14TH ANNUAL
MIT SLOAN
SPORTS ANALYTICS CONFERENCE
MARCH 6 - 7, 2020 | BOSTON, MA
presented by ESPN

Fair and Efficient Ranking in Incomplete Tournaments

November 29, 2022

Paper Track: Football

Paper ID: 171177

Abstract

I discuss basic desirable fairness standards for the case of incomplete tournaments. I present a parsimonious family of scoring methods that uniquely satisfies these standards. It includes the win percentage method as a special case. I analyze this family of scoring methods in terms of efficiency, defined as how close a scoring method comes to capturing what the teams' win percentages would have been, in a complete tournament. I show that efficient scoring methods are typically unfair. Finally, using data on betting odds, I calibrate the family of scoring methods to match, as closely as possible, the actual rankings that were used to determine the teams that would go on to compete for the championship of the NCAA division 1 football tournament between 2011 and 2017. I find that the rankings used by the NCAA were generally efficient and unfair, and I quantify the biases present in each year's ranking.

Keywords: Incomplete paired comparisons, tournament ranking, scoring methods.

42 ANALYTICS

1. Introduction

In May of 2016, against all odds (5,000 to 1 to be precise), Leicester City won the English Premier League football title. It did so by winning 23 games, drawing 12 and losing only 3, despite fielding a roster of players whose aggregate wage bill was the fifth-lowest in the 20-team league. The team that came in second had won 20, drawn 11 and lost 7. The title was uncontroversial. Just a year later another underdog (this time 1,000 to 1 odds), the University of Central Florida, did *not* win the NCAA division 1 (American) football title despite winning all 13 of its games in a league of 130 teams. The team that was awarded the title had won 13 games and lost 1. The title remains highly controversial. There are several factors that both fuel the controversy surrounding the outcome in the US while justifying the outcome in England. However, the most glaring one is given by the very basic structures of the two tournaments: In the first one there are 20 teams that play 38 games each (twice against every opponent) and in the second one there are 130 teams that play less than 15 games each (and teams don't even play the same number of games). The first is a case of a twice-complete tournament and the second a case of an incomplete tournament. The fundamental question is straightforward: How to establish a final ranking of the teams that play in a given tournament in a fair and reasonable way. Intuitively, it seems like a very simple task to do this for the first case but not necessarily to do so for the second.

In this work, I present basic criteria of fairness and efficiency for developing scoring methods that, through rewards and punishments for winning and losing, assign each team a final score which is then used to establish the final ranking of teams. I also present a particular family of scoring methods that is intuitively simple and satisfies the fairness and consistency criteria. It is then calibrated to match (as closely as possible) the actual rankings used by the NCAA (between 2011 and 2017) in order to quantify, among other things, how fair, efficient and/or biased these rankings were.

1.1 A Non-Trivial Scoring Problem

To be clear, a game is defined as a contest between two teams and the objective for any given team is to beat its opponent. That is, when one team wins, the other team necessarily loses and if no team wins then no team loses, which (if allowed) defines a draw. There is no added information that will be used to score teams, that is, there will be no way in which a win can be qualified as better or worse other than from knowing which team won and which team lost. Interestingly there seems to be overwhelming consensus across different leagues of different sports and competitions that, despite there being multiple ways of further qualifying a given win (goals or points difference, judges scores, speed of victory, etc), none of these qualifiers should be used other than to break a tie in win percentage at the end of a complete tournament. In other words, the way in which a win is secured does not matter. At the end of the match, one team walks away with the win and the other with a loss. Whether this is done to give appropriate closure to a match or to avoid teams running up scores against weak opponents or simply as a way to discourage cheating is irrelevant.

The operating assumption in this work is that the only information that can be used to score teams is which teams played against each other and who won each game. The results of all the games played in a given tournament of n teams can be summarized by an n × n matrix **W**, labeled the win matrix and also referred to as the matrix of tournament results, where any entry w_{ij} represents the total number of times i beat j. Thus, any game between any two teams i and j that has been played in the tournament gets recorded either as a win by i (adding 1 to w_{ij}) or a win by j (adding 1 to w_{ji}). A matrix that records the games played by each team but not the results of those games is referred to as the games matrix and is

defined as $\mathbf{G} \equiv \mathbf{W} + \mathbf{W}^T$ so that every entry g_{ij} shows the number of times that team i plays against team j. This matrix will also be referred to as the tournament fixture in reference to a tournament that has not been played yet.
Since teams must be scored using only the win matrix as a source of information then a win matrix **W** can also be interpreted as a scoring problem. A scoring method is a multivariate function $M_n(\cdot)$ that assigns any n × n scoring problem an n × 1 vector of scores $\boldsymbol{v}_n$.

1.2. A Benchmark Scoring Method

In a complete tournament all teams play each other once. As a result, there is no advantage by any team over any other in terms of the quality of opposition faced. For this reason, the scoring method that has been used in all types of competitions where a complete tournament is played is a system of points where all wins count equally and all losses count equally, but less than a win. The final scores are simply given by the sum of all points assigned to each team. Additionally, if, for example, we want to compare the scoring of two different complete tournaments where the number of teams is different, then a simple solution is to use points per game as a measure instead. This is also known as (or isomorphic to) the win/loss record, or, more precisely, the win percentage.
The simplicity of the win percentage scoring method is its main strength. When thinking about creating a scoring method that is reasonable, our first instinct is to reward winners and punish losers. Because we want to be unbiased and fair, we want our reward for a win to be the same for any team that wins and our punishment for a loss to be the same for any team that loses. Adding up rewards and punishments gives us a very natural way to compare different teams that played the same amount of games. Normalizing by the number of games played is just a minor adjustment that is reasonable when teams have not played the same amount of games.
However, a key to feeling comfortable with this simple method is that at the end of the tournament, all teams will have faced the same competition. Thus, if one team faces weak competition early on and as a result ranks high on the scoring table, we feel at ease because the remaining schedule will either pull it back down where it belongs or prove that this team is at least as good as it's current position suggests. In other words, our approval of the win percentage scoring method is directly tied to the notion that a complete tournament is itself a fair type of tournament.
When moving away from a complete tournament, the win percentage scoring method loses its appeal. The tournament itself is no longer fair. A team that faced weak competition got an unfair advantage and a team that faced tough competition received an unfair disadvantage so, naturally, the win percentage will provide a biased measure of performance. The problem becomes compounded by the fact that the strength of the competition is itself an unknown variable that must be obtained using the same information (the results of all the games played) as the scoring of each team.
With the understanding that when it comes to incomplete tournaments, the very nature of such tournaments is biased (and therefore unfair), the set of fairness axioms studied here seeks to correct the bias as much as possible, while preserving the very simple structure of additive rewards and punishments. In other words, I will discuss methods that assign points to teams for wins and losses, but these points are not necessarily confined to being 1 for every win and 0 for every loss.

1.3. Points-per-game framework:

We want the scoring method to be based on the sum of points received as a result of each win or loss. For this we define a scoring method to be a points-per-game method if it can be expressed through a points system where the points assigned to any team *i* can be decomposed as the following sum:

$$p_i(W) = \sum_{j=1}^{n} [w_{ij} F_{ij}(W) + w_{ji} G_{ij}(W)]$$

where F_{ij} and G_{ij} are functions that assign a number to any scoring problem **W**, with F_{ij} representing the points assigned to i for every win against j and G_{ij} the points assigned to i for every loss against j. The score v_i assigned to team i is simply its assigned points p_i divided by the number of games played g_i.

2. Fairness Axioms:

Our first fairness axiom is anonymity. It requires scoring methods to survive re-labeling of teams and it is discussed at length in the literature. The definition of surviving a re-labeling is very straightforward: Re-labeling team i as team j and vice-versa (plus appropriately changing the win matrix to accommodate this re-labeling of teams) should always result in the exact same scoring of all teams (including i and j but where the new v_i would equal the old v_j and vice-versa). Otherwise the scoring method is fundamentally unfair and of no practical use. Applied to the points-per-game framework, we require both F_{ij} and G_{ij} to survive the relabeling of teams. This will guarantee that the scores do so as well.

Axiom 1 *(Anonymity)***:** Functions F_{ij} and G_{ij} survive the re-labeling of teams.

Absent any other information to be used, it would not be fair to assign two different teams a different amount of points for beating the same opponent. For the same reason it would not be fair to assign two different teams a different amount of points for losing to the same opponent. Thus, the second fairness axiom is:

Axiom 2 *(Win/loss fairness)*: A win against opponent j is assigned the same points to any team that beat j and a loss against opponent j is assigned the same points to any team that lost to j.

Allowing wins and losses to be assigned different points requires some caution: If we want to interpret a win as a positive signal (in the sense of being superior to the opponent) and a loss as a negative signal (inferior to the opponent) then the former should result in more points assigned than the latter. Thus, our third fairness axiom is that any win by any team should be assigned more points than any loss by any team.

Axiom 3 *(Win Dominance)***:** Any win by any team is assigned more points than any loss by any team.

This criterion will be key to avoiding nonsensical scoring outcomes like a team that loses all its games (presumably to very strong opponents) being scored higher than a team that wins all its games (to weak opponents nonetheless). There is also a very practical reason for this type of criterion to be applied: If teams are allowed to influence the schedule of opponents they will face, then without win dominance there would be no incentives to schedule an a-priori very weak opponent against the possibility of scheduling a very tough one that would guarantee a higher score regardless of the result of the game between them. Notice that the win percentage method trivially satisfies the above criteria, but the criteria leave the door open to assigning different points for beating (or losing to) different opponents. This will be crucial to a scoring method that is applied to incomplete

tournaments, which requires qualifying a win and/or loss through the only sensible way: The strength of its opposition.
Of course, we want a scoring method to give higher rewards for beating better opponents. If, instead, a victory against a weak opponent were to be assigned more points than a victory against a stronger opponent then the scoring method would be giving an unfair advantage to teams playing against the weaker opponent with respect to those playing against the stronger one. The same can be said about losses to better opponents as compared to losses against weaker ones. In other words, we would like the scoring method to assign points for victories that are non-decreasing in opponent strength and to assign points for losses that are also non-decreasing in opponent strength. Notice that a reasonable measure of the strength of an opponent is given by the opponent's score, after all, the scoring method is designed to assign scores that reflect the relative importance of teams in order to rank them from best to worst. Therefore, we can naturally use the scores as measures of strength. As a result, the strength of an opponent that is used to define this axiom is endogenous to the scoring method that the axiom is being applied to. This concept is referred to in the literature as *self-consistency*. Thus, our fourth fairness axiom is:

Axiom 4 *(Self-consistent win/loss fairness)***:** The points assigned for victories and losses are non-decreasing in the opponent's score.

It was argued in the introduction that the win-percentage method is the benchmark for fairness when the tournament is complete. Thus, we don't want to use a scoring method that, in complete tournaments, delivers a different ranking of teams than the one resulting from using win percentages. The scores vector and the win percentage vector don't have to match in order to deliver the same ranking, only the implied ordering of teams have to. In other words, we would like the scoring method to deliver scores that are increasing in win percentages whenever the tournament is complete. The reason for this is that, other than playing each other, in a complete tournament any two teams face the same opposition so a higher score should be consistent with a higher win percentage. Taking this idea to the context of any incomplete tournament, a slightly stricter version of this property, labeled homogenous treatment of victories, requires win percentages to determine which team is scored higher whenever any two teams face the same opponents, regardless of whether the tournament is complete or not. Thus, define the schedules of two teams as equivalent if they include the same opponents, possibly including each other as opponents as well. Then our fifth axiom states the following:

Axiom 5 *(Homogenous treatment of victories)***:** Scores are increasing in win percentages when teams play equivalent schedules.

The idea of anchoring the scoring method to be consistent with win percentages under appropriate circumstances (in this case equivalent schedules) can be taken one step further to include the opponents of the opponents of two given teams. We can consider the case of two teams that played different opponents but for every opponent of the first team we can find an opponent of the second team that played an equivalent schedule (with the exception of the two original teams because every team is itself an opponent of any opponent). In this case, we would say that the two original teams have equivalent second-order schedules. Of course, the opponents of the first team could, in principle, be much stronger than those of the second team by virtue of having better records against their opponents. As a result, it would not be fair to require the score of the second team to be higher than the score of the first team by virtue of having a higher win percentage alone. Thus, let us further assume that the average win percentage of all the opponents of the first

team is the same as that of the second team's opponents. Now, at least on average, the opponents of the first team are as strong as those of the second team. In this case axiom six would require the team with the higher win percentage to be scored higher.

Axiom 6 *(Homogenous treatment of second-order victories)***:** Scores are increasing in win percentages when teams play equivalent second-order schedules and the average win-percentage of its opponents is the same.

In the following section, I present a family of scoring methods that uniquely satisfies the above fairness axioms.

3. The generalized win percentage method family:

The generalized win percentage *(GWP)* family of scoring methods assigns scores according to a system of *n* equation and *n* unknowns, where for any team *i*, its final score v_i is given by:

$$v_i = \alpha \sum_{j=1}^{n} \frac{w_{ij}}{g_i} + (1 - \alpha) \sum_{j=1}^{n} \frac{g_{ij}}{g_i} v_j ,$$

where g_{ij} is the number of games played between *i* and *j*, g_i is the total number of games played by *i* and α is a number between zero and one. Thus, a team's score is a weighted average α of it's win percentage and (1- α) of its strength of schedule, where the strength of schedule is defined as the weighted average score of all opponents with weights defined by the percentage of all games by *i* that were played against a given opponent *j*. This recursive formulation results in a unique and finite score v_i for any team *i.* Moreover, this family of scoring methods, which is governed by the parameter α, satisfies all six fairness axioms when $\alpha \geq 1/2$.

Theorem: The family of GWP methods uniquely satisfies axioms 1 to 6 if $\alpha \geq 1/2$**.**

What this theorem shows is not only that the family of GWP methods satisfies the fairness axioms, it also shows that no other method satisfies all six axioms at once. In other words, if we use a different method to score teams, it will necessarily run counter to at least one of the fairness axioms. The proof of this theorem can be accessed through a more formal version of this paper.
Expressing the GWP method as a points-per-game method also provides us with a very intuitive interpretation of its underlying structure: For every game played between team *i* and team *j*, whenever i beats j the following points are assigned:

Points assigned to winning team $i = \alpha \times 1 + (1 - \alpha) \times v_j$

Points assigned to losing team $j = \alpha \times 0 + (1 - \alpha) \times v_i$

That is, the winning team receives a weighted average between 1 and the score of the losing team and the losing team receives a weighted average between zero and the score of the winning team. Notice as well that for $\alpha = 1$, this method assigns one point per win and zero per loss, so it collapses to the simple win-percentage method.
This intuitive way of expressing the GWP method is useful for finding the sufficient

condition on the possible values of α, namely that $\alpha \geq 1/2$. From the above, we know that losing to a team of score $v_i = 1$ would award the loser $(1-\alpha)$ points whereas beating a team of score $v_j = 0$ would award the victor α points. Since no team can ever achieve a score higher than 1 or lower than 0, it follows that we must have $\alpha \geq 1/2$ to guarantee that axiom 3 (win dominance) is satisfied.

In the following section, the question of a lower bound for α is further discussed. First, to show that if we want this criterion to be satisfied for all tournaments (and regardless of the number of teams) then 1/2 is the appropriate lower bound. In other words, $\alpha \geq 1/2$ is both a sufficient and a necessary condition. I will refer to this sub-set of GWP methods as *globally fair*. Nevertheless, it is easy to show that when the number of teams in a given tournament is low, the corresponding lower bound is not as high. Thus, global fairness can be too restrictive for specific applications (for example, in a two or in a three team tournament, the lower bound is zero). Unfortunately, when the number of teams is high, the total number of possible combinations of results grows exponentially and calculating the actual lower bounds on α very quickly becomes impossible. Instead, the analysis in section 3 turns the question around by making explicit a practical way of ruling out values of α that do not satisfy this criterion.

4. Win dominance and the game matrix:

We know that the globally fair lower bound for α that guarantees that every win will award the victor more points than any loss will award the loser is 1/2. It relies on assuming that it is possible to have a team with a score of 1 and another with a score of 0. With a sufficient number of teams *n*, the maximum possible score for a team does indeed approach 1 and the minimum possible score for a team approaches 0. Thus, $\alpha = 1/2$ is the appropriate lower bound if the intention is to satisfy win dominance for any number of teams *n* that each play any number of games *g*.

However, even for fairly high values of *n*, a sufficiently low number of games played by each team *g* leads to much lower bounds on α. This is because teams cannot achieve scores close enough to 0 and 1 even under very favorable circumstances. For example, win dominance will typically not be satisfied for values of α lower than 0.25 (all that is required is 12 teams playing 5 games each) and for applications where the number of teams is not much higher (for example, more than 25 teams playing 5 games each) a value of α that is greater than 0.35 will be required. For the specific case of the NCAA College football tournament, we have about 130 teams playing between 10 and 12 games each (in the regular season). In this case, the lower bound for α is 0.41. I label this lower bound the *Ex-ante* fairness bound, because it is the lowest possible value of α that guarantees fairness *before* the games are played. Of course, this bound could be lowered even further if we compute it *after* the games are played and we simply check that win dominance is satisfied *ex-post*.

5. Obtaining the most efficient α:

A natural way of addressing the question of which α to use is by comparing the rankings that result from different values of α in the incomplete tournament to the ranking that results from the win percentages had the tournament been complete. One way to implement this is to normalize the scores for any α so that they are centered around 1/2

and have a similar standard deviation as that of the win percentages, regardless of α. We can then (through simulations) directly compare normalized scores to win percentages instead of indirectly comparing the rankings that each method generates.

5.1. Simulation Strategy

In this section I proceed as follows: Start by simulating a complete tournament and computing the corresponding win percentages. These will be referred to as the *true* win percentages. Next, generate a number of incomplete tournaments using randomly generated incomplete game matrices, but populating the corresponding win matrices with the results from the complete tournament win matrix. Then, for every possible value of α and each win matrix, obtain the corresponding normalized scores. Finally, assess how efficient a given value of α is at approximating the (true) win percentages of the complete tournament. We do this by calculating the sum, across all n teams, of the squared differences between the win percentage of the complete tournament and the average of the normalized scores of the incomplete tournaments (referred to as the expected normalized score of the incomplete tournaments). That is:

$$SS = \sum_{j=1}^{n} [\hat{w}_i - E(\tilde{v}_i)]^2$$

where the first term is the win percentage of team *i* and the second is its expected normalized score. The sum of squared differences is simply a natural way of evaluating the goodness of fit. Also, if we divide the sum of squares by n and take the square root, we can interpret it as the standard deviation from the true win percentage. Finally, it is important to note that the expected normalized scores of the incomplete tournament will depend on the way in which we randomly create the incomplete tournaments. This last step is done through Monte-Carlo simulations because, other than for very simplistic randomization methods, it is practically impossible to obtain expected scores theoretically. Thus, the most efficient α will depend on the following: The number of teams *n*, the results of the complete tournament **W**, the number of games played by each team *g* in the incomplete tournament (in the simulations all teams play the same amount of games) and the random process chosen to assign games in the incomplete tournaments. In the Monte-Carlo simulations, this random process is governed by parameter ρ, related to how likely it is that teams of similar strength play against each other in the incomplete tournament. It is loosely labeled the correlation parameter because a value of ρ = 1 represents an incomplete tournament where teams almost exclusively face other teams of similar complete-tournament win percentage, a value of ρ = 0 represents a uniformly random chance of playing different opponents and ρ = −1 represents an incomplete tournament where teams almost exclusively face opponents that have opposite complete-tournament win percentages.

5.2. Simulation Results:

A robust feature of the Monte-Carlo simulations is that they produce sums of squared differences that are U-shaped in α. Thus, a most efficient α^* exists and is unique, for a given $(n, \boldsymbol{W}, g, \rho)$. Moreover, the simulations show that for a given (n, g, ρ), α^* is fairly constant in W as long as the standard deviation of the win percentages σ_W is constant. For example, Figure 1 shows 15 different sets of 200 simulations. Each set of simulations has $n = 130$, $g = 11$, $\rho = 0$ and a unique complete-tournament win matrix **W** that shares an almost identical standard deviation of win percentages σ_W with that of the other sets of simulations.

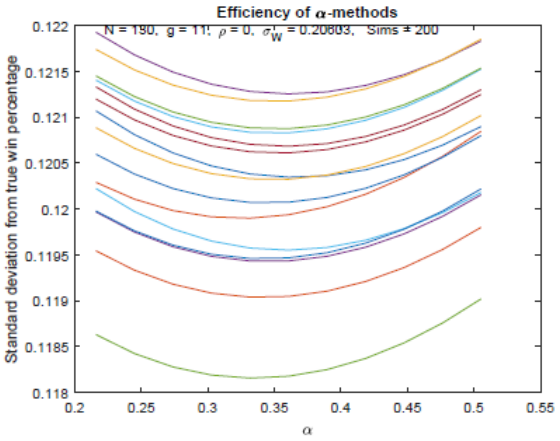

Figure 1: Efficiency

In each individual simulation within a set, a new incomplete tournament $\mathbf{G}_T$ is generated (and its corresponding win matrix $\mathbf{W}_T$ populated using win matrix $\mathbf{W}$) by selecting g games to be played by each team with the random pairing of teams governed by parameter ρ. It is clear upon inspection that, each time, α^* is close to 0.35 (the average α^* over the 15 sets of simulations is, more precisely, 0.3473).

The simulations allow us to perform comparative statics on the main parameters of the incomplete tournament. Each of the following graphs in Figure 2 shows what the most efficient α is as a function of the selected parameter, given benchmark values of the other ones. More precisely, a point along the full line represents, for the corresponding parameter value, the average over all 15 sets of 200 simulations of the most efficient α within each set. The dashed lines display +/- two standard deviations from the average and the point marked with dotted lines represents the benchmark value of the given parameter.

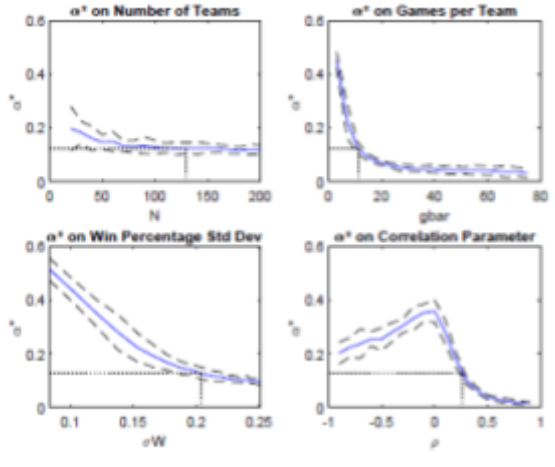

Figure 2: Comparative Statics

The benchmark values are n = 130, g = 11, ρ = 0.265 and σ_W = 0.204. They correspond to the average values when calibrating to the NCAA football tournament, as explained in the applied section.

The first graph shows how changing the number of teams n affects the most efficient α. It is

clear upon inspection that having more teams has almost no impact on the most efficient α (with the exception of having very few teams). The second graph shows how changing the number of games played by each team affects the most efficient α. It is decreasing in the number of games played because if teams play very few games then the strength of schedule itself is not very useful but as we increase the number of games played it becomes more reliable, so the true win percentage of a team is better predicted by giving the strength of schedule a higher weight. The third graph shows how changing the win percentage standard deviation affects the most efficient α. The bigger the difference between the stronger teams and the weaker teams the higher the win percentage standard deviation will be. This enhances the importance of the strength of schedule in correcting the bias that results from some strong teams playing other strong teams and weak ones playing other weak ones. Thus, the most efficient α declines with a higher win percentage standard deviation. Finally, and most importantly, the fourth graph shows how changing the correlation parameter affects the most efficient α. High and low values of ρ result in an α^* that is considerably lower than when $\rho = 0$. Intuitively, this is because when $\rho = 0$ opponents are drawn uniformly randomly, so this is when correcting for strength of schedule becomes least useful. But as the correlation increases in either direction (towards drawing teams of more similar true win percentages or towards drawing teams of less similar true win percentages) the usefulness of the strength of schedule adjustment increases, hence $\alpha*$ decreases.

One takeaway from these results is that it is critical to understand the way in which the incomplete tournament is created in order to determine which α is most efficient. It varies widely depending on how many games each team plays, how evenly matched the teams are, and the way in which games between teams are randomly assigned. Thus, it is important to have a grasp on these factors before deciding on what α to use under this efficiency metric. However, the most important takeaway from these results is that there is typically a conflict between what is fair and what is most efficient. Fairness dictates that we look at values of α that are mostly above 0.35 but maximum efficiency rarely occurs at those levels of α. Thus, if we want to maintain fairness while being as efficient as possible it becomes important to determine the minimum fair α because it will most likely be the most efficient one. Specifically, for the case for the College Football, the lower bound for fairness is 0.41 whereas efficiency is achieved at levels of α closer to 0.15.

6. Application: College Football Rankings

Every year between 2011 and 2017 there have been over 120 teams playing in the upper division of college football. They play between 11 and 14 games each season, depending on their success on the field. In order to advance to a bowl game, a team must finish with a non-losing record. In order to play for the championship (semi-finals and a final in the College football playoff era and just a final during the BCS era) a team must be selected to participate. We are currently in the College Football playoff era, where a committee selects the final four teams that will compete for the national championship. Prior to that, during the BCS era the Associated Press, Coaches and a set of ranking algorithms were weighted in order to determine the two teams that would play for the title.

Whether explicitly or not, all these rankings took into account (among other things) the win percentage and the strength of schedule of a team in order to score and/or rank it. But the rankings were created by means of aggregating the individual subjective rankings of

human beings. As a result one would expect biases and/or individual preferences for information other than results and strength of schedule to color the outcomes. Even then, we can establish how close any ranking is to that of some GP method's ranking. This will give us an objective way of determining, at a fundamental level, which of the two main components (win percentage or strength or schedule) is favored more by each individual ranking. Then, having done that for the rankings by a given entity over multiple years, we can ask three questions for that ranking entity: How fairly it ranks, how efficiently it ranks and whether its rankings include specific biases in favor of or against certain teams or conferences.

6.1. Best Fit metric used

In most of the rankings that are (or were) used by the NCAA only the top 25 teams are ranked. The metric favored in this work (labeled LR) in order to assess which GP method comes closest to matching a given ranking is the following:

$$LR = \sum_{i=1}^{25} \left| ln(x_i + \kappa) - ln(y_{i\alpha} + \kappa) \right|$$

where x_i represents the ranking position of a given ranked team (as ranked by a given entity), $y_{i\alpha}$ is the ranking position of that same team according to the GWP method used here and κ is a non-negative number. Notice that this means that $x_i \in \{1, ..., 25\}$ whereas $y_{i\alpha} \in \{1, ..., n\}$ because the GP methods result in a complete ranking of all teams. Notice as well that absolute values are used instead of squares. This is done to minimize the effect of outliers on the total sum, which avoids turning the best fit metric into a metric that best fits to the one or two outliers (with only 25 observations, this is a non-trivial matter). For robustness I also calculated the sum of squares in each case. Finally, notice that as $\kappa \to \infty$ the metric is equivalent to using the sum of the absolute value of the differences and when $\kappa = 0$ it is equivalent to using the difference in natural logarithms. Neither of these extremes is best. The difference in values gives the same weight to all ranking positions. But there is a sense in which a team that was ranked 24th by one method and 21st by another method was indeed closely matched by the two methods but a team that was ranked 1st by one and 4th by the other was not closely matched by the two methods. This is because teams at the top are on the tail of the distribution whereas teams at the bottom (25th out of more than 120 teams) are closer to the median. It presumably means that if two ranking methods are truly similar, it is more difficult to misalign the rankings of teams at the top of the 25 team ranking than it is those at the bottom. On the opposite extreme, using the difference in log values gives too much weight to the top ranking positions because very small misalignments there would be equivalent to extreme misalignments at the bottom of the 25-team ranking. In order to strike a balance we can calibrate κ to treat equally the average misalignment in each position. The calibrated value is $\kappa = 2.5$. With few exceptions, the results obtained are robust to changes in κ.

6.2. The Fairness Question

After analyzing the rankings of three different entities (BCS, AP and CFP) over 7 different years (14 different rankings because CFP came into existence as a replacement of the BCS) the results on the question of fairness were conclusive: Not a single one of the 14 rankings (even under any of the best fit metrics used as robustness checks) met the win dominance

standard. In other words, not a single one of the rankings was fair. This is because in order for a ranking to be considered fair it would have to have given no more than a 59% weight to the strength of schedule ($\alpha \geq 0.41$). Figure 3 shows the LR sum of absolute differences at $\kappa = 2.5$ between the ranking by the BCS/CFP and that of the GP method for different values of α for each of the 7 years with data. The stepwise nature of the graphs are the result of small differences in α not always changing the ordering of teams. The best fit occurs where the sum of absolute differences is at its lowest. For all years the minimum clearly occurs at values of $\alpha < 0.41$.

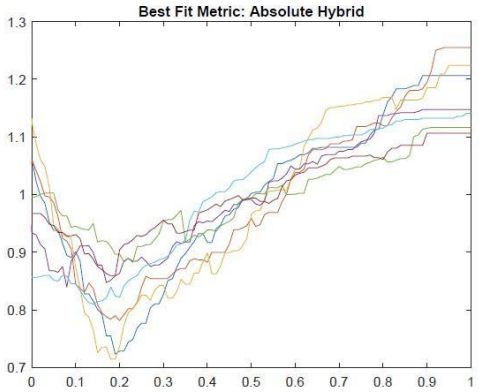

Figure 3: Sum of absolute κ−log differences for $\kappa = 2.5$

6.3. The Efficiency Question

In order to properly assess whether any ranking is efficient we must know the true win percentage of all teams. However, this is only possible through simulations. In any application, teams only play a limited number of games so it is impossible to know what the true win percentage is. To get around this problem (in this application) we can use data on betting odds for every individual game played in a given season. This allows us to obtain an implied strength vector s, under a very simple strength-based winning probability model discussed in Stern (1991). With this information in hand we can then obtain two important measures: First, an implied σ_W and second, through simulations, an implied ρ_W which can be contrasted to the actual σ_W and ρ_W from the data in order to validate or not the simple strength model used. Finally we can simulate the model, calibrated to each individual season, to obtain the most efficient α in order to contrast it to the one implied by any given ranking.

In Figure 4, the results of the calibration strategy are shown. The solid blue line labeled *Most Efficient* shows the most efficient levels of α for each year, given that year's calibrated parameter values (with the dashed lines representing +/- one standard deviation). The solid red line labeled *CFP/BCS* shows the implied values of α by the BCS (for the years 2011 to 2013) and by the CFP (for the years 2014 to 2017) using $\kappa = 2.5$ as the best-fit parameter. The solid green line uses the AP ranking as a control.

The remaining three solid lines show the lower bounds on α given by the respective fairness criterion used, that is: Global fairness, ex-ante fairness and, for completeness, ex-post fairness. Figure 4 clearly shows that ranking entities are much closer to ranking

teams in a way that is more consistent with efficiency rather than fairness.

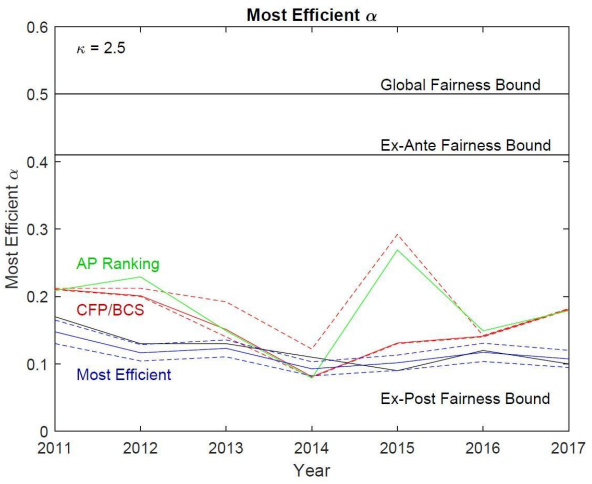

Figure 4: Most Efficient α by year

6.3. The Bias Question

Having established an implied α for each ranking of any given season, we can then take the rankings for all seasons by the same entity to try and determine if it consistently over or under-ranks any given teams or conferences. Of course, with more than 120 teams each season and only 7 seasons studied it is inevitable to find certain teams that will be consistently overrated or consistently underrated. The interesting question is a quantitative one: By how much the overrated teams are ranked above what the implied GWP method would have suggested. Figure 5 shows the 9 most overrated teams (of those ranked at least 4 times) in the past 7 years, with a quantitative assessment of how overrated they are: Using κ = 2.5 as our best-fit parameter, this figure shows how overrated each team was in every year that is was ranked by the entity that was in charge of determining which teams would qualify for the playoffs that year. The rankings used are those of the regular season (excluding bowl games and playoffs). Values above 1 mean that the team was overrated that year. More specifically, the ranking that a team should have received $r^* = (\gamma - 1)\kappa + \gamma r$, where r is the ranking it actually received and γ is the corresponding entry in the figure. Thus, for example, a value of 1.1 means that a team that was ranked 7th should have been ranked 8th or a team that was ranked 17th should have been ranked 19th. Similarly, a value of 1.35 means that a team that was ranked 6th should have been ranked 9th.

Team	2011	2012	2013	2014	2015	2016	2017	Total	#Obs
Michigan State	1.76	-	1.77	1.84	1.00	-	0.85	1.37	5
Florida State	1.36	1.19	1.29	1.00	1.87	1.08	-	1.27	6
Southern California	1.67	-	-	0.96	-	1.26	1.19	1.25	4
Oklahoma State	0.82	-	1.06	-	1.00	1.77	1.72	1.22	5
Baylor	1.00	-	1.12	1.27	1.24	-	-	1.15	4
Louisville	-	0.96	1.10	1.09	-	1.46	-	1.14	4
Wisconsin	1.61	-	1.09	1.10	1.20	1.00	0.88	1.13	6
Oklahoma	0.67	0.93	1.15	-	1.00	1.32	2.11	1.12	6
Clemson	1.24	1.06	1.00	1.15	1.29	1.00	1.00	1.10	7

Figure 5: Most overrated teams between 2011 and 2017

The same question can be asked of the different football conferences, which include anywhere from 10 to 16 teams each. Figure 6 shows how over or underrated every conference (with at least 4 ranked teams in the seven years studied) was. An interesting finding is worth mentioning: All power five conferences were overrated relative to all other smaller conferences. However, with the exception of the ACC and the Big12 the degree of overrating in comparison to the smaller conferences was not very strong.

Conference	2011	2012	2013	2014	2015	2016	2017	Total	#Obs
ACC	1.05	1.13	1.13	1.20	1.53	1.25	0.91	1.17	23
Big 12	0.74	1.22	1.11	1.01	1.04	1.38	1.71	1.11	23
Big 10	1.19	0.98	1.20	1.24	0.94	0.90	0.94	1.03	33
PAC	1.19	0.95	0.92	0.93	0.99	1.04	1.08	1.00	28
Indep	-	1.00	1.07	-	1.10	-	0.82	0.99	4
SEC	1.03	1.04	0.89	0.90	0.85	1.13	1.15	0.98	37
MWC	1.05	1.31	0.91	0.66	-	-	0.96	0.97	6
AAC	-	-	1.05	-	0.94	1.00	0.69	0.89	10
MAC	-	0.97	0.92	-	-	0.66	-	0.87	4

Figure 6: Most overrated conferences between 2011 and 2017

Another interesting question that can be answered relating ranking biases is whether the right teams were chosen for each season's semi-finals (or just the final in seasons prior to 2014). The following teams were left out of contention for the championship as a result of ranking biases: Oklahoma State in 2011, Florida in 2012, Texas Christian in 2014 and Central Florida in 2017.
Finally, and most glaringly, there was one of the above teams that despite being left out of the championship playoff, despite having had to play a lower-ranked opponent in its bowl game and despite the fact that the teams that did get to play for the championship got an extra boost from playing higher-ranked teams, still managed to rank number one among all teams as objectively measured by the implied ranking method used that year by the very ranking entity that chose to leave it out. That team was the 2017 Knights of the University of Central Florida.

7. Conclusion

Head-to-head match-ups in sports and other competitions conclude with the declaration of a winner and a loser (or the absence of both which defines a draw). After multiple matches involving multiple teams it is natural and customary to establish a final ordering of the teams. The win percentage scoring method is the most widely used, least disputed method to create such an ordering in complete tournaments. Using it as a benchmark, I presented a parsimonious family of scoring methods that uniquely satisfy basic fairness standards for the case of incomplete tournaments (the most important one being that no team should be awarded more points for a win than for a loss, regardless of the opponents). It includes the win percentage method as a special case. I then analyzed this family of scoring methods in terms of efficiency, defined as how close each scoring method comes to capturing what the teams' win percentages would have been had the tournament been complete. I showed that there is a clash between fairness and efficiency in that the most efficient scoring method will typically be an unfair one. Finally, using data on betting odds and results for the NCAA division 1 football tournament I calibrated the family of scoring methods to match as

closely as possible the actual rankings that were used to determine the teams that would go on to compete for the championship in each of the years ranging from 2011 to 2017. The main findings are that the rankings used by the NCAA were generally efficient (if anything the strength of schedule component was under-utilized under this metric) but clearly unfair (the strength of schedule component was over-utilized under this metric) and there were quantifiable biases present in the rankings, the most glaring one occurring during the 2017 season where the best team in the country was left out of the four-team playoff that ultimately determined that year's champion.

References

[1] Chebotarev, P. [1989] "Generalization of the row sum method for incomplete paired comparisons" *Autom Remote Control* 50, 1103-1113.

[2] Chebotarev, P. [1994] "Aggregation of preferences by the generalized row sum method" *Mathematical Social Sciences* 27, 293-320.

[3] Chebotarev, P. and Shamis, E. [1998] "Characterizations of scoring methods for preference aggregation" *Annals of Operations Research* 80, 299-332.

[4] Chebotarev, P. and Shamis, E. [1999] "Preference fusion when the number of alternatives exceeds two: indirect scoring procedures" *Journal of the Franklin Institute* 336, 205-226.

[5] Colley, W. [] "Colley's Bias Free College Football Ranking Method" *unpub- lished.*

[6] Csato, L. [2019] "Some impossibilities of ranking in generalized tourna- ments" *International Game Theory Review* 21(1) 15pgs.

[7] Gonzalez Diaz, J., Hendrickx, R. and Lohmann, E. [2014] "Paired compar- isons analysis: an axiomatic approach to ranking methods" *Social Choice and Welfare* 42, 139-169.

[8] Jackson, Matthew [2008] "Social and Economic Networks" *Princeton U. Press*.

[9] Keener, J. (1993), "The Perron-Frobenius Theorem and the Ranking of Football Teams" , *SIAM Review* 35 (1), 80-93.

[10] Leiva Bertran, F. [2019] "Scheduling incentives in incomplete tournaments" *working paper*.

[11] Palacios-Huerta, I. and Volij, O. [2004] "The measurement of intellectual influence" *Econometrica* 72(3), 963-977.

[12] Slutzki, G. and Volij, O. [2005] "Ranking participants in generalized tour- naments" *International Journal of Game Theory* 33(2), 255-270.

[13] Slutzki, G. and Volij, O. [2006] "Scoring of web-pages and tournaments-axiomatizations" *Social Choice and Welfare* 26(1), 75-92.

[14] Stern, H. [1991] "On the probability of winning a football game" *The Amer- ican Statistician* 45(3), 179-183.