Working papers series

WP ECON 22.11

Pairwise contests: wins, losses, and strength

Carmen Herrero
Department of Economics, Universidad de Alicante

Antonio Villar
Department of Economics, Universidad Pablo de Olavide & ISEAK

Keywords: sports competitions; pairwise comparisons; tournaments; dominant eigenvector, win-loss; strength; Premier League.

JEL Classification: : D71, Z19.

UNIVERSIDAD PABLO DE OLAVIDE SEVILLA
Department of Economics

Pairwise contests: wins, losses, and strength

Carmen Herrero

University of Alicante

Antonio Villar

Universidad Pablo de Olavide & ISEAK

ABSTRACT

Sports competitions represent an interesting family of evaluation problems involving pairwise comparisons. In this context, the alternatives are contending teams, and the comparison is made in terms of outcomes. Different evaluation protocols have been proposed in the literature, aimed at getting more robust estimates of the teams' worth or better predictions of their future achievements. We present here a new evaluation protocol that can be described in two steps. First, we modify the teams' outcomes by introducing a penalty function that ponders the points accrued by the points lost. This step already produces an interesting evaluation procedure, the *relative performance* rule (the ratio between points won and points lost). Second, we define a new procedure that also considers the strength of the teams in the evaluation. We call this new evaluation protocol the *relative strength* rule.

Keywords: sports competitions; pairwise comparisons; tournaments; dominant eigenvector, win-loss; strength; Premier League.

JEL numbers: D71, Z19

Acknowledgments: The first author acknowledges financial support for this research from the Spanish Ministerio de Ciencia e Innovación, (project PID2019-107081GBI00) and the Generalitat Valenciana (Prometeo 2021/073). The second author acknowledges financial support from the Spanish Ministerio de Ciencia e Innovación (project PID2019-104452RB-I00) and the Junta de Andalucía (projects FEDER UPO-1263503 and P18-RT-2135).

1. Introduction

The evaluation of alternatives is the essence of decision theory and social choice. It may derive from aggregating judgments, measuring characteristics, or comparing outcome variables, and result in a global ranking, a selection of best options, or a numerical evaluation. The standard framework consists, therefore, of a set of alternatives and a collection of partial evaluations (e.g., orderings or ratings provided by a number of "judges", or measurements of different characteristics), and the evaluation problem is that of transforming those several partial evaluations into a single global evaluation. The properties of transitivity and completeness are usually key to solve that problem.

Pairwise comparisons provide a more general framework for the evaluation of alternatives, that requires neither transitivity nor completeness, while one can keep track of how each alternative fares relative to each other. This is an approach with a long tradition, that we find in Condorcet's consistent rules (Fishburn, 1977) or the literature on tournaments (Laslier, 1997). Sports competitions may be regarded as a special family of evaluation problems based on pairwise comparisons, whose informational inputs correspond to outcomes of matches between teams or players. Different sports have different rules to compute victories and defeats and to determine the ranking of the competing teams. Yet, there is scope for the analysis of those competitions from a conceptual viewpoint, as they represent evaluation problems with a specific structure and particular features. This context helps thinking about introducing in the evaluation some elements that may refine the conventional way of rating alternatives.

In most sports competitions the team that wins a match gets credited some points whereas the team that losses the match gets zero. Then the evaluation of the teams in a competition is given by the total points accumulated. Yet, we find different ways of computing wins and losses in the literature on sports competitions, which propose richer ways to evaluate the teams' performance. Those evaluation protocols may have multiple purposes: help enhance competitiveness, provide estimates of the teams' commercial worth, obtain better predictions of future outcomes, define shadow prices of some assets (e.g. broadcasting rights), or gather competitors into comparable categories. Note that, besides the conceptual appeal of those considerations, there are many people interested in sports competitions and lots of money involved, so the evaluation issue has quite a bite.

There are two main aspects that those richer evaluation protocols introduce. On the one hand, considering the points won and lost by the teams in the competition, rather than the points won exclusively. On the other hand, pondering the points obtained by the strength of the competing teams. Both aspects affect the incentives of the teams in the competition and help obtain a richer evaluation of the contenders.

Perhaps the most popular evaluation rules that consider the points won and lost in competitions are the Elo (1978) rating and the rankings proposed by Massey (1997) and Colley (2002).

The Elo rating was initially conceived to classify chess players, then extended to many other competitions. The basic idea is to give player *i*, when playing against player *j*, a rating r_i that reflects the probability of winning against its opponent, based on the frequency of their victories. The player's evaluation is adjusted by making the change in the rating proportional to the difference between the actual score, s_{ij}, and the expected score, $E(s_{ij})$. That is, $r_i^t - r_i^{t-1} = K\left(s_{ij} - E(s_{ij})\right)$, where $K > 0$ is a sensitivity constant, and $E(.)$ the expected score is drawn from a logistic function of the rating.

The idea behind Massey's and Colley's procedures is that the evaluations of the competing teams should be proportional to their differential outcomes. In the case of Massey, the evaluation obtains as the vector $\boldsymbol{r}$ that solves the system $\mathbf{M}\boldsymbol{r} = \boldsymbol{p}$, where $\mathbf{M} = \{m_{ij}\}$ is the so-called Massey matrix, which is given by $m_{ii} = N$ (the total number of games played by team *i*), $m_{ij} = -n$ (where n stands for the number of games played between any two teams), and $\boldsymbol{p}$ is the vector of cumulative score differences of each team. Colley proposes to evaluate outcomes according to the solution to the system $\mathbf{C}\boldsymbol{r} = \boldsymbol{b}$, where $\mathbf{C} = \{c_{ij}\}$ is the Colley matrix, given by: $c_{ii} = 2 + N$, $c_{ij} = -n, i \neq j$. Vector $\boldsymbol{b}$ corresponds to the cumulative win-loss differentials for each team, that is, $b_i = 1 + 0.5(w_i - l_i)$, where w_i, l_i are the total games won and lost by *i*, respectively. In both cases, the evaluation derives from the solution of a linear system, applies a proportionality principle, and focuses on outcome differences.

An alternative stream of contributions introduces the idea of pondering outcomes on the strength of the teams. A key reference in this approach is Keener (1993), who states explicitly (p. 81), "To each participant in a contest we wish to assign a score that is based on the interactions with other participants. The assigned score should depend on both the outcome of the interaction and the strength of its opponents." To solve the evaluation problem by computing both the outcomes and the strength of the opponents derived from those outcomes, Keener recurs to a fixpoint argument, which in this context corresponds to the dominant eigenvector of a matrix $\mathbf{A} = \{a_{ij}\}$, where a_{ij} reflects what *i* gets in its confrontation with *j*. In the simplest case, he takes $a_{ij} = 1$, if team *i* wins, equal to ½ in case of a tie, and equal to 0 if it losses. By letting r_j stand for the strength of team *j*, the score of team *i* corresponds to the sum of the outcomes weighted by the strength of the competitors, $s_i = \sum_j a_{ij} r_j$, Keener proposes that the score be proportional to the strength. That is, $\mathbf{A}\boldsymbol{r} = \lambda\boldsymbol{r}$. He also discusses different ways of defining the outcomes that can be embedded within this evaluation procedure. So, here we find again a proportionality principle but now considering the strength of the teams at the heart of the evaluation.

A variant of Keener's contribution is that in which the evaluation procedure consists of a Markov chain (e.g. Govan, 2008, Govan, Langville & Meyer, 2009). This procedure derives from a "voting" matrix. In a pairwise confrontation between players *i* and *j*, each player gives a "vote" to the other. So, m_{ij} is the vote of *j* to *i*, and m_{ji} the other way around. The columns of the matrix are normalized so that it becomes a Markov matrix (dividing each element by the sum of the column). The evaluation is given by the stable distribution of this Markov matrix.

The reducibility of the outcome matrices, which is usually found in empirical applications, is regarded as problematic since it implies that some teams may receive the same evaluation even though they present different outcomes. To deal with this "inconvenience" (see below) Vaziri, Yih & Morin (2018) introduce the so-called $(1, \alpha)$ method. In this method, the winning team gives a vote of 1 to the losing team, and the losing one gives a vote of $\alpha > 1$ to the winning team. The Markov matrix becomes, therefore, irreducible. Yet, the final evaluation depends critically on that α parameter, which is chosen arbitrarily by the evaluator. Here we find similar features to Keener's evaluation protocol, but the cardinal evaluation depends on the choice of the α parameter.

Vaziri, Yih & Morin (2018) show that the $(1, \alpha)$ method satisfies three desirable properties, for a suitable range of values of the parameter α: (1) To consider the strength of the opponents (*comprehensiveness*); (2) To provide incentives to always win a match (*monotonicity*); and (3) Independence of the sequence of matches (*fairness*). They also show that other Markov procedures, Elo's, or those obtained as the solutions of linear equations systems, fail to satisfy at least one of these three properties.

We propose in this paper an evaluation rule that computes both wins and losses, and the relative strength of the alternatives, in the realm of pairwise contests. We keep the framework of sports competitions to facilitate the exposition, even though the rule can be applied to more general evaluation problems. This evaluation rule, called the ***relative strength rule***, provides cardinal evaluations of the teams in a competition. It can be described as applying Keener's proportionality principle to a modified outcome matrix in which the original outcomes have been transformed into relative outcomes, by pondering points won by points lost. This transformation already yields an interesting evaluation procedure, the *relative performance* rule, which corresponds to the ratio between total points won and total points lost.

The paper is organized as follows. Section 2 presents the evaluation protocol and provides an application that helps visualize how it works. It refers to the English Premier League, in which we compare the evaluation derived from our formula, that corresponding to the relative performance rule, and the official one (points accrued through the competition). Section 3 is devoted to discussing some relevant features of our evaluation rule. A few final words, in Section 4, close the work.

2. Sports competitions

2.1 The evaluation of teams in a competition

A sports competition consists of a set of teams, $M = \{1,2,\dots,m\}$, that compete between them in pairwise confrontations a number n of rounds in a given time period. To facilitate the discussion, we assume that competitions adopt the format of symmetric round-robin tournaments. That is, each team plays exactly n times against each other, and all rounds are equally worthy. For each pairwise contest between teams $i, j \in M$, and each round $h \in N = \{1,2,\dots,n\}$, the term $p_{ij}(h) \in [0, K] \subset \mathbb{R}$ describes the ***partial outcome*** obtained by i when playing against j in round h. There are several ways of defining this outcome variable (see below), but what follows does not depend on a particular choice. We shall refer to the units in which those outcomes are measured as "points". Since no team can obtain points by playing against itself, we let $p_{ii}(h) = 0, \forall\, i, h$.

We denote by p_{ij} the ***outcome*** of team i with respect to team j, which corresponds to the sum of all the points obtained when playing against team j along with the competition, i.e. after n rounds. That is,[1]

$$p_{ij} = \sum_{h=1}^{n} p_{ij}(h) \qquad [1]$$

All the relevant information of the competition can be summarized into a square m matrix $\mathbf{P}$ whose entries are the outcomes p_{ij} corresponding to the total points obtained by team i when playing against team j. We shall therefore identify an evaluation problem, involving m teams, with the matrix $\mathbf{P} = \left(p_{ij}\right)_{i,j=1}^{m} \in \mathbb{R}_{+}^{m^2}$ (a matrix that, by construction, has zeroes into the main diagonal). We aim to provide an evaluation of the teams in M that participate in a competition, based on those outcomes. More precisely, we look for

[1] When not all rounds have equal weight in a competition, the outcomes should be redefined as:

$$p_{ij} = \sum_{h=1}^{n} \delta_h p_{ij}(h) \qquad [1']$$

Where δ_h is the weight attached to round h. An example of those asymmetries appears in tennis tournaments, in which different rounds have different points that can be obtained.

an ***evaluation rule*** $f: \mathbb{R}_+^{m^2} \rightarrow \mathbb{R}_+^m$, such that to each problem $\mathbf{P}$ associates an m-vector $f(\mathbf{P})$, such that $f_i(\mathbf{P}) \geq f_j(\mathbf{P})$ implies that team i is regarded as better than or equal to team j.

The simplest evaluation rule is that in which we associate with each team the aggregate outcomes obtained along with the competition, i.e., the sum of all points achieved. We call this rule the ***performance rule***, denoted by f^P, which can be described as follows. Let $\mathbf{1}$ stand for the unit m-vector, then: $f^P(\mathbf{P}) = \mathbf{P1}$. The rule f^P is a well-defined continuous function, whose ith component is given by:

$$f_i^P(\mathbf{P}) = \sum_{j=1}^{m} p_{ij} \qquad [2]$$

This is a rather common evaluation rule, very intuitive, and easy to calculate. It tells us how many points each team has accumulated during the competition. Yet, it may be regarded as too elementary in some cases because it disregards aspects that might be relevant. We shall refer here to two of those aspects: letting the points lost play a role in the evaluation (relative performance) and treating differently the outcomes of pairwise contests with teams of different importance (computing the strength of the teams). We do so by applying a double transformation to the original outcomes: (1) pondering the points obtained by the points lost, and (2) introducing the strength of the competing teams.

First transformation: relative outcomes

The first transformation refers to introducing the points lost in the evaluation. We do so by computing the points obtained in each confrontation relative to the points lost in the competition. To formalise this idea, consider the function $f^*(\mathbf{P}) = \mathbf{1}^T\mathbf{P}$, where $\mathbf{1}^T$ is now the row-unit vector in $\mathbb{R}^m$. This function can be regarded as the dual function of f^P, whose ith component is given by:

$$f_i^*(\mathbf{P}) = \sum_{j=1}^{m} p_{ji} \qquad [3]$$

Function f^* is a penalty function, as the term $f_i^*(\mathbf{P})$ corresponds to the total points lost by team i during the competition. We now define the ***relative outcome*** of team i concerning team j, as follows:

$$r_{ij} = \frac{p_{ij}}{f_i^*(\mathbf{P})}, \qquad \forall\, i \qquad [4]$$

That is, the relative outcome r_{ij} is simply the outcome of team i vis a vis j, pondered by i's penalty function.

Let us call $\mathbf{R} = \left(r_{ij}\right)_{i,j=1}^{m}$ to the matrix of relative outcomes. Note that applying the performance rule over the matrix of relative outcomes already defines an interesting evaluation protocol, that we may call the ***relative performance rule***, f^R. This rule is given by: $f^R(\mathbf{P}) = \mathbf{R1}$. That is,

$$f_i^R(\mathbf{P}) = \frac{f_i^P(\mathbf{P})}{f_i^*(\mathbf{P})} = \frac{\sum_{j=1}^{m} p_{ij}}{\sum_{j=1}^{m} p_{ji}} \qquad [5]$$

This rule associates with each team the ratio between total points won and total points lost. This is also a well-defined continuous function, provided $f_i^*(\mathbf{P}) > 0, \forall\, i$, with a clear meaning and an easy computation. It takes values greater than, equal to, or smaller than 1, depending on whether total points accrued are larger than, equal to, or smaller than those conceded. Using this rule, instead of the conventional one, will modify the incentives of the teams enhancing competitiveness, since winning a match not only gives points to the winning team but reduces the worth of the points obtained by the losing team (and losing a match reduces the worth of the scores already accumulated).

Second transformation: relative strength

We now introduce a second transformation, to consider the strength of the alternatives when valuing the outcomes of pairwise confrontations. Note that, as the only data in a problem refer to the elements in matrix $\mathbf{P}$, the strength of a team must necessarily be a function of the points obtained in those confrontations. Introducing this aspect in the evaluation, thus, requires some sort of fixpoint argument.

Given an evaluation problem $\mathbf{P}$, let $f(\mathbf{P}) \in \mathbb{R}_+^m$ denote a vector of teams' ratings provided by an evaluation rule *f*. We define the ***relative strength*** of team i under rule *f*, as the weighted sum of the relative scores accrued in the competition, where the weights correspond to the evaluations of those alternatives. That is,

$$S_i(\mathbf{P}, f) = \sum_{j \neq i} r_{ij} f_j(\mathbf{P}) \qquad [6]$$

A team is thus stronger when it has better relative outcomes and/or when it obtains those outcomes from teams with higher evaluations. How do evaluation and strength relate? The next principle provides the key element to closing the circle:

Relative strength proportionality: The evaluation of each team is proportional to its relative strength, that is,

$$f_i(\mathbf{P}) = \lambda \sum_{j=1}^{m} r_{ij} f_j(\mathbf{P}), \qquad \forall\, i$$

Relative strength proportionality corresponds to Keener's proportionality principle applied to the relative outcomes. From a formal viewpoint this property holds if and only if there exists a mapping $f^{RS}: \mathbb{R}_+^{m^2} \rightarrow \mathbb{R}_+^m$ such that,

$$f^{RS}(\mathbf{P}) = \lambda \mathbf{R} f^{RS}(\mathbf{P})$$

That is, a mapping that corresponds to a dominant eigenvector of matrix **R**. As for each problem **P** the associated matrix **R**, is a non-negative square matrix, we can apply the Perron-Frobenius Theorem (e.g. Berman & Plemmons, 1994), which ensures the existence of a positive dominant eigenvalue $\lambda \in \mathbb{R}_{++}$, and an associated eigenvector, $\boldsymbol{y} \in \mathbb{R}_+^m, \boldsymbol{y} \neq \mathbf{0}$. The eigenvectors associated with the dominant eigenvalues are the only ones that belong to $\mathbb{R}_+^m$ (i.e. they are vectors of real numbers with non-negative components). Therefore, there is a well-defined mapping $f^{RS}(\mathbf{P})$ that associates to each problem **P** the dominant eigenvector(s) of the corresponding matrix **R**, with $f_i^{RS}(\mathbf{P}) = \lambda \sum_{j=1}^m r_{ij} f_j^{RS}(\mathbf{P}), \forall\, i$, as required. The same Perron-Frobenius Theorem ensures that if matrix **P** is irreducible (hence the associated **R** matrix), the dominant eigenvector is unique and strictly positive, so that the mapping f^{RS} is a function. Moreover, the components of the unique dominant eigenvector are continuous functions of the elements of **P**.

There is, therefore, an evaluation mapping, that associates to each team a weighted sum of the relative outcomes obtained along with the competition, where the weights correspond to their evaluations. We call this mapping the ***relative strength rule***. This is a mapping homogeneous of degree zero, so that we can freely choose the units of measurement, Moreover, the evaluations can be systematically calculated using a familiar algorithm.

We can write more explicitly the evaluation of team *i* as follows:

$$f_i^{RS}(\mathbf{P}) = \frac{\sum_{j=1}^m p_{ij} f_j^{RS}(\mathbf{P})}{\sum_{j=1}^m p_{ji}}, \qquad \forall\, i \qquad [7]$$

It is interesting to note the precise relationship between the relative performance rule and the relative strength rule. Indeed, the relative strength rule can be regarded as the limit of a process that starts from the evaluation obtained by the relative performance rule, adjusting sequentially the evaluation by the strength of the teams. That is if we let $f^R(\mathbf{P}) = \mathbf{R1}$, then we have:

$$f^{R_1}(\mathbf{P}) = \mathbf{R} f^R(\mathbf{P})$$
$$f^{R_2}(\mathbf{P}) = \mathbf{R} f^{R_1}(\mathbf{P}) = \mathbf{R}^2 f^R(\mathbf{P})$$
$$f^{R_3}(\mathbf{P}) = \mathbf{R} f^{R_2}(\mathbf{P}) = \mathbf{R}^3 f^R(\mathbf{P})$$
$$\dots \quad \dots \quad \dots$$
$$f^{RS}(\mathbf{P}) = lim_{t\rightarrow\infty} \mathbf{R}^t f^R(\mathbf{P})$$

Remark 2: *We assume implicitly that* $\sum_{j=1}^{m} p_{ji} > 0$ *for all i. Note that* $\sum_{j=1}^{m} p_{ji} = 0$ *implies that team* i *does not concede a single point to any other, which means that it is a winner "hors catégorie" (we discuss later the existence of different categories of teams regarding the decomposability of the relative outcome matrix).*

One may argue that the relative strength rule cannot be used as a standard evaluation procedure in competitions, as the season proceeds, because the necessary information is only available once the competition ends. This is only partly true, as one can give this rule a dynamic format that gets updated as the competition unfolds, as with all standard classification procedures. Let us briefly address this point here.

Suppose that the competition consists of a single round, to simplify the discussion. Each team must play $(m-1)$ matches in a certain order, which implies that there must be $(m-1)$ "days" of competition and each of those "days" will involve $(m-1)$ pairwise encounters.

Before the competition begins, we take a reference matrix $\mathbf{P}^0$ given by:

$$p_{ij}^0 = \frac{m-1}{m-1}, \forall\, i \neq j,\;\; p_{ii}^0 = 0, \forall\, i$$

Needless to say, all teams get the same evaluation at this stage.

After the first day of the competition, we build a new matrix $\mathbf{P}^1$ as follows:

$$p_{ij}^1 = \begin{cases} p_{ij}, if\ i\ played\ against\ j \\ \dfrac{m-2}{m-1}, otherwise \end{cases} \qquad , \forall\, i \neq j$$

With $p_{ii}^1 = 0, \forall\, i.$ Applying the relative strength rule to this matrix provides an evaluation of the teams after the first day.

We proceed along this path as the season goes on. After t days of competition, we will have a new matrix, $\mathbf{P}^t$:

$$p_{ij}^t = \begin{cases} p_{ij}, if\ i\ played\ against\ j\ on\ day\ t \\ p_{ij}^{t-1}, if\ i\ had\ played\ against\ j\ in\ a\ former\ day \\ \dfrac{m-(t+1)}{m-1}, otherwise \end{cases} \qquad , \forall\, i \neq j$$

With $p_{ii}^t = 0, \forall\, i.$

After completing the matches of the last day of the competition, we shall have: $\mathbf{P}^{m-1} = \mathbf{P}$.

This construction permits obtaining an evaluation of the teams that gets updated as the competition develops. Note how the relative strength rule produces an evaluation in which the outcome of each new pairwise encounter affects the evaluation of *all* teams, as each new match modifies the outcome matrix $\mathbf{P}^t$ and hence the associated solution.

2.2 An application: The Premier League

We now apply those evaluation procedures to a real sports competition: The 2021-2022 English football league, *The Premier League*. The first division of the English football championship consists of 20 teams that compete twice against each other (i.e., two rounds) along the season. A team obtains 3 points when it beats another, 1 point in case of a draw, and 0 points otherwise. The competition's outcome corresponds to the sum of all those points over the season (i.e., the performance rule). This is, therefore, an evaluation in which the number of defeats is immaterial (except for breaking ties), and the number of points is independent of the strength of the competing teams.

We apply here the relative performance rule and the relative strength rule, to the 2021-2022 English football championship, to illustrate how these evaluations operate and how they differ from the official outcomes.

Table 1 provides the evaluations according to those three criteria. We normalize all those evaluations to a common maximum of 100 points for the winner. Note that the relative performance rule and the relative strength rule discriminate much more than the official competition, with coefficients of variation of 1.125 and 1.118, respectively, for 0.359 for the official results (coefficients of variation more than three times higher than those of the official results). There are 9 changes in the associated ranking of the relative performance rule, and 13 in the relative strength rule. Those changes include another winner of the competition, for both alternative evaluation rules: Liverpool, rather than Manchester City, would get the first position due to its better performance vis a vis the top tiers.

Table 1: Normalized points and rank of the Premier League 2021-22, according to the official scoring system (P), the relative performance rule (RP), and the relative strength rule (RS)

	Official classification		Relative Performance		Relative Strength	
Team	Rank	Points	Rank	Points	Rank	Points
Arsenal	5	74	5	25	5	22
Aston Villa	14	48	14	11	16	9
Brentfor	13	49	13	11	13	12
Brighton	9	55	8	16	9	18
Burnley	18	38	18	8	17	8
Chelsea	3	80	3	40	3	41
Crystal Pal	12	52	10	14	8	19
Everton	16	42	17	8	15	9
Leeds Un	17	41	16	9	18	6
Leicester City	8	56	9	15	10	16
Liverpool	2	99	1	100	1	100

Man City	1	100	2	94	2	98
Man Utd	6	62	6	19	7	20
Newcastle	11	53	12	14	14	11
Norwich	20	24	20	4	20	3
Southampton	15	43	15	10	11	13
Tottenham	4	76	4	28	4	39
Watford	19	25	19	4	19	3
West Ham	7	60	7	17	6	21
Wolverhampton	10	55	11	14	12	13

Regarding the implications of using those evaluation rules, we can think of the effect on incentives and on the market value of some of the teams' assets. In terms of incentives, one would expect that applying the relative performance rule or the relative strength rule will induce tougher competition. This is so because losing not only implies that the team does not add points but also that it is penalized. Moreover, in the case of the relative strength rule, weak teams will have a premium when beating top tiers so it is to be expected that they will fight harder (and strong teams would know it, which also induces a higher effort as the distance between teams will shorten in a convex way). Concerning market values, those evaluation protocols may provide a basis to define the range of some financial variables, from the worth of a team to the allocation of broadcasting rights, say.

3. Discussion

3.1 Considering losses and strength

We have presented an evaluation rule for pairwise contests that incorporates two relevant aspects in the rating of alternatives: points won relative to points lost, on the one hand, and the relative strength of the alternatives, on the other hand. Both aspects affect the incentives of the teams in sports competitions and tend to improve the discrimination power of the evaluation rule.

The idea of introducing losses in the evaluation can be related to the need of considering complete rankings of candidates in voting procedures, rather than computing only the number of first positions. The critique of plurality voting goes back to Borda and Condorcet, who proposed different alternatives to that evaluation rule. Indeed, the outcomes p_{ij} ma be interpreted as a form of Condorcet numbers, whereas $\sum_{j=1}^{m} p_{ji}$ can be regarded as a sort of anti-Borda count.

There is not a unique way of introducing the losses in the evaluation. The one offered here presents some interesting features. First, it makes each victory worthier and each defeat more harmful, as the outcomes of the confrontations affect simultaneously to the numerator and the denominator, that move in opposite directions. Second, it links each particular outcome with all the rest, so losing a match will reduce the evaluation of the points obtained in former confrontations. Third, it keeps the outcome variable in the positive orthant, which helps making comparisons. And fourth, being a round-robin tournament, there is no negative implication (i.e. loss of relevant information) regarding the degree zero homogeneity of those transformed variables. The example in Section 2.2 shows that introducing the losses in the evaluation changes both the distance between the teams' evaluations (points) and also the rankings.

The notion that the evaluation of an alternative relative to another must consider the strength of the competitor is a familiar principle in many evaluation problems. It expresses the principle that beating a top tier in a pairwise confrontation is more relevant than beating a low-level alternative. Think, for instance, of the evaluation of the impact of academic journals. If we use the bare citation impact, all citations are given the same relevance, irrespective of the position that the citing journal occupies in the ranking. A finer evaluation obtains when we ponder citations by the relevance of the citing journal (e.g. Pinsky & Narin, 1976, Liebowitz & Palmer, 1984, Palacios-Huerta & Volij, 2004, or Albarrán et al, 2017), as now implemented by the Eigenfactor (see http://www.eigenfactor.org). A similar principle is applied in the celebrated Google search engine (Brin & Page, 1998) and appears in many evaluation protocols regarding tournaments (Daniels, 1969, Moon & Pullman, 1970, Laslier, 1997, Saaty, 2003, Slutzski & Volij 2006, Boccard, 2020).

3.2 Outcome variables

Let us briefly address the question of the nature of the outcome variables. Note that most round-robin tournaments apply a lexicographic criterion to determine the competition results. That is, winning or losing the matches is the key criterion, with the specific results of the matches used only as a secondary principle, if at all (e.g. football, basketball). So, we typically find that winning a match gives the winner some points, a draw gives a smaller number of points (when draws are allowed) and losing gives the loser no point. All those points are added along with the competition and the final classification is determined by those sums, while the specific results achieved in the matches might be used to undo the ties.

Following this criterion, let us assume that the scores p_{ij} correspond to the result of applying a scheme by which the team that wins the match gets $\alpha > 0$ points, the team that losses the match gets 0 points, and both teams get β points, with $0 < \beta < \alpha$, when there is a draw. So here we have $p_{ij}(h) = \alpha$ with $p_{ji}(h) = 0$ if i beats j in round h, $p_{ij}(h) =$

0 with $p_{ji}(h) = \alpha$ when i is beaten by j, and $p_{ij}(h) = p_{ji}(h) = \beta$ in case of a draw. Common schemes of this type are: (1, ½, 0), as in Keener's baseline model, (3, 1, 0) as in The Premier League and most European football competitions, or (4, 2, 0) as in rugby contests.

Let w_{ij} denote the number of victories when *i* plays against *j*, in all the rounds, and d_{ij} the corresponding number of draws. Then, we can write $p_{ij} = \left(w_{ij}\alpha + d_{ij}\beta\right)$ and the relative performance rule adopts the form:

$$f_i^R(\mathbf{P}) = \frac{\sum_{j\neq i}\left(w_{ij}\alpha + d_{ij}\beta\right)}{\sum_{j\neq i}\left(w_{ji}\alpha + d_{ij}\beta\right)}\ , \qquad i,j = 1,2,\dots,m \qquad [1']$$

So that when $d_{ij} = 0$ for all $j \neq i$ (no draws), $f_i^R(\mathbf{P})$ is just the ratio between the number of victories and the number of defeats.

Similarly, the relative strength rule becomes now:

$$f_i^{RS}(\mathbf{P}) = \frac{\sum_{j\neq i}\left(w_{ij}\alpha + d_{ij}\beta\right)f_j^R(\mathbf{P})}{\sum_{j\neq i}\left(w_{ji}\alpha + d_{ij}\beta\right)}\ , \qquad i,j = 1,2,\dots,m \qquad [2']$$

In this case, when there are no draws, we get:

$$f_i^{RS}(\mathbf{P}) = \frac{\sum_{j\neq i} w_{ij}f_j^R(\mathbf{P})}{\sum_{j\neq i} w_{ji}}\ , \qquad i,j = 1,2,\dots,m$$

That is the number of victories weighted by the strength of the corresponding teams, over the number of defeats.

One can think of different ways of introducing the results of the matches in the evaluation (see Keener, 1993, Redmond, 2003, Dabadghao & Vaziri, 2021). A simple way of doing that, while preserving the lexicographic nature of victories, which somehow extends the scheme used in the Six Nations rugby contest,[2] is the following. Let s_i, s_j denote the "goals" obtained by teams *i* and *j* in a pairwise encounter. Then define:

$$p_{ij} = w_{ij}\left(\alpha + g\left(s_i, s_j\right)\right) + d_{ij}\beta$$

Where $g\left(s_i, s_j\right)$ is a function increasing in s_i and decreasing in s_j. An obvious example of this function is $g\left(s_i, s_j\right) = max\left\{0,\ \tau\left(s_i - s_j - k\right)\right\}$ for some scalars $\tau, k > 0$, that regulate the number of points that will be added, depending on the score difference. In the case of football competitions, for instance, we can think of $k = 1, \tau = 1/2$ so that the winning team gets an extra point when there is a difference of three goals, and proportionally

[2] The Six Nations rugby tournament is a yearly international competition involving the best European rugby teams: England, France, Ireland, Italy, Scotland, and Wales. It adopts the form of a league, that is, each team competes against all others in pairwise encounters. Since 2017 the scoring system for the matches is the following: a victory yields 4 points, a draw 2 points, and a defeat 0 points. There are also some bonus points, to make the competition livelier. There is an extra (offensive) point awarded when more than 4 essays are realized in the match, and another extra (defensive) point if the defeat is by 7 or less points.

more or less depending on that difference (with no extra point when there is a single goal of difference).[3]

3.3 The reducibility problem

Note that $p_{ij} = 0,\ \forall j$ implies $f_i^{RS}(\mathbf{P}) = 0$. Yet, the converse is not true, that is, some teams can be given a zero rating and still get $p_{ij} > 0$ for some *j*. This will happen when the matrix **R** is reducible, which implies the existence of different subsystems that are globally ranked (i.e., all teams in the top subsystem are better than those in a different subsystem because no team of the lower subsystem has been able to beat -or to score against- a team from the top one). The dominant eigenvector of a decomposable matrix of this type gives value zero to all teams that conform to the dominated subsystem. So, in the overall comparison, all teams in the inferior categories are equally valued and rated zero.

The reducibility of the outcome matrix has been regarded as problematic by many authors, as mentioned above because it implies making indistinguishable all those teams in the dominated subsystem and induces that winning or not be irrelevant for those teams regarding the final evaluation. Moreover, those procedures based on solving linear equation systems must ensure the existence of the inverse matrix to get unique solutions. Yet, in our view, reducibility is a structural property of the system that provides relevant information on the competition, which can be easily handled. Indeed, when we have a reducible matrix, we can apply the same evaluation procedure to a dominated subsystem, considered in isolation, and obtain the evaluation of those teams *within* that group. The teams of an inferior league will typically get non-zero ratings, as we may find values $p_{ij} > 0$ for some *j* within its category.

When there is a subset of dominated teams, therefore, we can proceed to rate them in two different groups. First, those teams that in the joint evaluation appear with positive ratings, which would define the top class. Then, the teams that appear initially with zeroes in the dominant eigenvector of **R**, which would define the bottom class, in the understanding that all teams in the top class are regarded as better than anyone in the bottom class.[4]

An extreme case of reducibility appears when $\sum_{j=1}^{m} p_{ji} = 0$, which will be observed when team *i* beats all other teams in all their pairwise confrontations. In this case, neither the relative performance rule nor the relative strength rule are defined for *i* and this indicates that there is a single team that is above all others (a strong Condorcet

[3] The official webpage of the Premier League provides information on the aggregate goals difference of each team, as a complementary data of interest.

[4] It may be that the subset of alternatives that are not in the top class define in turn a decomposable submatrix, which indicates that there are more than two "categories".

winner) and stands on its own in a different category. So, it is not meaningful to attach any relative value to that team, or one can say that its relative value is +∞, as a way of expressing the notion of *hors catégorie*. The evaluation can be applied to the remaining teams, in the understanding that they all are below *i*.

Note that many sports competitions define different divisions so that only the teams that belong to the same division play against each other throughout the season. A specific protocol determines how to move up and down between those divisions. This setting can be regarded as an application of the notions of computing the strength of the competitors and using a form of reducibility. In this case the notion of strength is introduced implicitly in a very elementary way, which determines that only teams with similar strengths (those in the same division) play against each other. The configuration of the different divisions can also be regarded as the application of a form of reducibility, since all teams playing in a higher division are regarded as better than those playing in a lower one, no matter their relative outcomes.

3.4 Comparison with other rules

As already mentioned, the relative strength rule can be regarded as the application of Keener's formula to the matrix of the relative scores, **R**, rather than to the matrix of absolute scores, **P**. Therefore, we can interpret that what our evaluation rule does is incorporate the role of the losses to Keener's strength proportionality principle.

One may wonder whether the relative performance rule, f^R, and the relative strength rule, f^{RS}, satisfy the three criteria proposed by Vaziri *et al* (2018), referred to in the Introduction. In the context of symmetric round-robin tournaments, both rules satisfy fairness (i.e., they are independent of the sequence of the matches). Moreover, the relative strength rule does satisfy comprehensiveness (i.e., it is sensitive to the strength of the competitors), whereas the relative performance rule does not. Finally, the relative performance rule satisfies monotonicity whereas the relative strength rule only satisfies weak monotonicity. That is, in both cases, we can say that no team will find it beneficial to lose a match or to score fewer points, and in the case of the relative performance rule there will always be incentives to win.[5]

There are some common features between the relative strength rule and the Elo rating system, worth mentioning. Elo's rating is a dynamic process that keeps changing the teams' evaluations as they compete against each other. The key idea is that if team *i* performs as expected against team *j*, it gains nothing, whereas if it performs better (resp. worse) than expected, it is rewarded (resp. penalized). This principle induces an

[5] This weak monotonicity property is present in many real competitions when there are teams that before the end of the season have already lost their category but still have to play some matches.

adjustment process in which ratings are modified after each match, proportionally to the difference between the actual and the expected score. That is, when team *i* confronts team *j* at time t, the outcome will affect *i*'s rating as follows:

$$r_i^t - r_i^{t-1} = K\left(s_{ij} - E(s_{ij})\right)$$

where s_{ij} is the outcome of the match (typically 1 for a victory, 0 for a defeat, and ½ for a draw), and $E\left(s_{ij}\right)$ the expected outcome, which is, in turn, a function of the teams' rankings.[6] Note that the adjustment process depends on two constants that are to be determined from the outside. Also observe that none of the evaluations of the rest of the teams will be affected by the outcome of the match between *i* and *j* (hence, avoiding the competition may be a dominant strategy for both teams in some cases).

We have already mentioned that the relative strength rule can be given a sequential format, which facilitates the comparison with the Elo's rule. From this perspective, the relative strength rule exhibits an updating procedure such that the outcome of each new pairwise encounter affects the evaluation of *all* teams, as each new match modifies the outcome matrix $\mathbf{P}^t$ and hence the associated solution. As for the way of computing wins and losses, note that relative scores can be interpreted as follows: in each pairwise confrontation, we credit team i with $1/m$ points when the corresponding score equals the average points lost in the competition, and credit values proportionally larger or smaller than $1/m$ depending on whether p_{ij} exceeds or falls short of that average.

4. Final remarks

We have considered here a type of evaluation problems based on making pairwise comparisons of alternatives and presented an evaluation rule that considers both wins and losses and the strength of the alternatives in the evaluation. Sports competitions is a particularly clear setting to discuss this evaluation rule, but the criterion can also be applied to more general contexts. For instance, if we are comparing alternatives in terms of the ordinal evaluations provided by a collection of judges, then the terms p_{ij} may be associated with the number of judges who prefer alternative *i* to alternative *j* (plus ½ of those who are indifferent, if we admit weak orderings). This actually corresponds to an extension of the Borda-Condorcet rule (Herrero & Villar, 2021), dispensing with the transitivity of the judges' evaluations.

---

[6] The expected score is calculated by (where ζ is a constant):

$$E\left(s_{ij}\right) = \frac{10^{r_i/\zeta}}{10^{r_i/\zeta} + 10^{r_j/\zeta}}$$

The relative strength rule is a meaningful evaluation criterion that can be computed using standard algorithms and can also be applied sequentially. Moreover, this is a rule that enhances competitiveness in sports and discriminates more between the alternatives than the standard methods. This may be relevant for several reasons: To have a better picture of alternatives that may appear close together; to provide different shadow prices for the alternatives, depending on the features regarded as relevant; to define a range of values for those alternatives, that provide a robustness check of their evaluation; or to provide incentives to foster competition between contenders.

Let us conclude by noting that there are always concerns about the possibility of manipulating the results of sports competitions. Recently, the United Nations Office of Drugs and Crime, together with the International Olympic Committee, published a guide to tackling such manipulation (UNODC, 2021). Different rules exhibit different degrees of manipulability, in the sense that simpler rules facilitate computing what to do to get a given result. For instance, it has been proved that by using the performance rule, manipulation can be computed in polynomial time (Russell & Walsh, 2009). Increasing the complexity of the rules may help reduce manipulation, as it is more difficult to anticipate the precise implications of individual actions on the evaluation (Russell & van Beek, 2012). In this respect, therefore, introducing losses and strength in the evaluation will render manipulation more difficult.

REFERENCES

Albarrán. P., Herrero, C., Ruiz-Castillo, J, Villar, A. (2017). The Herrero-Villar approach to citation impact. **Journal of Informetrics**, 11: 625-640.

Berman, A. & Plemmons, R. J. (1994), **Nonnegative matrices in the mathematical sciences.** Society for Industrial and Applied Mathematics.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. **Computer Networks and ISND systems,** 30: 107-117

Boccard, N. (2020), Pairwise Comparison Ranking. The Landau Method, mimeo, University of Girona.

Colley, W.N. (2002), *Colley's Bias Free College Football Ranking Method: The Colley Matrix Explained* [online] http://www.colleyrankings.com/matrate.pdf .

Daniels, H. E. (1969), Round-robin Tournament Scores, **Biometrika**, 56: 295-299.

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

Fishburn, P.C. (1977), Condorcet Social Choice Functions, **SIAM Journal on Applied Mathematics**, 33(3): 10.1137. https://doi.org/10.1137/0133030

Govan AY (2008) Ranking theory with application to popular sports. **ProQuest**

Govan AY, Langville AN, Meyer CD (2009) Offense-defense approach to ranking team sports, **Journal of Quantitative Analysis in Sports**, 5(1). https://doi.org/10.2202/1559-0410.1151

Herrero, C. & Villar, A. (2021), Group decisions from individual rankings: The Borda-Condorcet rule, **European Journal of Operational Research**, 291 : 757-765.

Keener, JP. (1993), The Perron-Frobenius Theorem and the ranking of football teams, **SIAM Review**, 35 (1): 80-93 .https://doi.org/10.1137/1035004

Laslier, J. (1997), **Tournament solutions and majority voting**. Springer, Berlin, Heidelberg, New York.

Liebowitz, S. J., Palmer, J. P. (1984), Assessing the relative impacts of economics journals, **Journal of Economic Literature**, 22 : 77- 88.

Massey, K. (1997), **Statistical Models Applied to the Rating of Sports Teams**, Bachelor's thesis, Bluefield College, Bluefield, VA.

Moon, J.W., N.J. Pullman, (1970), On Generalized Tournament Matrices, **SIAM Review**, 12(3): 384-399.

Palacios-Huerta, I., O. Volij (2004), The Measurement of Intellectual Influence, **Econometrica**, 72(3): 963-977.

Pinsky, G., F. Narin (1976), Citation Influence for Journal Aggregates of Scientific Publications: Theory, with applications to Physics, **Information Processing, and Management**, 12: 297-312.

Russell, T., & van Beek, P. (2012). Detecting Manipulation in Cup and Round Robin Sports Competitions. *2012 IEEEE 24th International Conference on Tools with Artificial Intelligence*, (pp. 285-290).

Russell, T., & Walsh, T. (2009). Manipulating tournaments in cup and round robin competitions. **First International Conference on Algorithmic Decision Theory**, (pp. 26-37).

Saaty, T.H. (2003), Decision-making with the AHP: Why is the principal eigenvector necessary, **European Journal of Operational Research**, 145: 85-91.

Slutzski, G, O. Volij, (2006), Scoring of web pages and tournaments: axiomatizations, **Social Choice and Welfare**, 26 (1): 75-92.

UNODC. (2021). **Legal approaches to takling the manipulation of sports competition: a resource guide**. Viena: United Nations.

Vaziri, B., Dabadghao, S., Yih, Y., & Morin, T. L. (2018), Properties of sports ranking methods, **Journal of the Operational Research Society**, 69(5), 776-787. https://doi.org/10.1057/s41274-017-0266-8

Vaziri B, Yih Y, Morin T (2018) A proposed voting scheme to reduce the sensitivity of the Markov method, **International Journal of Operations Research,** 32(1):24–40