

Journal of Quantitative Analysis in Sports

Volume 4, Issue 3

2008

Article 2

Reweighting the Bowl Championship Series

Susan Buchman*

Joseph B. Kadane†

*Carnegie Mellon University, sbuchman@stat.cmu.edu

†Carnegie Mellon University, kadane@stat.cmu.edu

Reweighting the Bowl Championship Series*

Susan Buchman and Joseph B. Kadane

Abstract

The majority of statistical work on college football's Bowl Championship Series (BCS) has involved proposing or categorizing computer ratings of teams. Computer algorithms, a coaches' poll, and a media poll make up the three ratings systems that are currently equally weighted to produce an overall BCS rating, which ultimately determines which schools will compete in lucrative post-season BCS bowls. We focus on investigating the performance of the BCS as implemented for the 2004, 2005, and 2006 seasons to determine whether equal weights are appropriate. Our Bayesian analysis shows that while the posterior mode places more than half the weight on the media poll, the 95% HPD credible interval contains the equally-weighted scheme. We relate our work to the ongoing controversies over the BCS.

KEYWORDS: college football, weighting expert opinion, MCMC

*The authors would like to acknowledge Patrick L. Larkey, whose curiosity motivated this article.

1 Introduction

In 1998, the Bowl Championship Series was implemented in order to ensure that the top two Division I-A college football teams in the country faced each other in a post-season bowl game. Previously, the national champion was crowned based on the agreement of two human polls – the Associated Press media poll and the Coaches’ poll – after the completion post-season play; when the two polls disagreed, “split championships” resulted. There was a belief that requiring the top two teams to play each other would reduce controversy over the assignment of a national champion (of course, controversy over which two teams are selected to play in that championship bowl game followed).

The methodology for the BCS has changed frequently since its inception, but has remained fairly stable since 2004. The BCS rankings now consist of three equally-weighted parts: a media poll component, a coaches’ poll component, and a computer poll component, and it is this current implementation that we will be considering.¹ Each component produces a “rating” – a cardinal number between 0 and 1 (unlike in “rankings,” which are considered to be ordinal).

In the human polls, ratings are determined by taking the ratio of votes earned (higher is better) to the maximum possible number of votes. The computer component is a composite of six computer polls; each poll ranks its top 25 teams from 25 (highest) to 1 (lowest), and ranks all other teams zero.² A team’s highest and lowest computer rankings are discarded and the sum of the remaining four is divided by 100; that average is the computer rating component for the team. The ordered average ratings of the three components produce the overall BCS rankings. For the remainder of this paper, a team is considered “BCS-ranked” for a particular week if it is among the top 25 overall BCS ratings.³

¹In 2005, the AP poll was replaced by the Harris Interactive College Football Poll.

²The six computer polls are Anderson & Hester, Billingsley, Colley Matrix, Massey, Sagarin, and Wolfe.

³One reviewer raises concern over treating ordinal rankings as numeric; because each individual voter is restricted to assigning teams score of 1 through 25, there is no way for the voter to indicate that he feels that the skill level between teams one and two is different than that of teams two and three. This distinction will matter less among the human polls, because the BCS uses the cumulative point totals, not just the rankings they induce; if two and three are much closer in talent, some voters might instead swap those two teams in their vote, resulting in closer point totals than one and two. This might be more of a concern for the computer component, which is only aggregating six “votes”. There are models designed to deal with this issue, such as modified isotonic regression.

This formula is described in detail in Table 1. These ratings are produced weekly starting in the middle of the season. The two highest-ranked teams in the final week are selected to play in the national championship game, and the winner of that game is crowned as the national champion, regardless of the performance of teams in other bowl games.⁴ Clearly, the rankings are very high-stakes.

Most statistical investigation of the BCS has involved proposing new algorithms for ranking teams, discussing what should and should not be included as inputs to those systems, and metrics for comparing computer ratings (Callaghan et al., 2007; Martinich, 2002; Coleman, 2005; Mease, 2003). A thoughtful discussion of the statistical and contextual issues surrounding the development of a ranking or rating system can be found in Stern et al. (2004). But designing a new system for ranking college football is not our focus. Instead, we investigate the performance of the BCS *as implemented* for the 2004, 2005, and 2006 seasons. The appeal of this approach is that it both informs us about the performance of existing components and provides a policy prescription. For example, although Martinich (2002) compared the accuracy of the 10 components that made up the BCS in the 1999 and 2000 seasons, it is not clear how, were it so inclined, the BCS should have incorporated the results. Should only the most accurate components be used for the overall BCS rankings? The top five components? Should the weight of a component be proportional to its accuracy?

The advantage of our analysis is that it simultaneously evaluates each component *and* provides clear direction on what effect the evaluation should have on the construction of the overall BCS rankings. We do so by investigating the following questions:

Should the three components of the BCS be equally weighted? If not, how should they be weighted?

To answer these questions, we use the overall ratings' predictive power to infer the optimal weighted average.

1.1 Motivation and defense of a predictive criterion

As Stern has often and convincingly stated, the major weakness with the BCS is its lack of a clear role for the computer component (Stern et al., 2004; Stern,

⁴For example, if BCS #1 plays BCS #2 in the championship and BCS #3 plays BCS#4 in a different bowl game, a BCS #3 routing of BCS #4 cannot earn them the national championship, even if many people felt BCS #3 should be in the championship game and even if the winner of the national championship won via sloppy, uninspired play.

Buchman and Kadane: Reweighting the Bowl Championship Series

School	Harris Interactive			USA Today			Computer Rankings							BCS Average
	Rank	Points	%	Rank	Points	%	AH	B	CM	M	S	W	%	
USC	1	2806	0.9933	1	1540	0.9935	25	25	24	24	25	25	0.990	0.9923
Texas	2	2725	0.9646	1	1492	0.9626	24	23	22	25	24	24	0.950	0.9591
Virginia Tech	3	2596	0.9189	3	1428	0.9213	23	22	20	21	22	23	0.880	0.9067

Table 1: **Rankings explanation.** This table replicates the rankings of the top three overall BCS teams for week one of the 2005 ratings. USC earned 2806 of a possible 2825 in the Harris Interactive poll, so their media rating is $\frac{2806}{2825} = 0.9933$. They earned 1540 of a possible 1550 in the USA Today poll, so their coaches component is 0.9935. They were ranked 25 by four of the computer polls and 24 by the remaining two, so after dropping one 24 and one 25, the computer rating is produced by dividing the sum of the four remaining polls by 100, returning $\frac{25+25+25+24}{100} = 0.99$. Finally, each rating is multiplied by a third and summed, producing the overall BCS rating of $\frac{1}{3} \cdot 0.9933 + \frac{1}{3} \cdot 0.9935 + \frac{1}{3} \cdot 0.990 = 0.9923$ for USC.

2006). Is its goal primarily to comport with the human polls, thus confirming them? History would suggest so, as the computer components are usually judged to be a failure whenever they differ from the human polls. But as Harville states, “a possible reaction is that if computer ranking systems are to be judged primarily on their ability to duplicate the results of the polls, then why bother to include them in the BCS?” (Stern et al., 2004).

But even when computers are not considered failures when they do not act just like human polls, there is disagreement as to what is and should be measured. One often discussed split is “prediction” versus “performance” (Stern, 2006; Martinich, 2002). Prediction ratings are those which reflect the rater’s opinion on how teams would fare if they played against each other in the future given their performance up to the present, and performance ratings are those which quantify the team’s past performance. This distinction matters beyond the world of algorithm authors; sportswriters also feel tension between rewarding past performance and being realistic about future performance Mandel (2007).

This distinction was recently made clear in the case of the University of Oregon Ducks. The Ducks were 8-1 going into their November 15, 2007 game against the University of Arizona, were ranked second in the BCS, and their senior quarterback, Dennis Dixon, was considered to be a contender for the Heisman Trophy (Dufresne, 2007). However, Dixon was injured early in the game and Arizona upset Oregon with a 34-24 win; it was later announced that Dixon was out for the season. The prediction versus performance question translates to whether Oregon should have been ranked with its 8-2 cohort in the next week’s ratings, or whether the ratings should reflect a recognition that the transformed Oregon team, less its star quarterback, would have been unlikely to earn that 8-2 rating on its own. (In fact, Oregon lost its remaining two regular season games.)

But these examples stand out because they are so rare. Occasionally a star player will be hurt, or suspended, or need to travel home to attend a parent’s funeral. But a team’s fortunes rarely change drastically for reasons unrelated to its play, and a rating system should not be designed around such events. While there are isolated cases in which it makes sense to distinguish rewarding performance from making prediction, *on average* any system that is supposedly about evaluating performance must also do a respectable job of prediction. If not, how can we claim that its assessment of performance is of any value?

Therefore we conclude it is unimportant to know to what extent any of the eight components of the BCS reflect a performance metric versus a prediction metric; whether explicitly designed as a prediction metric or not,

any reasonable rating system should have predictive power. As Martinich (2002) writes, “A rating scheme that, late in the season, retrodictively predicts correctly which teams would have won the first games of the season but does a poor job at predicting the upcoming games seems fundamentally deficient.”

1.2 Algorithm restrictions

An additional consideration is the BCS mandate against using margin of victory in the computer rankings. Unlike the human voters, which are free to use all available information in producing their rankings (subconsciously if not explicitly), after the 2001 season the computer components were forbidden from using the final score in their calculations in response to a belief that Oregon was mistakenly kept out of the championship game against the University of Miami because the computers too highly valued the large margin of victories that the University of Nebraska had earned over its season (Carey, 2002). One wonders how well the computers might have performed over the 2004-2006 span had the BCS responded in a way that was less severe than to forbid the use of margin-of-victory information altogether. In fact, the merits of the restriction were drawn into question the very next season, when it was widely held that USC was unfairly denied a spot in the championship game because the computers *did not* use margin of victory.

2 Model specification and data

Our model is concerned with inferring the best weights for the weighted average comprising the overall rankings. Specifically, the goal is to find $\underline{w} = (w_1, w_2, w_3)$, where w_1 is the weight on the media poll, w_2 is the weight on the coaches’ poll, and w_3 is the weight on the computer component, and $\sum_{i=1}^3 w_i = 1$; currently the BCS fixes these at $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$.

The likelihood shown in Equation 2 is motivated by considering all games in the 2004-2006 seasons ($n = 82$) in which two top 25 BCS-rated (for a particular week) teams play each other as separate independent Bernoulli(p) experiments⁵ where p is the relative weight of the overall BCS rankings for the two teams.

⁵i.e. coin flips

Specifically, define

$$p_{i,j,k} = \frac{\sum_{s=1}^3 w_s \pi_{i,k,s}}{\sum_{s=1}^3 w_s \pi_{i,k,s} + \sum_{s=1}^3 w_s \pi_{j,k,s}} \quad (1)$$

where k ranges over the eight weeks of the season in which the BCS releases rankings, i and j range over teams that were both ranked in week k and playing each other, s ranges over the three components (media poll, coaches' poll, and computer polls), w_s is the weight for component s , and $\pi_{i,k,s}$ is the rating for team i during week k under component s .

Then the joint likelihood is

$$\prod_k \prod_{(i,j)} (p_{i,j,k})^{v_{i,j,k}} (1 - p_{i,j,k})^{1-v_{i,j,k}} \quad (2)$$

where $v_{i,j,k}$ is 1 if i beat j in week k , 0 otherwise.

As an example, consider the contribution made to the likelihood by the Texas Tech versus Texas game on October 22, 2005. As of game day, the most recent BCS rankings were the 2005 week one rankings, based on games through October 15, 2005, and both Texas Tech and Texas were BCS-ranked, with overall rankings of seven and two, respectively. Texas Tech had Harris Interactive, Coaches' poll, and computer ratings of 0.6641, 0.6961, and 0.75; Texas had ratings of 0.9646, 0.9626, and 0.95. Thus for $i = \text{Texas Tech}$, $j = \text{Texas}$, $\sum_{s=1}^3 w_s \pi_{i,k,s} = w_1 \cdot 0.6641 + w_2 \cdot 0.6961 + w_3 \cdot 0.75$ and $\sum_{s=1}^3 w_s \pi_{j,k,s} = w_1 \cdot 0.9646 + w_2 \cdot 0.9626 + w_3 \cdot 0.95$. Because Texas beat Texas Tech, $v_{i,j,(2005,1)} = 0$.

Thus

$$\begin{aligned} & (p_{i,j,k})^{v_{i,j,k}} (1 - p_{i,j,k})^{1-v_{i,j,k}} \\ = & 1 - p_{i,j,k} \\ = & \frac{\sum_{s=1}^3 w_s \pi_{j,k,s}}{\sum_{s=1}^3 w_s \pi_{i,k,s} + \sum_{s=1}^3 w_s \pi_{j,k,s}} \\ = & \frac{0.9646w_1 + 0.9626w_2 + 0.95w_3}{(0.6641w_1 + 0.6961w_2 + 0.75w_3) + (0.9646w_1 + 0.9626w_2 + 0.95w_3)} \end{aligned}$$

would be the game's contribution to the overall likelihood.

2.1 Simulation

We proceed with a Bayesian model, with the posterior sample simulated via a Metropolis-Hastings with independence proposals (Tierney, 1994). Because the weights must sum to 1, we model the prior on the weights $\pi(\underline{w})$ as a Dirichlet distribution with parameters $(1, 1, 1)$, which corresponds to a uniform prior over the simplex. The proposal distribution is chosen to be the same as the prior distribution. The simulation was performed in the language R, and the data and the code can be found on StatLib (<http://lib.stat.cmu.edu/>). The results are drawn from 30,000 simulations from the posterior, with a burn-in of 3,000. The standard convergence assessments were made with the R package `boa` (Smith, 2005).

3 Results

The three-dimensional joint posterior density of the weights is represented below as a density over a simplex. The interpretation of the density over the simplex is as a bijection between all triplets $s \in \mathbb{R}^3$ such that $\sum_{i=1}^3 s_i = 1$ and the closure of an equilateral triangle. If each of the components of s were placed at the vertices of an equilateral triangle, t_s is the corresponding point within the triangle on which it would balance. A different interpretation is shown in Figure 1. If a gradient from 100% to 0% were extended from each vertex, t_s can be fixed using any two of the components of s , as we would expect given the summation constraint (Allen, 2002). The general idea is the closer density is to a vertex, the higher the weights corresponding to the component at that vertex.

The posterior density is given in Figure 2. Note that if we were to select the posterior mode as a point estimate, we would have $\underline{w} = (.576, .125, .299)$; in other words, relative to the current implementation, we would almost double the weight on the media poll, more than halve the weight on the coaches' poll, and slightly lower the weight on the computer poll. If we study the joint posterior more closely, we note that the contours of the posterior are not concentric around the mode. Instead, the contours are roughly constant along fixed values of the computer weight, w_3 . In other words, but for weights very close to the posterior mode, for a fixed w_3 the likelihood does not vary too much regardless of the combination of w_1 and w_2 .

This is not too surprising; the human polls often comport very closely with each other. This can be confirmed by considering the correlation between ratings: it is .995 between media and coaches; .823 between media

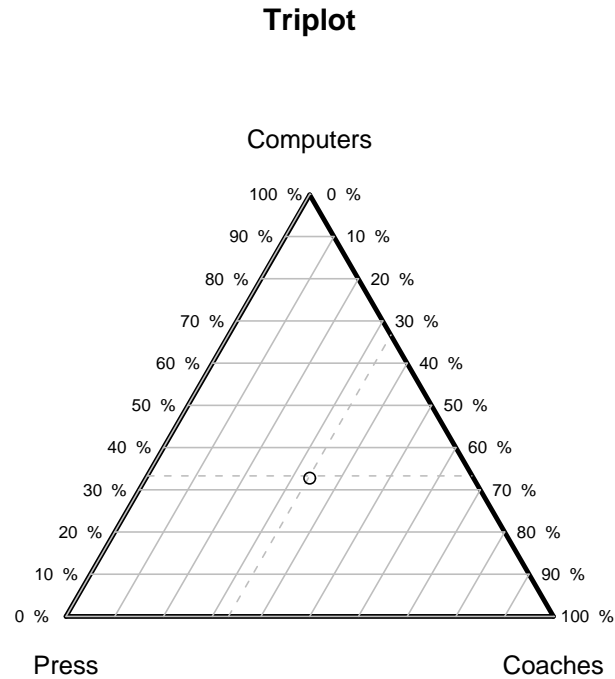


Figure 1: **The location of $\underline{w} = (1/3, 1/3, 1/3)$.** The figure above shows the gradient for computers, starting at 100% at its vertex with the percentages continuing down the left side of the triangle, and for coaches, starting at 100% at its vertex with the percentages continuing up the right side of the triangle. \underline{w} can be fixed with just two points, namely at the intersection of the 1/3 lines for computers and coaches (shown as the dashed lines). This point is clearly in the middle of the triangle, exactly where the “balancing” intuition would put it. Conversely, if one were given a point somewhere in the triangle, it is clear how to extract the associated weights.

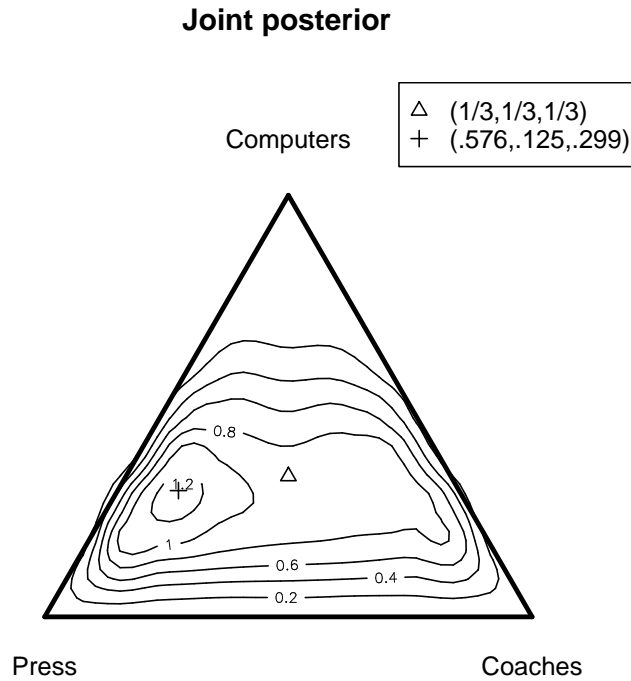


Figure 2: **The joint posterior density.** The + marks the mode of the joint posterior density of $(media, coaches, computers) = (.576, .125, .299)$, and the Δ marks the current BCS weighting of $(media, coaches, computers) = (1/3, 1/3, 1/3)$.

and computers; and .81 between coaches and computers. This suggests that coaches and media rate teams essentially the same, and when they do not the media poll makes the better departure on average.

4 Conclusion

The BCS will most likely always remain highly controversial; perhaps that is part of its appeal. Should the BCS move away from equally-weighted averages? Although we would put forward a point estimate of $\underline{w} = (.576, .125, .299)$, the posterior density for equal weighting $\underline{w} = (1/3, 1/3, 1/3)$ is reasonably high; at a density of .94, it is two-thirds the density of the posterior mode and falls within the 95% highest posterior density credible interval. We do not consider it urgent that the BCS adjust its weights.

References

- Allen, T. (2002). Using and Interpreting the Trilinear Plot. *Chance* 15, 29–35.
- Callaghan, T., P. J. Mucha, and M. A. Porter (2007). Random Walker Ranking for NCAA Division I-A Football. *American Mathematical Monthly* 114, 761–777.
- Carey, J. (2002, June 25). Another BCS computer ranking to drop margin of victory. *USA Today*.
- Coleman, B. J. (2005). Minimizing Game Score Violations in College Football Rankings. *Interfaces* 35(6), 483–496.
- Dufresne, C. (2007, November 16). Dixon, Oregon go down; Heisman-contending quarterback hurts knee in the first quarter and No. 2 Ducks' fortunes turn soon after in a 34-24 loss to Arizona. *Los Angeles Times*.
- Mandel, S. (2007). No. 1 ... for now. *College Football Power Rankings Blog*. Weblog post: http://sportsillustrated.cnn.com/2007/writers/stewart_mandel/10/16/power.rankings8/index.html.
- Martinich, J. (2002). College Football Rankings: Do the Computers Know Best? *Interfaces* 32(5), 85–94.
- Mease, D. (2003). A Penalized Maximum Likelihood Approach for the Ranking of College Football Teams Independent of Victory Margins. *The American Statistician* 57, 241–248.

- Smith, B. J. (2005, March 23). Bayesian Output Analysis program (BOA), version 1.1.5. <http://www.public-health.uiowa.edu/boa>.
- Stern, H. S. (2006). In Favor of A Quantitative Boycott of the Bowl Championship Series. *Journal of Quantitative Analysis in Sports* 2(1). Available at: <http://www.bepress.com/jqas/vol2/iss1/4>.
- Stern, H. S., K. Massey, D. A. Harville, and R. B. et al. (2004). Statistics and the College Football Championship. *The American Statistician* 58(3), 179–195.
- Tierney, L. (1994). Markov Chains for Exploring Posterior Distributions. *Annals of Statistics* 22, 1701–1762.