# College Football Rankings: Do the Computers Know Best?

By

Joseph Martinich

College of Business Administration

University of Missouri - St. Louis

8001 Natural Bridge Road

St. Louis, MO 63121

e-mail: joseph.martinich@umsl.edu

Final Version: May 7, 2002

Key Words: Recreation and sports

Statistics - Data analysis

## Abstract

The bowl championship series (BCS) committee uses ten ranking schemes, including eight computer rankings, to select college football teams for bowl championship series bowl games, including the national championship game. The large financial benefits of participating in BCS bowl games makes it imperative that the selection process accurately select the best teams. I evaluated the performance of the ten ranking schemes the BCS Committee used during the 1999 and 2000 seasons to select bowl teams. I found that almost all are equally accurate, but the *Seattle Times* scheme clearly underperforms the others. In addition, two proposed changes to the BCS selection formula, (1) to prohibit computer ranking schemes from considering the margin of victory in their rankings, and (2) to include explicitly the outcomes of head-to-head games among teams being considered for BCS bowls, could do more harm than good and could decrease the likelihood of the committee selecting the best teams for the BCS bowls.

This Fall 115 Division 1A college football teams (those colleges with the largest and most costly football programs) will compete for spots in 28 post-season bowls. Most teams are selected for bowl games based upon prearranged contracts with college football conferences; for example, the Alamo Bowl selects one team from the Big Ten conference and one from the Big 12 conference. Four of the bowl games, the Rose, Fiesta, Orange, and Sugar bowls, are affiliated with the bowl championship series (BCS), for which the participating teams are selected by the BCS committee. These four bowl games are the most prestigious of the bowl games, and they pay by far the largest amounts of money to participating teams, approximately $10 million per team. One of these games is designated the national championship game, which is intended to match the two best U.S. college football teams to determine a national champion. Selection to participate in a BCS bowl, especially the national championship game, is of great importance to U.S. colleges because of the immediate financial benefits, and because of the increases in financial contributions and student applications that result from participation. Since its inception the BCS committee has used human polls (by coaches and sports writers) and computer-based rankings as major elements in the selection formula. In recent years there has been considerable criticism of the selection procedure. Given the substantial financial implications, as well as the desire to select the best teams for the championship and other BCS bowls, it is imperative that the ranking systems included in the selection formula be the most accurate at ranking teams and that any bad ranking systems be dropped from consideration.

I initiated this research with the hypothesis that, because of biases, conflicts of interest, and a lack of knowledge (especially by coaches who do not see many other teams play during the

season), the *USA Today* /ESPN Coaches= Poll and, to a lesser degree, the AP Writers= Poll, would be inferior to more Aobjective@ computer rankings. Consequently, the large weight given to these two polls in the BCS selection formula might justifiably be reduced. In fact, I began this research with a preconceived title: AThe computers know what the coaches and writers don=t.@ A second hypothesis was that because a wide variety of approaches are used in the computer rankings, one or two computer ranking schemes would probably stand out from the rest as clearly superior or clearly inferior to the rest.

During the final stages of this research, in the Summer of 2001, two additional issues became of interest because of proposed changes to the BCS selection formula. One proposal was to incorporate explicitly the results of head-to-head games among the top teams as part of the bowl selection rules. This proposal was motivated specifically by the fact that Florida State University was selected for the January, 2001 championship game rather than the University of Miami, even though the teams had the same won-lost record and Miami had beaten Florida State in a head-to-head game. The second proposal was to drop computer rankings from the selection formula that gave too much weight to margin of victory. Specifically, the BCS committee considered requiring the developers of the computer ranking schemes either to eliminate margin of victory in computing their rankings or to cap the margin of victory at some level, such as 14 to 20 points. The rationale was to eliminate any incentive for teams to run up the score in blow-out games. Ranking schemes that did not adhere to this requirement would be dropped from the BCS formula. Although the basis for this requirement is sound - to discourage poor sportsmanship and to reduce the risk of injury to first-string players in games that have already been decided - the question is whether this would be effective. The goal of the BCS formula is to identify and

match the best teams at the end of the season. If the changes to the ranking schemes resulted in inaccurate rankings, the cure might be worse than the disease. These two proposals could have a substantial impact on the selection process, so the research was expanded to address them.

My research results appear to resolve the two hypotheses, although my first hypothesis was not supported. No single ranking system stands out clearly at the top. Almost all the rankings used in the BCS formula, including the coaches= and writers= polls, are approximately equally good, except for one clearly inferior ranking system (that of the *Seattle Times*) and one slightly inferior one (that of the *New York Times*). Because the coaches= and writers= polls are about as good as the best computer ranking systems, and they represent the thinking of 130 so-called experts (59 coaches, chosen by *USA Today* and ESPN, and 71 writers, chosen by the Associated Press) rather than eight computer rankings, assigning these polls higher weight in the BCS formula can be justified. (Currently the average of the writers= and coaches= polls receive the same weight as a modified average of the eight computer polls.) The distinctly poor performance of the *Seattle Times* computer ranking supported my second hypothesis, that one or two schemes would be noticeably inferior, and it also shed light on the margin-of-victory proposal. Of the eight computer rankings the BCS used in 1999 and 2000, the *Seattle Times* ranking was the only one that did not consider the margin of victory in its methodology. So if the BCS committee prohibited using the margin of victory in computer rankings, it would be mandating revisions that would make the rankings less accurate. (Yet this would not totally eliminate margin-of-victory considerations from the BCS formula, because the writers and coaches could still consider that factor in their polls.) A better alternative appears to be revising the ranking schemes to put a cap on margin of victory. Some previous research indicates that

using a reasonable cap on margin of victory causes only a slight loss in accuracy, and in fact, some of the BCS computer schemes already do this, and they appear to be as accurate as those that do not. Finally, my research shows that any head-to-head game criterion must consider two crucial issues: possible intransitivity of game outcomes and home-field advantage.

The BCS committee has modified its methodology several times already in the past four years, for example, by adding additional computer rankings, by including a strength of schedule measure, and by including the number of losses in the formula. My main conclusion from this research is that the BCS constantly needs to monitor the ranking schemes used in its selection formula, to revise the formula as the data warrant, but to be careful about changing the formula based on good intentions unless the data indicate that the cure is not worse than the disease.

**Background and Previous Literature**

Researchers have proposed many ranking schemes for a variety of competitions over the past few decades, for example, Bassett (1997), Elo (1986), Knorr-Held (2000), Leake (1976), Stern (1995), Wilson (1995b). Many seem to have been motivated by college football to develop ranking schemes, possibly because they think the standard polls do not rank their favorite teams high enough, or they just like the challenge of it. The ACollege football ranking comparison@ Website, http://www.mratings.com/cf/compare.htm, maintained by Kenneth Massey, currently tracks over 70 polls and rankings for college football.

For the most part computer ranking systems fall into two categories: those based on optimization or statistical models and those based on partially or completely subjective

heuristics. The former category includes models based on least-squares estimation, linear programming, paired comparisons, maximum-likelihood estimation, and neural networks (Bassett 1997, Harville, 1977, Leake 1976, Stern 1995 , and Wilson 1994, 1995a). Researchers have developed numerous ranking schemes from these general models by modifying the game outcome functions (for example, will the outcome of a game be reflected by the point differential, a function of the point differential, or will the outcome of a game simply reflect victory or defeat); by including home-field advantage; and by providing differential weights to game data based on the recency of the games. Although these models are based on clearly stated criteria, the authors proposing them have typically supported their use by simply presenting the rankings that would have resulted from their use in a specific year and sometimes showing their similarity to or difference from the writers= or coaches= polls ( Leake 1976, Wilson 1995a, 1995b). The implication is that the rankings resulting from the proposed scheme are obviously superior to those commonly used (the writers= and coaches= polls) or that the proposed computer scheme is almost as accurate as the writers= and coaches= polls and therefore should be used because of its objectivity or transparency. Some authors have presented data on prediction accuracy for alternative versions of their ranking schemes to demonstrate the effects of using different outcome functions (Harville 1977, Stern 1995 ) or of including a home field advantage (Stefani 1980), but with the exception of Stefani (1977), they have generally shied away from comparing the accuracy of their methods with those of competing schemes.

Ranking schemes based on ad hoc or subjective heuristics have generally not been documented precisely in publicly available literature, possibly because of the supposedly proprietary nature of the rankings (which may be used for gambling recommendations), but more

likely because their subjective nature makes them difficult to describe precisely and because they may constantly change. Some of these schemes, such as those of Richard Billingsley, the *Seattle-Times*, and the *New York Times*, are based largely on numerical formulas, but the exact number, form, and rationale of the formulas are not published. (Bob Kirlin nicely satirizes the proliferation of so-called computer ranking schemes based on subjective methods in his article, AHow to fake having your own math formula rating system to rank college football teams@ (http://www.cae.wisc.edu/~dwilson/rsfc/ history/kirlin/fake.html.). In spite of this lack of public documentation of methods, several of the computer polls the BCS formula includes are based on subjective, heuristic methods. The only rationale for including these methods would be if they are more accurate than the public objective computer methods and the longstanding writers= and coaches= polls.

**The Evaluation Criterion**

Probably the key issue in comparing ranking schemes is the criterion used for measuring performance. Hopkins (1997) distinguishes between measures of Aprediction@ (how well the ranking scheme predicts future game outcomes) and measures of Aretrodiction@ (how well the scheme predicts or explains past game outcomes, which were used in creating the rankings). Hopkins argues for using a retrodictive measure, such as the percentage of game outcomes consistent with the rankings (that is, the percentage of already-played games in which the higher ranked team defeated the lower ranked team). One problem with this approach is that at the time the game was played the rating scheme may have rated team A higher than B, but after B wins

the game, it may rate B higher than A, in which case the scheme gets credit for correctly predicting the outcome of the game on an ex post basis. Optimizing retrodictive measures, for example, by minimizing least squared error or maximizing the number of correct ex post outcomes, forms the foundation of many rating schemes. However, using such a measure to compare ranking schemes becomes tautological because once we have specified a retrodictive measure, we can usually derive mathematically the ranking that will optimize that measure for a specified set of contests or games. Also, using retrodictive measures presents the same problems in comparing football ranking schemes and in comparing business forecasting models. A forecasting model that explains or fits the past data well often fails to forecast the future well. Retrodictive measures may be useful in constructing and screening forecasting models, but the appropriate way to compare forecasting models is to compare their accuracy in forecasting the future (typically by using hold-out data), not simply by comparing their goodness of fit with respect to the data which were used to construct the model.

The fundamental issue, as presented by Stern (1995), is whether the goal of the rankings is to determine which team would most likely win in a head-to-head matchup or simply to reward season-long performance, including difficulty of schedule. Should the ranking identify the best team or teams at that time or the best over the entire season? If we look at how and why people use football ranking schemes (and why they are so widely published), it is that they want to measure how good teams are at that point in time and to predict the outcomes of upcoming games. Most people would expect a good ranking scheme to have the property that a higher ranked team should be expected to defeat a lower ranked team at that time on a neutral field. Specifically, the final prebowl BCS ranking is intended to represent the quality of the teams at

that time, and the goal is to select the two best teams at that time for the bowl championship (and to select at-large teams for the other BCS bowl games). 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. For these reasons the criterion I used for this research was the percentage of outcomes of immediately upcoming games correctly predicted by the ranking scheme (Bassett (1997) calls these Aout-of-sample forecasts@). A ranking scheme that is most accurate at prediction may be based on optimizing a retrodictive measure, such as least squares or percentage of game outcomes correctly predicted retrodictively, and may, in fact, be the best scheme as measured by retrodictive criteria, just as some forecasting models are good at both prediction and retrodiction. (However, many of the computer rankings the BCS uses have a retrodictive accuracy of well over 90 percent for games involving top 25 teams but a predictive accuracy of only 70 to 75 percent for the same teams.)

**Games Used in the Evaluation**

In this research I wanted to compare the performance of the ranking schemes the BCS used in selecting bowl teams. Prior to the 1999 season, it expanded the number of computer rankings it used from three to eight: those developed by Billingsley, Dunkel, Massey, Matthews, the *New York Times*, Rothman, Sagarin, and the *Seattle Times*. (The BCS gives a modified average of these eight rankings the same weight in the selection formula as the average of the

coaches= and writers= polls.) It has used this set of rankings for only the past two seasons, 1999 and 2000, so I performed my analysis using data for the 1999 and 2000 seasons.

I had to consider two major issues in deciding which games from these two seasons to use for comparison purposes.

(1) Several of the computer models require enough played games to connect all the teams (that is, enough games (arcs) to connect all the competing teams (nodes) to form a connected graph). In fact, some rankings are not published until this is the case, which usually takes five weeks. Even ranking schemes that do not require connectness need a reasonable amount of evidence from the current season; otherwise, they would rely on performance in previous seasons or be essentially subjective. Therefore, I made comparisons only using games played during the last six regular weeks of each season (games beginning on October 21 in the 1999 season and October 19 in the 2000 season), plus the conference championship week, plus the bowl games. (Stern (1995) used a similar approach; he used games from the second half of the season to evaluate his model for professional games.)

(2) Ranking systems do not all rank the same number of teams. Some systems (for example, Dunkel) rank all college football teams at any level (over 700 teams). Others rank only Division 1 teams or Division 1A teams; and some rank only the top 25 or the top 50 Division 1A teams. Especially troublesome is that the coaches= and writers= polls officially rank only the top 25 teams but in fact include all teams that receive votes, so from week to week these two polls typically rank from 35 to 40 teams, sometimes more, sometimes fewer. It would be unfair to

evaluate one ranking method using only games played by the top 25 teams because it ranks only those teams, while evaluating another scheme using all games played by Division 1 teams. Clearly, we would expect a higher accuracy rate in the former case because in most games, a top-level team would play a low-level team, whereas in the latter case a larger proportion of games would be between closely matched teams. I wanted to compare a common number of teams each week, but I also wanted to use as many games as possible. I decided to make two sets of comparisons: (1) using all games played by the top 25 teams each week, because all polls ranked the top 25, and (2) using all games played by the top 35 teams each week because all but the *New York Times* ranked the top 35. The *New York Times* published only its top 25 rankings until midway through the 2000 season, when it expanded to the top 50. The coaches= and writers= polls always ranked at least 35 teams except for a few weeks when only 34 received votes.

Some might disagree with the sets of games I used in this research, but I think it is a fair set, and I had no conscious bias in selecting the games.

**Accuracy of the Rankings**

For my first comparisons, I computed the percentage of game outcomes correctly predicted by each ranking method (Table 1). Specifically, for each of the weeks and teams considered, I determined the number of games in which the higher ranked team defeated the lower ranked team, using the rankings published the preceding Monday. (I made no adjustment for home field advantage.)

---

Table 1 Here

---

For games between two top 25 teams, the Dunkel and Sagarin rankings were the best, with accuracies of 72.7 and 72.4 percent, respectively. Except for the *Seattle Times* ranking, all of the computer rankings had higher accuracies than the writers= and coaches= polls. However, we should not read too much into this because the number of games played between two top 25 teams is relatively small. In looking at the accuracies for all games played by Top 25 teams, I found that eight of the ten rankings were almost indistinguishable, with accuracies ranging between 75.2 and 76.9 percent. The *New York Times* was slightly below these eight at 74.1 percent. But once again, the *Seattle Times* was clearly below the other nine rankings, with an accuracy of only 70.6.

Extending the population of teams and games to those ranked in the top 35, I got similar results (Table 2). For games between two top 35 teams, the Rothman ranking was the best, with 69.8 percent accuracy, and all the other rankings, except for the *Seattle Times*, had accuracies between 63.2 and 66.7 percent. Again the *Seattle Times* accuracy is well below the rest, at 60.0 percent. For all games played by top 35 teams, the writers= poll was most accurate, but the rankings were amazingly close, except for the *Seattle Times*. Seven of the rankings were less than 1 percent apart in their accuracies. The Billingsley ranking was slightly below the rest, while the *Seattle Times* was again substantially below the rest. Although the accuracy orderings among the ranking methods changed slightly depending upon whether I used games played by the top 25 or the top 35 teams, the conclusion appears to be the same: all the ranking schemes

were approximately equally accurate, except for the *Seattle Times* (and possibly the *New York Times*).

---

Table 2 Here

---

To some extent the similarity in performance among the ranking methods is not surprising. A large majority of the games used in Tables 1 and 2 were clear mismatches between high-ranked teams and low-ranked teams, for which all the ranking schemes predicted the same winner. (In fact, a person with only a modest knowledge of college football could probably correctly predict the outcome of games played by the top 25 or 35 teams with 65-70 percent accuracy.) A large portion of the time the schemes were correct, but sometimes there were upsets, in which case all the schemes failed to predict the upset. Games for which all the rankings agreed on the predicted outcome provided no information to differentiate among the accuracies of the ranking methods. Comparing only games for which the ranking schemes did not agree on the predicted outcome would be more informative. Consequently, I also looked at only those games where at least one ranking scheme disagreed with the others. The main drawback of this approach is that it drastically reduces the set of games used for the comparison. I evaluated games for which at least eight of the 10 schemes ranked at least one of the teams in the top 35 (that is, I could infer a game prediction), and at least one of the schemes made a prediction different from the other schemes.

---

Table 3 Here

---

The results showed a more pronounced difference in performance among the schemes than those of previous comparisons (Table 3). For games for which rankings disagreed in predicting the winner, seven of the schemes were very close in accuracy, correctly predicting over 60 percent of the outcomes. The *New York Times* and Dunkel rankings were somewhat less accurate with accuracies slightly above 50 percent. On the other hand, the *Seattle Time*s was clearly out of step with the other rankings in these games, correctly predicting the winners in fewer than 43 percent of these games.

An obvious question is whether the *Seattle Times*= difference in performance from the other schemes is statistically significant. Answering that question is difficult because the sets of games predicted by each ranking scheme summarized in Tables 1 and 2 are neither independent nor totally identical for each ranking scheme. So I could use neither standard tests based on independent samples nor tests based on paired samples. The games I used in the comparison were determined by which teams each scheme ranked in the top 25 or 35. However, by limiting the set of games to only those played by teams that were ranked in the top 35 by all the schemes, the data would now be implicitly paired, and I could use a test of proportions for paired observations (McNemar (1962, pp. 52-54)). For this set of games the difference in prediction accuracy between the *Seattle Times* ranking and the top seven other rankings was statistically significant at the 0.05 level. The only case tested that was not significant was between the *Seattle Times* ranking and the Dunkel ranking. I did not compare the *Seattle Times* and *New York Times* because the latter did not report top 35 teams for the entire two years. These statistical results strengthen the claim that the performance of the *Seattle Times* ranking has been inferior to that of the other schemes.

**Head-to-Head Games and Home-Field Advantage**

At the end of the 2000 regular season, the BCS formula=s selection of Florida State as the second-ranked team (after Oklahoma), thereby giving Florida State a spot in the national championship game, created considerable controversy. Some people argued that because Florida State and the University of Miami had identical 10-1 records (10 wins and one loss), but Miami beat Florida State in a head-to-head game, Miami should have been ranked ahead of Florida State (which it was in the coaches= and writers= Polls). Taking this argument further, others claimed that the University of Washington should have been ranked second because it was also 10-1, and had beaten the University of Miami, which had beaten Florida State. What is missing in these arguments is that these head-to-head games were not played on neutral fields. Miami beat Florida State by three points, but the game was played at the University of Miami, and Washington beat Miami by five, but the game was played at Washington. It has been mathematically demonstrated that there is a home-field advantage of approximately two to five points, and in fact, most of the BCS computer rankings claim to consider home-field advantage in their computations, although only Sagarin and Massey publish their home field adjustments explicitly. (At the end of the 2000 season, Sagarin had the average home-field advantage as 4.4 points, while Massey, which adjusts on a team-by-team basis, had the home-field advantage for Miami and Washington at just under 3.0 points.)

From a predictive point of view, including home-field information in computing rankings seems well justified. For example, Stefani (1980) found that including a home-field advantage factor in least-squares prediction models improved their accuracy for professional games by 1.9

percent and for college games by 1.0 percent (over four seasons of data). I tried to verify the effect of including home-field information in predicting game outcomes using the Sagarin ratings. In the results I report in Tables 1 to 3 for the Sagarin system, I ignored home-field information; that is, I considered the higher ranked team the predicted winner, regardless of where the game was played. I recomputed these results using the home-field-advantage adjustment reported by Sagarin each week. This adjustment changed the predicted winner in fewer than 20 of the more than 300 games played by top 35 teams during the 1999 through 2000 seasons. For games played by top 35 teams (those considered in Table 2), including the home-field adjustment increased the number of correct predictions by two, which increased the percentage of correct predictions from 74.3 to 74.9 percent (a record of 233-78). Looking only at games for which the ranking schemes disagreed (games considered in Table 3), including the home-field advantage increased correct predictions by one, thereby raising the prediction accuracy from 62.5 to 64.6 percent. Though these results are far from conclusive, they are consistent with previous research that indicated approximately a one to two percent improvement in predictive accuracy by including home-field advantage.

Given these results, I think that any bowl selection criterion based on the outcome of head-to-head games needs to consider the margin of victory and home-field advantage. The head-to-head games played by Florida State, Miami, and Washington seem too close to resolve the issue. One proposal made to the BCS committee in 2001 for teams that have played head-to-head, was to rank the losing team ahead of the winning team only if it had a superior won-lost record. (Apparently the potential for intransitivities has not been considered. For example, had Washington=s lone loss been to Florida State, a ranking paradox would have occurred.) Once

again, simply ruling that the losing team in a head-to-head match cannot be ranked ahead of the winning team, without regard for home-field advantage, could lead to inferior ranking schemes. Although the computer rankings typically include home-field information, it is not clear whether or how writers and coaches include this information in their decisions. For example, the writers= and coaches= polls just before the 2000 bowl games seemed to follow the proposed head-to-head rule in some cases but not in others. Both polls ranked Miami ahead of Florida State and Virginia Tech (teams with the same record as Miami that both lost to Miami at Miami); and ranked Washington ahead of Oregon State, whose only loss was by 3 points at Washington. But both polls ranked Miami ahead of Washington even though the teams had the same record and Washington had beaten Miami at Washington, which conflicts with the proposed rule.

**Margin of Victory**

One of the complaints made about the BCS formula=s choice of Florida State over Miami in 2000, was that seven of the eight computer rankings consider margin of victory, and that the BCS formula ultimately ranked Florida State higher than Miami because it piled up large victory margins in its games. However, this argument ignores the fact that *all eight* computer rankings, including the *Seattle Times,* which does not include margin of victory, ranked Florida State above Miami. In spite of this, during the Summer of 2001, the BCS committee was considering whether to require the computer rankings to eliminate margin of victory entirely from their ranking schemes so as not to reward the poor sportsmanship of running up the score. I assume that most all of the developers of the computer rankings the BCS uses have tested out the

benefit of including various factors in their models. Seven of them include margin of victory to some extent, an indication that including margin of victory helps accuracy, and some of the developers have stated this publicly and forcefully. In fact, the developers of published models have determined quite clearly that dropping margin of victory entirely from football ranking schemes results in the loss of useful predictive information. For professional games, Stern (1995) found that a model that included the margin of victory improved accuracy by two to three percentage points over one that considered only who won. Harville (1977) found essentially the same result. My results (Tables 1 through 3) support this earlier research. The only computer ranking that does not consider margin of victory, the *Seattle Times*, was the only one that was clearly inferior in its predictive accuracy. It is therefore ironic that the BCS would propose using only the type of scheme that least accurately ranks the quality of teams.

A far superior alternative is to incorporate a margin-of-victory cap in the computer rankings; that is, victories by more than x points would be treated the same as a victory by x points. This would discourage teams from running up the score. In addition, the data indicate that if the capping (the selection of x) is done carefully, the ranking has almost no loss in predictive accuracy. For example, Harville (1977) found that capping the margin of victory at 15 points had almost no effect on predictive accuracy (from 0 to 0.4 percentage points, depending on the pool of games used). Stern (1995) similarly found that, with his least-squares approach, using an adjustment to reduce blow-out scores had almost no effect on forecast accuracy for professional games (63.6 versus 63.5 percent accuracy). This finding conforms to the beliefs of some of the developers of the BCS computer rankings. Several of them were already capping the margin of victory in their models (though they don=t reveal the details), while others were willing to make

this change if the BCS required it. In August 2001 (while this paper was under review), the BCS committee accepted the arguments of the computer ranking developers and decided to allow the computer rankings to use margin of victory in their models as long as it was capped at a reasonable level (20 points). The developers of two of the computer rankings, Dunkel and the *New York Times*, refused to change their models and were replaced by the BCS in August, 2001. It is ironic that the BCS retained the ranking that was clearly the least accurate, the *Seattle Times*, simply because it did not include the margin of victory, whereas it replaced the far more accurate Dunkel scheme for not changing the methodology that contributed to its accuracy.

**Conclusions**

In my research I addressed several issues. First, I showed that the writers= and coaches= polls were as accurate in predicting future outcomes of college football games as were the six or seven best computer models the BCS used. Although the writers, coaches, and computer-model developers may be using different information and processing it differently, they seem to produce equally accurate rankings. Second, not all computer rankings are equally good. The *Seattle Times* ranking is clearly inferior to the other ranking schemes. The most pronounced difference in methodology between the *Seattle Times* rankings and the other nine was that it did not include margin of victory. Third, the BCS committee should keep in mind the old saying, AThe road to hell is paved with good intentions.@ My results indicate that the committee=s desire to discourage teams from running up the score, and to avoid rewarding them for it, could cause bigger problems if it does not implement the right solution. Capping the margin of victory

considered by the computer rankings appears to address the problem well, without harming the accuracy of rankings, whereas eliminating margin of victory completely as a factor in such rankings could have serious harmful consequences. Fortunately the BCS committee accepted this approach. However, the BCS committee continues to receive proposals from interested parties to change the selection procedure. For example, some people still want the committee to establish rules concerning the ranking of teams that met head to head. I found that any such rules should consider home field advantage as well as possible intransitivities; otherwise, the result could be worse than the apparent problem. In August 2001 the BCS committee decided not to establish explicit rules concerning the ranking of teams that meet head to head, but instead has added a bonus component for victories over highly ranked opponents. Unfortunately, this bonus component does not consider home field advantage.

## References


Bassett, Gilbert W. 1997. Robust sports ratings based on least absolute errors. *The American Statistician*, 51(2) 1-7

Elo, Arpad E. 1986. *The Rating of Chess Players Past and Present*. 2nd edition, Arco Publishing, New York.

Harville, David. 1977. The use of linear-model technology to rate high school or college football teams. *Journal of the American Statistical Association*, 72(358) 278-289.

Hopkins, Mark. 1997. Letter on rating systems comparison, on Website www.cae.wisc.edu/~dwilson/rsfc/rate/hopkins2.txt, letter dated Nov. 1, 1997, accessed May 6, 2002.

Kirlin, Bob. 2000. How to fake having your own math formula rating system to rank college football teams. Website http://www.cae.wisc.edu/~dwilson/rsfc/ history/kirlin/fake.html, accessed May 6, 2002.

Knorr-Held, L. 2000. Dynamic rating of sports teams,@ *The Statistician*, 49(2) 261-276.

Leake, R.J. 1976. AA Method for Ranking Teams: With an Application to College Football,@ in *Management Science in Sports*, ed. R.E. Machol et. al., Amsterdam, North-Holland Publishing., pp. 27-46.

Massey, Kenneth. 2002. The college football ranking comparison website, www.mratings.com/cf/compare.htm, accessed May 6, 2002

McNemar, Quinn. 1962. *Psychological Statistics*, 3rd edition. John Wiley and Sons, New York.

Stefani, Raymond T. 1977. Football and basketball predictions using least squares. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-7(2) 117-121

_____ 1980. Improved least squares football, basketball, and soccer predictions. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-10(2) 116-123

Stern, Hal. 1995. Who=s number 1 in college football?...and how might we decide? *Chance*, 8(3) 7-14

Wilson, R. L. 1994. A neural network approach to decision alternative prioritization. *Decision Support Systems*, 11(5), 431-447.

_____, 1995a. The >real= mythical college football champion. *OR/MS Today,* 22(5) 24-29

_____. 1995b. Ranking college football teams: a neural network approach. *Interfaces,* 25(4) 44-59

**Table 1**: I determined the prediction accuracy of the ten ranking methods used by the BCS for games played by top 25 teams for the 1999 and 2000 seasons. The first row, top 25 vs top 25, shows the number of game outcomes the ranking methods predicted correctly and incorrectly for games played between two teams ranked in the top 25 by that ranking. The third row, 25 vs UR, shows the number of game outcomes correctly and incorrectly predicted in which only one team was ranked in the top 25 by that ranking. The fifth row, all top 25, shows the number of game outcomes predicted correctly and incorrectly for games in which at least one team was ranked in the top 25 by that ranking. The ranking methods were the AP writers= poll (AP), the *USA Today* /ESPN coaches= poll (US), and the Billingsley (Bill), Dunkel (Dun), Massey (Mas), Matthews (Math), *New York Times* (NYT), Rothman (Roth), Sagarin (Sag), and *Seattle Times* (ST) computer rankings. The number of games evaluated for each method is not the same because the methods did not all rank the same teams in the top 25.

| | AP | US | Bill | Dun | Mas | Math | NYT | Roth | Sag | ST |
|---|---|---|---|---|---|---|---|---|---|---|
| top 25 vs top 25 | 41-24 | 41-23 | 45-23 | 40-15 | 38-16 | 45-20 | 38-20 | 42-20 | 42-16 | 37-27 |
| % correct | 63.1 | 64.1 | 66.2 | 72.7 | 70.4 | 69.2 | 65.5 | 67.7 | 72.4 | 57.8 |
| 25 vs UR | 138-35 | 142-35 | 130-32 | 148-44 | 147-43 | 130-37 | 136-40 | 141-35 | 144-43 | 129-42 |
| % correct | 79.8 | 80.2 | 80.2 | 77.1 | 77.4 | 77.8 | 77.3 | 80.1 | 77.0 | 75.4 |
| All top 25 | 179-59 | 183-58 | 175-55 | 188-59 | 187-59 | 175-57 | 180-63 | 183-55 | 186-59 | 166-69 |
| % correct | 75.2 | 75.9 | 76.1 | 76.1 | 76.0 | 75.4 | 74.1 | 76.9 | 75.9 | 70.6 |

**Table 2**: I determined the prediction accuracy of the ten ranking methods used by the BCS for games played by top 35 teams for the 1999 and 2000 seasons. The ranking methods were the AP writers= poll (AP), the *USA Today* /ESPN coaches= poll (US), and the Billingsley (Bill), Dunkel (Dun), Massey (Mas), Matthews (Math), *New York Times* (NYT), Rothman (Roth), Sagarin (Sag), and *Seattle Times* (ST) rankings. The first row, top 35 vs top 35, shows the number of game outcomes the ranking methods predicted correctly and incorrectly for games played between two teams ranked in the top 35 by that ranking. The third row, 35 vs UR, shows the number of game outcomes correctly and incorrectly predicted in which only one team was ranked in the top 35 by that ranking. The fifth row, all top 35, shows the number of game outcomes predicted correctly and incorrectly for games in which at least one team was ranked in the top 35 by that ranking. The number of games evaluated for each method is not the same because the methods did not all rank the same teams in the top 35.

| | AP | US | Bill | Dun | Mas | Math | NYT | Roth | Sag | ST |
|---|---|---|---|---|---|---|---|---|---|---|
| top 35 vs top 35 | 63-32 | 60-35 | 71-38 | 64-32 | 66-34 | 66-36 | N/A | 74-32 | 66-34 | 60-40 |
| % correct | 66.3 | 63.2 | 65.1 | 66.7 | 66.0 | 64.7 | N/A | 69.8 | 66.0 | 60.0 |
| 35 vs UR | 183-51 | 178-47 | 149-44 | 171-50 | 168-48 | 166-44 | N/A | 154-46 | 165-46 | 162-53 |
| % correct | 78.2 | 79.1 | 77.2 | 77.4 | 77.8 | 79.0 | N/A | 77.0 | 78.2 | 75.3 |
| All top 35 | 246-83 | 238-82 | 220-82 | 235-82 | 234-82 | 232-80 | N/A | 228-78 | 231-80 | 222-93 |
| % correct | 74.8 | 74.4 | 72.8 | 74.1 | 74.1 | 74.4 | N/A | 74.5 | 74.3 | 70.5 |

**Table 3**: I determined the prediction accuracy of the ten ranking methods used by the BCS for games played by the top 35 teams during the 1999 and 2000 seasons. The ranking methods were the AP writers= poll (AP), the *USA Today* /ESPN coaches= poll (US), and the Billingsley (Bill), Dunkel (Dun), Massey (Mas), Matthews (Math), *New York Times* (NYT), Rothman (Roth), Sagarin (Sag), and *Seattle Times* (ST) rankings. I only considered games for which the ranking methods did not all predict the same team to win. For 49 games at least eight ranking schemes made inferred predictions. All but the *New York Times* made predictions in at least 48 of them. The *New York Times* made predictions for only 43 of them, because it did not publish rankings beyond the top 25 until the middle of the 2000 season. The first row, all top 35, shows the number of game outcomes the ranking methods predicted correctly and incorrectly for games in which at least one team was ranked in the top 35 by that ranking.

| | AP | US | Bill | Dun | Mas | Math | NYT | Roth | Sag | ST |
|---|---|---|---|---|---|---|---|---|---|---|
| All top 35 | 30-18 | 29-19 | 31-18 | 26-23 | 31-18 | 30-19 | 22-21 | 32-17 | 30-18 | 21-28 |
| % correct | 62.5 | 60.4 | 63.3 | 53.1 | 63.3 | 61.2 | 51.2 | 65.3 | 62.5 | 42.9 |