Applying Occam's Razor to the Prediction of the Final NCAA Men's Basketball Poll

John A. Trono

Saint Michael's College
One Winooski Park
Colchester, VT 05439 (USA)
jtrono@smcvt.edu

Abstract

Several approaches have recently been described that attempt to match how the coaches vote in the final poll that is taken after the men's NCAA basketball championship concludes (Trono and Yates 2014, 2015). The new strategy presented here: is more straightforward than the prior approaches; has demonstrated reasonably high correspondence with the actual, final polls; and its results can be generated much more easily than the most accurate of the aforementioned approaches.

1 Introduction

Once the National Collegiate Athletic Association (NCAA) selection committee has completed its deliberations, and that year's men's basketball tournament bracket has been made public, the following three weeks of tournament games generate more interest (at least within the United States) than any other sporting event that spans this length of time. During the first few days after the bracket is announced, millions of individuals will each spend a varying amount of time agonizing over how to complete their bracket – in an attempt to be the winner of various types of pools.

Strategies for making successful predictions of all 63 tournament games, which are played over six 'rounds', are as unique as the teams competing therein. However, this paper will *not* focus on the prediction of games that occur during the tournament, but instead, it will focus on the prediction of what the final rankings will be – as generated by the coaches who vote in the final poll which is taken *after* the championship game has been played.

2 The first prediction strategy

In a previous study, a weighted, least squares regression model was used to predict the number of votes each team would achieve in the aforementioned final poll (Trono and Yates 2014). The chosen equation used three objective quantities to generate a team's predicted vote total: its winning percentage (PCT), multiplied by 100; the number of NCAA tournament wins (W) + 1 earned by that team; and the team's power rating (PR), as determined by the power rating system (Carroll, Palmer and Thorn 1988). The plus one, in W + 1, positively distinguishes those teams who received an NCAA tournament bid, and lost their only game, over teams who were not invited. National Invitation Tournament (NIT) wins were also deemed to be worth one quarter of an NCAA tournament win in that study.

Even though this model was reasonably accurate, in both a team's rank – and its predicted vote count, the weights in this model were improved, using Monte Carlo techniques, to try and match each team's final rank more accurately by eliminating the goal to match the predicted final vote total as closely as

possible. Using the final polls from 1993-2007 as the training set, the most accurate set of weights that were found appear in the following, 'best' Monte Carlo equation (MCB): 6.68605 * PCT + 17.64763 * PR + 88.24644 * (W + 1). The constant term in the original regression equation was dropped because an accurate final vote total is no longer relevant since the accurate order of the teams in the predicted final poll is now the primary goal.

The accuracy of this model was determined by the Spearman Correlation Coefficients (SCCs) for the top 15, 25 and 35 teams in the final polls. The goal was to maximize the sum of these three values, and when applied to the training set, the MCB model produced an average, yearly sum of 2.57721, with the three individual, average SCC values ranging between 0.85 and 0.865, respectively. For the nine subsequent years, the average sum is 2.51756, and the small drop in performance comes from a 0.04 decrease in the average SCC-25 value, and another 0.02 decrease in the average SCC-35 observed result.

3 Subsequent prediction strategy

While the article describing how the MCB model came to be was still in production, during April of 2013, several revelations occurred which eventually led to a more accurate approach (Trono and Yates 2015). The primary idea behind the first improved model is that teams with the same number of NCAA tournament wins typically remain in the same relative order as they appeared in the penultimate coaches' poll, which occurs right after all of the conference tournaments – that determine the teams who will receive each conference's automatic bid to the NCAA tournament – have finished. The number of teams ranked in between such teams will usually shrink, depending upon where these teams reside in the penultimate ranking, but most teams don't usually leap frog over other teams – with the same number of NCAA tournament wins – from the penultimate to the final vote.

While the vote totals which appear in the penultimate coaches' poll are certainly not truly '100% objective', once these totals are known, a prediction which incorporates them is deterministic, i.e. reproducible. And because past behavior is typically a reliably accurate predictor of future behavior, those coaches who thought that team A should be ranked higher than team B will probably continue to think so unless significant, new evidence is produced, i.e. team B shows that it actually *is* a better than team A by winning one or more NCAA tournament games than team A wins. Therefore, it seems acceptable to incorporate this penultimate coaches' poll information into a prediction model – in an effort to improve the overall accuracy of such a model.

The secondary idea was to simply add to the penultimate coaches' poll vote total a bonus reward quantity which is associated with the number of NCAA tournament wins each team accrues. This reward should probably increase more as the number of such wins increases, i.e. the difference between the bonus that is associated with four tournament wins, instead of three, should be larger than the bonus difference between three and two wins, respectively.

To test this approach, the first reward strategy was to simply use the sequence inherent in Zipf's Law, starting with 1/7 as the bonus for those teams who were invited to the NCAA tournament, but lost their only game. (The recently added category of 'play-in' games are excluded from the total representing a team's number of NCAA tournament wins.) The bonus for a team with one win would be 1/7 + 1/6, and so on, up to the bonus for the champion: 1/7 + 1/6 + 1/5 + … + 1/2 + 1/1. The initial vote count used with this approach would also first be normalized into the zero to one range by dividing the number of votes a team receives in the penultimate coaches' poll by the maximum vote total that a team could receive. The final, bonus increment, i.e. + 1, for the champion then guarantees that they will be the top team in the final poll, using this particular reward approach – and that specific outcome has occurred in all previous final polls since 1993, when this poll was reintroduced after its previous manifestations in 1953, 1954, 1974 and 1975.

A basic investigation of this idea produced an SCC sum for this approach (ZPF) above 2.7 (where the SCC sum for the MCB model was almost 2.58), and every one of three SCC values were around the 0.9 level, or higher – on average, over the fifteen years in the training set. This initial, encouraging result fueled a deeper study including other possible bonus reward quantities that might predict the final poll more effectively than how well these rewards, that mimicked Zipf's Law, performed.

The individual SCC sums were all well above 2.7, for the ZPF model, in the following years (2008-12, which were the years available when the first study – describing the MCB model – was submitted), except for the SCC sum in 2010. The primary cause of this lower SCC sum (in 2010) was because Xavier, a #6 seed who won two NCAA tournament wins, only garnered four votes in the penultimate coaches' poll (and were ranked as the #33 team), and was therefore predicted to be the #26 team in the final poll, according to the ZPF model, but was actually ranked #14. (The MCB model predicted them to be #15.)

This large disagreement dramatically impacted the SCC-15 value. A similar situation occurred in 2013, where the 9th seeded (and unranked) Wichita State team played their way into the Final Four, and eventually was ranked as the #4 team in that year's final poll. Even with four NCAA tournament wins to their credit, being unranked meant that the only quantity contributing to their team's total, that would determine their predicted rank – as far as the ZPF model was concerned, would be the specific tournament reward for their four NCAA tournament wins. So, the ZPF model predicted them to be the #13 team (whereas the MCB model predicted #8).

On average, there were 46.8 teams that appeared in the penultimate coaches' poll during the 15 year timespan in the training set, and so over 300 division one teams received zero votes in that poll, and are therefore all equal in some sense, at least with regards to models like ZPF – from the strong teams like Wichita State (26-8) down to a team with zero wins (e.g. Grambling, which was 0-28 in 2013). It seemed like the final poll predictions would be more accurate if each team could be assigned some quantity that would help to distinguish, i.e. separate somewhat, all the teams that reside in the unranked majority.

Thankfully, the tournament selection ratio (TSR) seemed to be an appropriate metric for the task (Trono 2013). The TSR value for any team is between zero and one, and half of this quantity is based on the two penultimate polls (populated by coaches and sportswriters, i.e. the AP poll for the latter, and, both polls are weighted equally), and so the highest TSR value a team can attain, that is not ranked in both polls, is 0.5 Eight computer based rating/ranking systems contribute the other 0.5. A trimmed Borda count is used amongst the eight, where four rating systems include each game's margin of victory (MOV), and the other four ranking systems do not. Wichita State's TSR value in 2013 was 0.44952, and their rank was therefore improved to #6 once this component was incorporated into a modified ZPF model. (Xavier also moved up to #16 – at the end of the 2010 season – when including this component.)

A total of five prediction models were deemed the most accurate in the follow-up study, where other patterned, bonus reward sequences were evaluated against the training set – and were added to either the normalized, penultimate coaches' poll value, or the TSR (Trono and Yates 2015). The first five rows in Table 1 contains the results for these models (where OCC will be described shortly). ZP2 follows the same bonus reward pattern as ZPF, except that the first denominator is 8, and the last is 2. PR2 uses prime numbers as denominators (17, 13, 11, 7, 3, 2), but uses two as the numerator instead. LN2 begins with 0.1, and then adds 0.2, then 0.3 and so on until finally adding 0.7 to the championship runner-up's bonus reward value in the sequence (2.1), so that the champion's bonus value becomes 2.8. Finally, the 50T approach attempts to use 50% as the amount of increase from one term to the next. The initial value was chosen by trying to maintain a two digit, decimal value for each term as well as have the final bonus reward value be roughly one larger than the runner-up's value. With these constraints in place, the first value was chosen to be 0.24, which produced the following bonus sequence: 0.24, 0.36, 0.54, 0.81, 1.21, 1.81 and 2.71. The baseline model was created for comparison purposes, and it essentially orders teams

using the number of NCAA tournament wins that they earned – utilizing the sum of a team's PCT and PR values to break the ties that would occur when teams have the same number of tournament wins in this simple, benchmarking model. Therefore, the bonus reward sequence for the baseline model was simply the values from one through seven.

To be able to produce the MCB model predictions, or any of the three other models in Table 1 that rely on the TSR, one would need to have access to at least the PR values, because the power rating system contributes two of the eight values in the TSR's computer-based component: one with, and one without, MOV. The rating system devised by Jeff Sagarin, whose ratings can be found in *USA Today*, is also one of the eight TSR computer based systems, but needless to say, without access to the TSR values, these three models would not be beneficial to someone attempting to make predictions about the final coaches' poll.

Table 1 – Spearman correlation coefficients (for a variety of models)

Model Acronym	Model-type, or, uses Pen/TSR	Training Set 93-2007	Predictions 2008-16
ZPF	Pen.	2.77810	2.77184
ZP2	TSR	2.80695	2.84642
LN2	Pen.	2.79890	2.84439
50T	TSR	2.78121	2.86102
PR2	TSR	2.77118	2.84387
MCB	Lin. Regress.	2.57721	2.51756
OCC	Multiplicative	2.75766	2.73884
Baseline	Win-based	1.80075	1.53183

At least for the ZPF and LN2 models, all that is required, to make predictions, is the penultimate coaches' poll, and knowledge of the bonus reward sequences used therein. The OCC system also essentially only needs the penultimate coaches' poll's actual ranks for each team, as the normalization step, which is required for ZPF (and LN2), is unnecessary.

4 One basic operation

While studying the distribution of how NCAA tournament wins match up with the penultimate and final polls – both with regards to lists of final ranks, and which integer pairs (the penultimate rank, and NCAA win total) earned that specific rank, as well as lists of which final ranks were derived from the same penultimate poll rank (and accompanying NCAA tournament win count) – a very basic (and simple) strategy seemed worth investigation.

Historically, most teams with one NCAA tournament win, who were ranked between roughly #25 and 'the teens', in the penultimate coaches' poll, have tended to end up very close to the same final rank, whereas those same teams moved up in the final poll with more wins, and fell in the ranking, with zero tournament wins.

It would be nice if one could say that a team ranked #P in the penultimate coaches' poll, and who earned W NCAA tournament wins, would end up being ranked as the #R team in the final coaches' poll. However, a team's final rank is also related to how other teams fared in the NCAA tournament. For instance, three wins will typically move a team higher in the rankings if the teams above said team only won two or fewer tournament games.

In an attempt to model this particular behavior, the strategy was to simply divide the team's penultimate rank by $2^{(wins-1)}$, and then order the teams by these new rank values to determine the final ranking. It is hard to see how this prediction can be determined in a simpler fashion, so, as long as this approach is fairly accurate, it appears that the Gordian knot of the aforementioned, computationally sophisticated strategies, will have been severed by this idea – and explains why it is called OCC, which is short for Occam.

5 Implementation decisions for the OCC approach

Before this particular approach can be effectively evaluated, several parameters must be established. Since a team's penultimate rank is required to produce the rank value, which will determine said team's predicted final rank, some value must be used for all teams that are unranked, i.e. those teams who receive zero votes in the penultimate coaches' poll. Of all the years in the training set, 1993 had the most teams (57) that received at least one vote. So, for initial investigation purposes, 60 seemed to be a reasonable value for the penultimate rank of all the unranked teams. The initial results after making this temporary decision were somewhat promising; however, several minor adjustments seemed necessary before the optimal value for the penultimate rank could be determined.

First off, this particular approach produced some ties since there are instances where a team, with a penultimate rank of P, that also earned W NCAA tournament wins, will have the same rank value as any other team with W+1 wins that also had a penultimate rank of 2*P. To alleviate such situations, a small value should be added to P before the rank altering division operation is performed. Using the training set, adding a small positive value produced somewhat more accurate results than adding a small negative value, so 0.05 was chosen though most any value under one half should yield the same results as the positive quantity selected here.

Secondly, it seemed appropriate to add a small integer constant to the penultimate rank, before the rank altering operation is applied, to lessen the numerical advantage that the #1 team (in the penultimate coaches' poll) might have – given the specifics of this particular strategy. Simply adding one to the penultimate rank maximized the SCC sum results for the years in the training set, so now the best rank constant, for the unranked teams, can be searched for.

So, if a team's penultimate rank is P, then P + 1.05 will be divided by $2^{(wins-1)}$, to produce that team's ranking value, which determines its final rank, after all the rank values are placed into ascending order. For unranked teams, the most accurate value for P, using the training set, was found to be 67, though variations in the average SCC sums were typically quite small (0.002 to 0.011) when changing this unranked value by plus or minus one. For instance, when examining the values of P in the range from 57 to 76, the average SCC sums varied from 2.7204 up to 2.75866 – the latter, when using 67.

Once 67 was chosen, the two previous parameter evaluations were revisited, and it was confirmed that it was still best to add one to the penultimate rank as well as adding a small constant (for tiebreaking purposes, regarding the teams' rank values) before performing the appropriate division operation. The NIT divisor of four, as chosen for the aforementioned MCB model, also worked well for the OCC model as that divisor gave the NIT champion the equivalent of 1.5 NCAA tournament wins. The OCC model tended to place a previously unranked NIT champion somewhere between #30 and #40 – which is where the NIT champion has usually ended up in the final coaches' poll except on five occasions (all before 2002); three of those occasions, the NIT champion was #28 or #29, and in 1993 they were #25 as well as #24 in 1997. Only 1.9 and 2.1 were evaluated instead of 2, in the OCC's rank-altering division operation, and neither produced more accurate results. Teams that receive votes in the penultimate coaches' poll, and who are invited to the NIT, are therefore sometimes predicted to be ranked above the NIT champion, as occurred in 2016, and is illustrated in Table 2, where NIT wins are designated by "#" in the wins column.

Table 2 – Various ranks and rank values (for 2016)

Pen. Rank	NCAA Wins	Rank Value	Pred. Rank	Final Rank
1	3	0.513	3	3
2	0	6.100	8	7
3	5	0.253	2	2
4	3	1.263	5	6
5	3	1.513	6	5
6	6	0.220	1	1
7	4	1.006	4	4
8	0	18.100	21	14
9	1	10.050	15	11
10	0	22.100	22	19
11	2	6.025	7	8
12	2	6.525	9	9
13	1	14.050	17	16
14	1	15.050	18	20
15	2	8.025	11	12T
16	0	34.100	29	22
17	2	9.025	13	12T
18	0	38.100	33	24
19	2	10.025	14	15
20	2	10.525	16	18
21	0	44.100	37	29
22	0	46.100	38	31
23	0	48.100	39	28
24	1	25.050	23	25
25	1	26.050	24	27
26	3	6.763	10	17
27	1	28.05	25	33
28	1	29.05	26	30
29	1	30.05	27	26
30.5	2	15.775	19	21
30.5	"2"	36.271	31	40T
32	2	16.525	20	23
33	1	34.050	28	34T
34.5	1	35.550	30	32
36	1	37.050	43	37
37.5	1	38.550	34T	34T
37.5	1	38.550	34T	40T
40	0	82.100	50T	38
40	1	41.050	36	40T
67	4	8.506	12	10
67	"5"	48.119	40	34T
67	"4"	57.223	41	39

Table 2 illustrates the predictions that were made by the OCC model after the 2016 NCAA tournament completed. Within this table, a penultimate rank of 67 indicates that that team was unranked, while teams listed at 37.5 implies that they were both tied for the #37 rank. (Dayton, which was also ranked #34, did not receive any votes in the final coaches' poll, and therefore does not appear in Table 2.) Three teams received the same number of votes, and since all three were tied for 39th in the penultimate coaches' poll, the rank of 40 represents their average rank: (39 + 40 +41) / 3 (though only two of them appeared in the final poll). The number of NIT tournament wins, for a team, are enclosed in double quotes in Table 2, and a 'T' within the final rank column indicates teams that were tied for that rank – in that poll. (The SCC values for OCC in 2016 were: 0.86339, 0.88827, and 0.90640, for the top 15, top 25 and top 35 teams respectively.)

6 Conclusion

This study has illustrated that a very straightforward, exponential update of a team's integral rank position, in the penultimate poll – which is taken just before the NCAA men's basketball tournament begins – models quite accurately the position where teams will be ranked in the final poll, which is taken once that tournament is completed. The magnitude of these updates relates directly to a team's performance during the NCAA tournament (and even less so for the NIT) where a team's regular season accomplishments are purportedly captured in the rank they are assigned in the penultimate poll. This new model (OCC) compares quite favorably with the slightly more accurate, but much more time consuming to generate, previous models that were briefly described here, and are described in more detail in the last two references below.

References

[1] Carroll, B., Palmer, P., and Thorn, J. (1988) *The Hidden Game of Football*, Warner Books.

[2] Trono, J. (2013) *Evaluating Regional Balance in the NCAA Men's Basketball Tournament using the Tournament Selection Ratio*. In Proc. of the 4th International Conference on Mathematics in Sport.

[3] Trono, J., and Yates, P. (2014) *How Predictable is the Overall Voting Pattern in the NCAA Men's Basketball Post Tournament Poll?* Chance, 27(2):4-12.

[4] Trono, J., and Yates, P. (2015) *Predicting the NCAA Men's Postseason Basketball Poll More Accurately*. 27th European Conference on Operational Research.