Can Statistical Models Out-predict Human Judgment?: Comparing Statistical Models to the NCAA Selection Committee

Luke Stanke
University of Minnesota – Twin Cities,
Minneapolis, Minnesota, United States, 55455
Email: stan0337@umn.edu

Abstract

The NCAA selects teams for their championship tournaments using a selection committee. For the Men's Division I NCAA Basketball Tournament the committee must factor in the results of over 5,000 regular season games, as well as other qualitative factors. One of the key summary statistics used by the committee during the process is the Rating Percentage Index (RPI). This paper evaluates the committee's selection process, as well as the RPI and winning percentage statistics, from the 2002-03 through the 2010-11 seasons. In the process of evaluating the committee, RPI, and winning percentage, a series of Bradley-Terry Models are developed to use as comparison models. Results suggest that the RPI statistic and the committee predictions are significantly correlated. Additionally, the all Bradley-Terry Models predict more games correctly than the committee, suggesting that a statistical model might be better suited in seeding teams for the NCAA tournament.

1 Introduction

Every March, millions of Americans experience "March Madness". These people often cheer for their favorite college basketball teams, root for the underdog, and participate in office pools to predict winners of the NCAA Division I Men's Basketball Tournament, a single-elimination event used to determine the top Division I college basketball team.

This prominent event features the top sixty-eight Division I college basketball teams. Teams must automatically qualify or be selected by a panel of experts to compete in the tournament. Thirty teams automatically qualify for the tournament by winning their conference tournament, a single team from the Ivy League automatically qualifies by winning the regular season conference championship, and the additional thirty-seven spots are determined by a group of experts who use quantitative and qualitative information for determining the selection and seeding of teams.

Conferences also feel the pressure of the high-stakes decisions from the selection committee. The athletic conferences each receive an annual payment from an NCAA basketball fund of more than $140 million. The payout to each conference is based on the number of games that teams from a conference participate in during the NCAA tournament over a six-year period [1]. This means the more teams that make the tournament from a conference, the potential for more earnings for a conference. Additionally the exposure of teams and conferences may result in additional TV revenue, merchandising sales, and royalties.

If teams are not evaluated appropriately during the selection process, a team could miss out on an opportunity to win tournament games and the national championship. As a result, the school

and the conference would miss out on additional earnings. The ranking procedure completed by the selection committee could also affect the outcome of the tournament. If the committee misranks two teams, it could result in two teams of greater "ability" playing early in the tournament, resulting in the elimination of a stronger team earlier in the tournament process.

This paper will evaluate the committee's ability to seed teams in the NCAA tournament. For this evaluation, a series of Bradley-Terry models will be develop and compared to the predictions of the committee. From this, we can determine which method is best for predicting the outcomes of games, and therefore likely a better method for selecting and seeding teams for the NCAA tournament.

2 Understanding the Tournament Selection Process

Teams are selected and ranked for the NCAA tournament by a formal selection committee. This selection committee consists of ten athletic directors and conference commissioners throughout Division I Men's Basketball. Members of the committee are representative of the conferences around the country. Additionally, the committee is geographically balanced. At any given time, the committee must consist of at least two members representing the East, Midwest, South, and West regions of the United States [2].

After the thirty-one automatic bids are earned, the members of the committee discuss the merits of each of the remaining teams and determine the top thirty-seven teams that have not automatically qualified for the tournament. The major components used to select at-large teams by the committee are not entirely clear. However, the committee does use the Nitty Gritty Report [3], which includes Rating Percentage Index (RPI), an index created by the NCAA to assist in post-season selection. Other factors in the Nitty Gritty Report include: conference winning percentage, record in road and neutral games, record against teams in the RPI top fifty, and the strength of schedule of conference and non-conference teams. Also, committee members take into consideration "subjective concepts; e.g., how well a team is playing down the stretch, what the loss or return of a top player means to a team, or how emotional a specific contest may be." [4]

After selecting the at-large teams, the sixty-eight teams are seeded, a process where teams are grouped in sets of four to determine their true seed. After seeding at sixteen different levels, teams are placed in the championship bracket using a pre-defined procedure. There are some limits in the placement of teams, for instance no more than two teams from a conference can be in the same region of a bracket, and teams cannot play games at a location in which they played more than three games during the regular season. As a consequence of this procedure, teams can be moved one seed up or down from their true seed location. [5]

3 Understanding the Ratings Percentage Index (RPI)

One key component in ranking teams by the selection committee is the RPI. The RPI is a statistical calculation used to rank every Division I school and continues to receive a lot of weight from the selection committee (West, 2006). The index takes into account each team's winning percentage (WP), the winning percentage of its opponents (OWP), and the winning percentage of its opponents' opponents (OOWP). The ratings percentage index (RPI) for a given college basketball team *i* is calculated by the NCAA is located in Equation 1.

$$\text{RPIi} = 0.25\ WP_i + 0.50\ OWP_i + 0.25\ OOWP_i \tag{1}$$

Slight changes were made to this equation in 2004, when the selection committee approved weights to be applied to the formula. When calculating RPI, home game wins and road game losses would receive a weight of 0.6, while road wins and home losses received a weighting of 1.4.

The RPI formula is designed to reward a team that plays teams with a higher winning percentage and penalizes a team that plays teams with a lower winning percentage. Despite this, the RPI statistic contains a major flaw. This flaw is illustrated when a team's RPI score actually decreases by defeating a poor opponent, even if the superior team defeats the team by a large point margin, which would be expected. This issue also works the other way when team plays a team with a better record, the teams RPI score can actually increase, even if the team loses by a large number of points. These scenarios illustrate only a win or loss is taken into account when ranking teams, and that the margin of victory is not taken into consideration. It should be expected that if two teams play the same opponent, the team with greater ability would win by a greater margin. In actuality, this might not occur, but if these two teams win, their RPI score will be affected identically.

4 The Bradley-Terry Model: An Alternative to the RPI

Game data fit to a Bradley-Terry model is one possible alternative to the RPI. Bradley-Terry models allow for the comparison of a pair of stimuli, and from that pair, one of the two stimuli must be chosen over the other. Choices are based on which stimulus has a greater value on a unidimensional latent scale. The stimulus with the greater value at a particular moment will be the chosen, where the latent values of the individual are normally distributed. For basketball, the latent scale is overall team basketball ability.

A Bradley-Terry model looks to measure where a decision is made on the basis of a pair (i,j). For instance, the preference for team i over team j can be expressed as $Y(i,j) = 1$, and a choice for the alternative results in $Y(i,j) = 0$. In an illustrated form, if i is Duke and j is North Carolina, and Duke wins, $Y(i,j) = 1$, however if North Carolina wins, $Y(i,j) = 0$. The main assumption for the Bradley-Terry model is that the probability of i is proportional to a constant π_i, which is independent of all other alternatives. Therefore, the probability of team i winning over team j is equal to:

(2) $$P_{ij} = \frac{\pi_i}{\pi_i + \pi_j}$$

Thus, the probability of team i or team j winning is equal to 1 ($\pi_i + \pi_j = 1$). Equation two can be expanded from work completed by Glickman [6]. Glickman suggested that every team produces a true latent score, S_i, which is independent of the score of an opponent, S_j. The winner of a game is determined by the team with the greater latent score. Meaning if $S_i > S_j$, then we can expect team i is more likely to defeat team j. The probability of choosing stimulus i over stimulus j, P_{ij}, be calculated by using a formula similar to the rating scale model used in item response modeling. This basic logistic model can be written as:

(3) $$P_{ij} = \frac{\exp(S_i - S_j)}{1 + \exp(S_i - S_j)}$$

where S_j is equal to the log-odds true score of stimulus j, and S_i is equal to the log-odds true score of stimulus i.

In the most basic form, a Bradley-Terry Model might only take into account wins and losses. That simple model predicts winners of games very well, while only using a small amount of information [7]. However, this measure does not take into account points scored, and this additional information may provide a better estimate of ability for each of the Division I teams. If each point scored is treated as a game between two teams, then we would have much more information about the estimated ability of teams. Instead of having 30 data points for each team (the approximate number of regular season games a team plays), each team now has thousands of data points, since each game will typically produce more than 120 points. While there is no previous literature supporting the use of this method in basketball, it has been used to predict the abilities of soccer teams[8].

5 Is the Selection Committee Getting It Right?

Since we cannot go back and have a statistical model seed teams, and then have teams play out the tournament, we will never have a direct way of saying that a statistical model does a better job at seeding teams than the selection committee. We can, however, look to see if statistical models predict the outcome of past NCAA tournament games better than the selection committee. Using regular season data, latent rating scales can be developed, and then values on the rating scale can be used to predict the outcomes of NCAA tournament games.

The effectiveness of the selection committee can actually be evaluated. During the tournament selection process, the committee ranks teams in an order from greatest ability (#1 seeds) to lowest ability (#16 seeds). From this process, it could be interpreted that the committee expects higher seed (#1 seeds) to beat all lower seeds (#2 to #16 seeds). Using this idea, the committee's predictions can be compared the actual results over eight years of NCAA tournaments and also compared to other statistical models.

6 Analysis of Eight Years of NCAA Tournament Data

Data from the 2003-2004 through the 2010-2011 NCAA seasons were collected from an online source [9] and double checked [10]. Using regular season data from these eight years, four Bradley-Terry Models were fit to the data to estimate the ability of Division I teams using BradleyTerry2 [11] package designed for R [12]. Total regular season games in a given season ranged from 4,888 games in 2003-04 to 5,635 games in 2009-10. The total number Division I teams in one season ranged from 325 teams in 2006-07 to 345 teams in 2010-11.

The four Bradley-Terry models that were fit to the data can best be illustrated from a game where North Carolina beat Duke on their home court by a score of 81 to 67. The first model, called Win Model, will account for a team's wins and losses versus other teams, that is to say only data about North Carolina beating Duke will be used in the model.

The second model, called Win + Home Model, uses the same win and loss data from the first model, but will also include a home-court advantage covariate. In this model, data also includes information about North Carolina playing at home, which may produce an advantage. In this case, home court advantage is considered to be a value equal for all teams across Division I basketball. It should be noted that during the NCAA there is no home court advantage, and that this covariate makes adjustments for a team's ability during the regular season at home.

The third model, the Points Model, will incorporate total points scored for each team in a game. In this model, it is assumed that each point acts as an individual match. Using our example

above, this means that the model will show that North Carolina beat Duke 81 times and Duke beat North Carolina 67 times.

The fourth model, Points + Home, will include total points scored for each team and include a home court advantage covariate. Once again, points from both North Carolina and Duke will act as several matches being played, but this model will consider that North Carolina might have had a home-court advantage that must be considered in predicting the team's overall ability, a home field advantage that is constant for all teams.

These four Bradley-Terry models were then compared to three other models for predicting the outcome of the NCAA tournament. The first comparison model, called the Committee Model, uses a team's tournament seed as a predictor. This is the committee's prediction for each game, where the teams with higher seeds (#1) are expected to beat lower seeds (#2, ect.) in a game. In some situations teams with the same seed could potentially play each other. In this case, these games have been omitted from the analysis, as it was unclear which team was rated higher by the committee.

The second model will use RPI to predict the outcome of the NCAA tournament. In a given game, the team with a higher RPI value will be expected to defeat the other team. The final comparison model will use a team's winning percentage, called Winning Percentage Model, to predict the outcome of the games from the NCAA tournament. In this case, teams with a higher regular season winning percentage would be used to predict the outcome of a tournament game.

After the models were developed, the total number of NCAA tournament games predicted correctly was calculated for each of the seven models (4 Bradley-Terry models, 3 alternatives) to determine the model that best fits the data. A total of 515 NCAA tournament games were played from 2004 through 2011. All of the games were used for 6 of the models. The Committee Model, used only 499 total games, as16 games were omitted from the analysis due to teams with the same seed playing each other. After the total number of correct games predicted is calculated for each model an ANOVA will then be used to determine if a model is predicting the number of games correctly above and beyond other models in a way that could not otherwise be attributed to chance alone.

7 Results

The seven models were then fit to the NCAA tournament data to determine which model best predicts the outcomes of games. Results (table 1) suggest that the Bradley-Terry Models outperform the other models (a further breakdown of the models by season can be found in Appendix A). In fact, models seem to cluster into three groups. The highest performing models are the Bradley-Terry models using only win/loss data. These two models correctly predicted approximately 89% of games in the NCAA tournament games from the past eight seasons. The next group of models is the Bradley-Terry Models using points a method for ranking teams. These models predicted over 82% of games correctly. The third group is the alternative models, the Committee Model, The RPI Model, and the Winning Percentage Model. These models range from 69.1% of games correctly picked to 72.9% of games correctly picked.

An ANOVA was then run to determine if any of the models significantly out predicts other models. The results of the ANOVA, $F(6, 3582) = 23.39$, $p<.0001$, suggests that at least one method is significantly different in the ability to predict the outcome of games. To gather more information, a series of contrasts were run. Based on the ANOVA contrasts, the Win, Win + Home, Points, and Points + Home Models all out predicted the Committee Model ($p<.0001$). This suggests that the Bradley-Terry developed in this paper do a better job of picking the winners of the NCAA tournament than the seeds that were determined by the committee.

Table 1 Total Correct and Percent Correct of Tournament Games Predicted Correctly By Model

Model	Total Correct	Percent Correct
Win	459	89.1%
Win + Home	460	89.3%
Points	436	84.7%
Points + Home	425	82.5%
Committee	364*	72.9%
RPI	374	72.6%
Win Pct.	356	69.1%

* The Committee Model was only determined from 499 games.

The contrasts between the RPI or Winning Percentage and Committee models failed to show significant differences in predictive ability. This means that using winning percentage of team does no worse than using the seed of teams to predict games. Looking closer at the relationship between RPI and Committee models, the correlation between the models is .742 (p<.0001) suggesting the models are very closely related. In contrast, the correlation between the Committee Model and model with the most correct predictions, the Win + Home Model, is an insignificant 0.064. (Additional model correlations are provided in Appendix B).

7 Conclusion

Selecting and seeding teams for the NCAA Division I Men's Basketball Tournament has always been a high-stakes process for fans, teams, and conferences. The selection process heavily relies on a selection committee to determine the teams and seeds of teams that will be a part of the tournament. It is clear from the results that the suggested Bradley-Terry models out predicted the committee's seeding procedure. In fact, the results suggest that the committee's ability to take into account injuries, recent play, and individual game emotional context does not aid in the correct seeding of teams. It could be that the committee overweighs certain subjective factors, or that these subjective factors do not really matter for selecting teams. Further, the use of the seeds does no better in predicting games than using a simple winning percentage.

It appears that there is a strong connection between the committee's predictions and RPI predictions, suggesting that there is heavy use of this statistic when ranking teams. The RPI statistic has been revised but it still carries some serious flaws. The use of the Win + Home Bradley-Terry Model may be a suitable alternative to the RPI, as it also takes into account the strength of a team, while still only using win-loss data.

While statistical models are used in other collegiate sports to determine the championship game have faced great scrutiny, this paper shows that at least a small set of statistical models seem to do a better job at choosing teams for a championship tournament than human evaluators. While no changes in the selection process have been suggested for the foreseeable future, it makes sense for the selection committee to at least think about using other statistical models for NCAA tournament selection.

9 References

[1] Schwartz, Peter J. "The Most Valuable College Basketball Teams." *Forbes.com*. Forbes, 16 Mar. 2010. Web. 5 May 2011.
[2] National Collegiate Athletic Association (NCAA). "NCAA Division I Men's Basketball Championship Principles and Procedures for Establishing the Bracket." *NCAA.org*. NCAA, n.d. Web. 5 May 2011.
[3] Johnson, Greg. "Basketball committees offer more access to selection materials." *NCAA*.org. National Collegiate Athletic Association, n.d. Web. 5 May 2011.
[4] "NCAA Rating Percentage Index" *NCAA.org*. NCAA, n.d. Web. 2 Jan 2012.
[5] "NCAA Division I Men's Basketball Championship Principles and Procedures for Establishing the Bracket." *NCAA.org*. NCAA, n.d. Web. 2 Jan 2012.
[6] Glickman, Mark. E. *Paired comparison models with time-varying parameters*. Diss. Harvard University, 1993. Web. 3 Mar. 2011.
[7] Matthews, Gregory J. *Improving paired comparison models for NFL point spreads by data transformation*. Diss. Worcester Polytechnic Institute, 2005. Web. 14 Feb. 2011.
[8] Leitner, Christoph, Zeileis, Achim, and Hornik, Kurt. "Bookmaker Consensus and Agreement for the UEFA Champions League 2008/09." *IMA Journal of Management Mathematics*. 22(2): 183 - 194. Web. 14 Jan. 2011.
[9] Pomeroy, Ken. *Kenpom.com*. The Forecast Factory, 2011. Web. 10 Mar 2011.
[10] *Sports Reference/College Basketball*. Sports Reference, 2012. Web. 23 Mar 2011.
[11] Turner, Heather. and Firth, David. "Bradley-Terry models in R: The BradleyTerry2 package. R package version 0.9-6." *R-project.org*. R-project, 2 Jan. 2012. Web. 9 Jan. 2012.
[12] R Development Core Team (2003). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, 2003. Web. 1 Apr. 2011.

9 Appendicies

Appendix A - Number of Correctly Predicted Games For Each Model By Tournament Year.

		Bradley-Terry Models				Alternative Models		
Year	Total Games	Win	Win + Home	Points	Points + Home	Committee* (correct/total)	RPI	Win Pct
2004	64	56	56	53	52	48/63	45	39
2005	64	58	59	56	54	45/63	47	44
2006	64	60	58	56	55	42/63	41	45
2007	64	58	56	54	53	50/62	47	45
2008	64	58	59	54	52	47/60	47	46
2009	64	57	59	55	55	47/63	51	44
2010	64	54	55	52	50	42/62	50	49
2011	67	58	58	56	54	43/63	46	44
Total	515	459	460	436	425	364/499	374	356

* Games where two teams with the same seed were excluded from the analysis for the Committee

Appendix B - Correlation Matrix of Tournament Predictions by Model Type

	Win	Win + Home	Points	Points + Home	Committee	RPI	Win Pct.
Win	1.000	.828	.388	.398	.108	.121	.145
Win + Home	---	1.000	.341	.371	.065	.112	.150
Points	---	---	1.000	.897	.170	.125	.054
Points + Home	---	---	---	1.000	.173	.165	.102
Committee	---	---	---	---	1.000	.742	.420
RPI	---	---	---	---	---	1.000	.457
Win Pct.	---	---	---	---	---	---	1.000