis on the observed residuals, reported in Figure 4, suggests that the normal distribution could be a satisfactory model for the error term in (14) and that the associated variance $\sigma^2(t)$ changes considerably throughout the season. Since the time evolution of the variability does not follow a simple functional pattern, the estimates $\widehat{\sigma}^2(t)$ are obtained using a moving variance procedure, which returns the sample variance of the observed residuals over a sliding window of length w (an odd integer value) centred about $t \in \{(w+1)/2, \ldots, 38-(w-1)/2\}$. The values related to the first and the last $(w-1)/2$ days are assumed to be equal to the first and the last computable estimates, respectively.

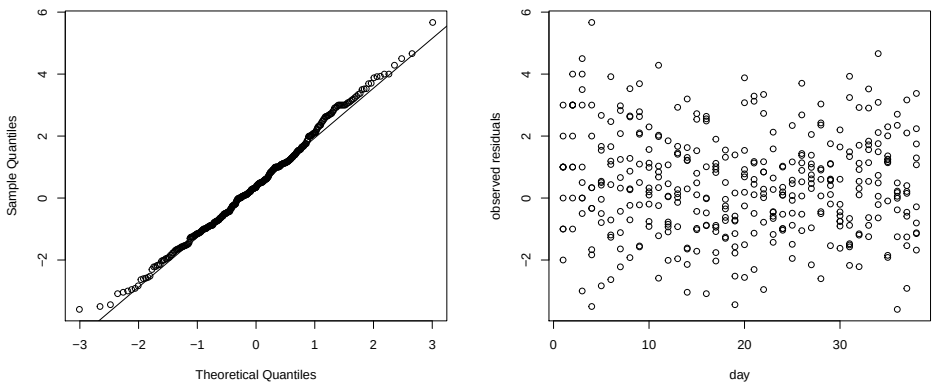

Figure 4: Italian Serie A soccer league 2015-2016. Normal qq-plot (left) and temporal evolution throughout the season (right) of the the observed residuals.

We find out that a window length $w = 5$ assures a reasonable amount of smoothing in the estimated sequence.

Table 2 shows the teams of the Italian Serie A soccer league 2015-2016 ranked according to the ratings at the end of the season (day 38), as estimated by the temporalized Massey's method. Furthermore, we report the parametric bootstrap estimates for the bias and the standard deviation of the associated ratings estimators and the bootstrap confidence intervals, with confidence level $1-2\alpha = 0.95$, specified according to equation (15). The use of the normal approximation for the confidence limits is empirically validated by the normal qq-plot of the bootstrap simulated estimates for the ratings. Moreover, almost the same conclusions can be obtained using the non-parametric bootstrap procedure described at the end of Section 3.2. This close similarity in the final results confirms the validity and the robustness of the conclusions drawn from the parametric analysis.

Notice that, for some teams, the estimates for the bias are substantial and also the sign is not always the same along the column. For this reason, the confidence intervals are computed using equation (15), where the bias-corrected estimates for the ratings are considered as interval midpoints. In our opinion, these unexpected bias values may reveal that the basic model (14) does not provide a complete explanation for the score differences. Improving the model, by introducing, for example, suitable additional explanatory variables, can possibly reduce the value of the bias term.

We emphasize that the 95% confidence intervals have to be only interpreted

Rank	Team	Rating	Bias	St Dev	95CI
1	Juventus	1.422	-0.587	0.268	(1.483, 2.535)
2	Napoli	1.240	-0.178	0.263	(0.903, 1.933)
3	Roma	1.013	-0.261	0.265	(0.755, 1.793)
4	Fiorentina	0.498	0.312	0.256	(-0.315, 0.688)
5	Inter	0.261	0.125	0.259	(-0.372, 0.645))
6	Sassuolo	0.141	0.061	0.264	(-0.437, 0.598)
7	Milan	0.101	-0.138	0.265	(-0.280, 0.758)
8	Lazio	0.090	-0.197	0.262	(-0.226, 0.802)
9	Chievo	-0.038	0.436	0.261	(-0.985, 0.037)
10	Torino	-0.040	0.328	0.263	(-0.883, 0.148)
11	Genoa	-0.079	-0.183	0.260	(-0.407, 0.614)
12	Atalanta	-0.142	0.034	0.266	(-0.698, 0.344)
13	Empoli	-0.267	-0.193	0.261	(-0.585, 0.439)
14	Bologna	-0.306	-0.056	0.257	(-0.754, 0.253)
15	Sampdoria	-0.380	0.457	0.260	(-1.346, -0.328)
16	Carpi	-0.390	-0.366	0.262	(-0.538, 0.489)
17	Udinese	-0.627	0.045	0.264	(-1.189, -0.154)
18	Verona	-0.689	0.005	0.261	(-1.206, -0.181)
19	Palermo	-0.705	0.288	0.262	(-1.506, -0.481)
20	Frosinone	-1.106	0.066	0.260	(-1.682, -0.663)

Table 2: Italian Serie A soccer league 2015-2016. Teams ranked according to ratings at the end of the season estimated by the temporalized Massey's method, boostrap-based estimates for the bias and the standard deviation, and 95% bootstrap confidence intervals.

as interval estimates for the team ratings and they can not be considered for pairwise comparison of the estimated team strengths. The non-overlap criterion, according to which, if two intervals fail to overlap, the corresponding ratings are interpreted as significantly different, is more conservative and less powerful than the standard testing procedure at the 5% significance level based on the difference of the ratings (Schenker and Gentleman, 2001). With equal standard errors, and under the normality and the independence assumptions, the non-overlap criterion achieves the required 5% significance level when we consider a normal quantile equal to $1.96\sqrt{2}/2$ instead of 1.96. Thus, in order to justify this criterion for pairwise ratings comparison with a 5% significance level, the confidence level of the intervals has to be reduced to $1-2\alpha = 0.834$ (Goldstein and Healy, 1995).

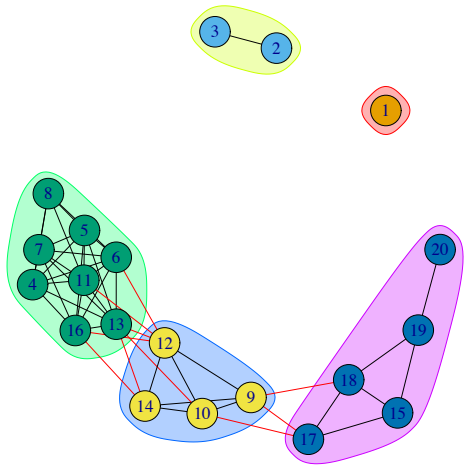

Figure 5: The 83.4% bootstrap confidence interval graph with highlighted communities of teams. Black edges connect nodes within the same group, red edges run between nodes of different groups.

The graph in Figure 5 gives a relational view of the 83.4% bootstrap confidence intervals of the teams. The graph is as follows. The nodes are the 20 teams numbered according to the temporalized Massey ratings (the rank given in the first column of Table 2). We used the Jaccard index to draw the edges between nodes. Recall that the Jaccard index measures similarity between two finite sets, and is defined as the size of the intersection divided by the size of the union of the two sets. Initially, we traced an edge between two teams if the Jaccard index of the corresponding confidence intervals is positive, that is, if the two intervals intersect. Then, we removed all edges with a Jaccard index less than the median of the edge Jaccard scores, which turns out to be 0.4. Finally, we ran an *optimal* community detection procedure on the resulting graph (Newman, 2010). The procedure outputs a partition of the nodes into the communities highlighted in Figure 5. Informally, the solution partition nodes into cohesive groups maximizing the number of edges that connect nodes within the same group. A community hence corresponds to a set of teams that might be considered equivalent with respect to the strength as measured with the temporalized Massey procedure. Notice the placement of teams 11 (Genoa), 13 (Empoli) and 16 (Carpi): the Massey ratings underestimate

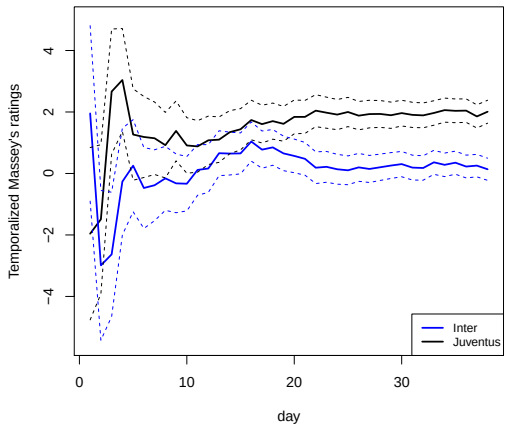

Figure 6: Italian Serie A soccer league 2015-2016. Temporal dynamics of the 83.4% bootstrap confidence intervals for the ratings of Juventus and Inter given by the temporalized Massey's method.

these teams, while the uncertainty analysis determines their inclusion into a higher-ranked community. Indeed, these teams have the largest negative bias among those of the second part of the ranking. On the other hand, teams 9 (Chievo), 10 (Torino) and 15 (Sampdoria) are overestimated by their Massey ratings and are placed into a lower-ranked community by the uncertainty analysis. Notice that these teams have the largest positive bias among those of the second part of the ranking.

Finally, in Figure 6, we consider the temporal dynamics of the 83.4% bootstrap confidence intervals for the ratings of Juventus and Inter, as estimated by the temporalized Massey's procedure. As emphasized before, this method points out the superiority of Juventus with respect to Inter well before the end of the first round, but this supremacy becomes evident, and statistically significant at the approximate 5% level for each single pairwise comparison, only at the beginning of the second round, when the intervals turn out to be well-separated.

5 Conclusion

We introduced a temporalized version of the popular Massey's method for rating actors in sport competitions. The idea of the new method is quite simple: rate matched teams with respect to the time when the match was played. We showed that

the resulting method can be described as a dynamic process in which the rating of any team is modified when the team plays according to the performance of the team during the game and the strength of the matched team before the game. We applied the new method to the Italian soccer league showing a good foresight prediction accuracy.

A future research line concerns the generalization of the basic statistical model, considered in this paper for the specific aim of producing a preliminary quantification of the uncertainty related to the rating estimates. As emphasized in Section 4, these estimates present an unusual bias, which may indicate that model (14) does not provide a fully satisfactory explanation for the score differences. The introduction of suitable covariate information, such as the home field advantage, and the specification of a more flexible temporal modelling for the ratings $r_i(t)$ and the variance $\sigma^2(t)$ of the error term might improve both the descriptive and the forecasting accuracy of the temporalized Massey's procedure.

References

Brouwer, A. E. and W. H. Haemers (2012): *Spectra of graphs*, Universitext, Springer, New York.

Cattelan, M., C. Varin, and D. Firth (2013): "Dynamic Bradley-Terry modelling of sports tournaments," *Journal of the Royal Statistical Society Series C Applied Statistics*, 62, 135–150.

Chartier, T., E. Kreutzer, A. Langville, and K. Pedings (2011): "Sports ranking with nonuniform weighting," *Journal of Quantitative Analysis in Sports*, 7.

Chartier, T. P., J. Harris, K. R. Hutson, A. N. Langville, D. Martin, and C. D. Wessel (2014): "Reducing the effects on unequal number of games on rankings," *IMAGE The bullettin of the International Linear Algebra Society*, 52, 15–23.

Colley, W. N. (2002): "Colley's bias free college football ranking method: The Colley matrix explained," Available at `http://www.colleyrankings.com/matrate.pdf`.

Davison, A. C. and D. V. Hinkley (1997): *Bootstrap methods and their application*, Cambridge, UK: Cambridge University Press.

Elo, A. E. (1978): *The Rating of Chess Players, Past and Present.*, New York: Arco.

Glickman, M. E. and H. S. Stern (1998): "A state-space model for national football league scores," *Journal of the American Statistical Association*, 93, 25–35.

Glickman, M. E. and H. S. Stern (2017): "Estimating team strength in NFL," in *Handbook of Statistical Methods and Analyses in Sports*, CRC Press, chapter 6, 113–136.

Goldstein, H. and M. J. R. Healy (1995): "The graphical presentation of a collection of means," *Journal of the Royal Statistical Society Series A Statistics in Society*, 158, 175–177.

Harville, D. (1980): "Predictions for national football league games via linear-model methodology," *Journal of the American Statistical Association*, 75, 516–524.

Harville, D. A. (2003): "The selection or seeding of college basketball or football teams for postseason competition," *Journal of the American Statistical Association*, 98, 17–27.

Langville, A. N. and C. D. Meyer (2012): *Who's #1? The science of rating and ranking*, Princeton University Press, Princeton, NJ.

Massey, K. (1997): *Statistical models applied to the ratings of sports teams*, Bachelor's thesis, Bluefield College.

Newman, M. E. J. (2010): *Networks: An introduction*, Oxford: Oxford University Press.

Schenker, N. and J. F. Gentleman (2001): "On judging the significance of differences by examining the overlap between confidence intervals," *The American Statistician*, 55, 182–186.

Stern, H. (1991): "On the probability of winning a football game," *The American Statistician*, 45, 179–183.