Proceedings of the 2005 Winter Simulation Conference
M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines, eds.

VALIDATING A DIVISION I-A COLLEGE FOOTBALL SEASON SIMULATION SYSTEM

Rick L. Wilson

William S. Spears School of Business
Department of Management Science and Information Systems
408 Business
Oklahoma State University
Stillwater, OK 74074, U.S.A.

ABSTRACT

NCAA Division I-A college football remains the only major collegiate sport not to have a true playoff to determine its national champion. A controversial and flawed process termed the Bowl Championship Series (BCS) has emerged with the goal of matching two teams in a so-called 'national championship' game. The BCS has employed a varied set of computer-based ranking models to help determine the two participants. The effectiveness of these computer models has never been empirically investigated in a controlled setting. This paper presents the development of a controlled test-bed of simulated college football seasons, a necessary preliminary step. A simulation model is developed and validated against past college football season data. Results show promise. It is hoped that this simulation process will allow future systematic study of the various mathematical models used to rank football teams.

1 INTRODUCTION

NCAA Division I-A college football remains the only major sport that does not utilize a playoff to determine its champion. There are a variety of reasons for this, not the least of which is money.

For years, voter polls determined the so-called 'Mythical National Champion'. A quick look through college football history will show many seasons where one voter poll (say sportswriters) had a different national champion than another voter poll (say coaches). This phenomenon has seemed to trouble football fans much more during the last 20 years.

One can hypothesize that the increased focus in modern society on being No. 1 and a series of controversial endings to the college football season in the late 1980s and early 1990s led influential people in college football to attempt to develop a process in which a single champion could be named. This attempt led to the creation of the Bowl Coalition process, which started in 1992. This process involved adding together the points received by teams in each of the two major voter polls to determine the top two teams, and matching them (if possible) in a major bowl game.

Unfortunately, not all major collegiate conferences participated in this process, the most notable omissions being the Big Ten (with its 11 football teams) and the Pac-10. This led to split or controversial national champions in 1993, 1994 and 1997. The continued controversy and the hope for a more robust method for selecting the participants led to the creation of the BCS in 1998, which included 'objective' measures of team strength by utilizing computer rankings such as the New York Times computer. However, controversy has been the rule rather than the exception since the BCS's inception, and it has seemingly changed its approach in determining the top two teams each year it has existed. The controversy has been so pervasive that congressional hearings were held during the last two years to discuss the 'BCS Mess'.

The root of the problem is the BCS leadership's failure to explicitly state criteria for determining the final two teams. For instance, should the championship game involve the best two teams during the entire season, or those teams who are strongest at the end of the season? Or is there some other objective or metric of interest?

Additionally, the mathematical models used in the BCS approach have often come under criticism and scrutiny. There has been a fair amount of research in the academic literature on college football ranking methods that date back to the 1950s (and perhaps farther - see Wilson 1995). Most of the published research has focused on how a particular approach would rank teams in a given season, and then argue, using some form of face validity, that one approach is superior to another.

As the BCS and academic researchers continue looking for the holy grail of ranking methodologies, one of the challenges faced is how to objectively determine which ranking method performs best. The fundamental concepts behind the different quantitative approaches can be examined, but most would be concerned with how 'good' a specific ranking approach appears to a knowledgeable college football fan. Thus, the major dilemma in systematically studying different ranking methods is that the 'true' best ranking is not explicitly known. Results of methods cannot be compared to some exemplar ranking because the 'true' strengths of the football teams are unknown.

A simulation system of a college football season would help mitigate this dilemma by creating a priori known team strengths and applying realistic game outcome mechanisms that allow for home field advantage, upsets and the general random nature of athletic contests; thus, allowing the systematic evaluation of ranking methodologies. This could lead to the simplification and improvement of the BCS system; hence, the motivation behind the work described in this paper.

2 RELEVANT PAST RESEARCH ON COLLEGE FOOTBALL GAME OUTCOMES

Three main studies help shape the development of the college football game/season simulation system described here. First is the 1991 research by Stern that deals with the use of point spreads, or betting odds, to determine the likely outcome of (primarily) National Football League games. The research stream that examines the efficiency of the gambling markets in college football betting is the second area used to develop the simulation system. Finally, a study spawned by Stern's research that attempted to find a simpler, more effective way of determining the strength of a team's schedule is also valuable to the system development. We will briefly review pertinent findings of these studies, which will help guide the development of the simulation system.

2.1 Stern's Law

Stern's 1991 study that deals with the probability of winning an NFL football game is a relevant paper, as this same approach has been proposed as valid for college football - although never explicitly validated (see Carlin and Stern 1999, Stern 1995).

Stern's empirically developed premise was that the probability of winning a football game is a random normal variable with a mean equal to the Las Vegas determined betting 'point spread' and a standard deviation of 14. The point spread is a surrogate indicator of the difference in team strength but, in practice, it also considers additional factors. It is well known that Las Vegas casinos set the odds, or point spreads, for football games such that the amount of money bet on each team is approximately equal. As those who accept bets receive a certain percentage of all winning bets (and keep all losing bets), this approach ensures steady Casino profits.

As an example of this premise, consider two teams - Michigan and Ohio State - and the Vegas odds have chosen Michigan as a seven-point favorite. Stern's premise would indicate that Michigan would have a $Z(7/14) = 0.6914$ or 69.14% chance of winning the game, while Ohio State would have a $1 - 0.6914 = 0.3086$ or 30.86% chance.

Stern's research is the foundation for using a normal distribution as a way of operationalizing the probabilistic outcome of college football games. Given some measure of team strength for two opposing teams, the difference of these strengths can be used (as Stern used point spreads) to determine a probability that each team will win. Thus, when implementing a simulation of a game, a random number determines game outcome according to this probability.

2.2 Point Spread vs. Actual Outcome

Numerous studies have appeared in the marketing literature that analyze the professional and college football betting markets (see Golec and Tamarkin 1991). Most studies have approached betting market analysis as a way of testing market efficiency, using the analogy with securities markets. Some previous research used statistical approaches such as regression to test for market efficiencies (see Gandar 1988), while others have explored specific betting strategies to determine if they lead to unusual profits.

In an attempt to gain additional insight into what determines market efficiency, a recent study using 1995-1997 college football data explored not only point spread data but also other factors such as home field location (Wilson 1998). Not surprisingly, results indicated that point spreads were a major determinant in actual game outcome and that other possible variables explained very little of the variance in game outcomes. A total of 48-55% of the point differential variance was explained by the regression models, with point spread being the only consistently significant variable. Residual tests also showed that the normality assumptions were met in the analysis as well.

While point spreads also involve perception of bettors as well as actual team strength, the lessons gained from this and other betting studies are twofold. First, we have a benchmark (48-55%) for the actual point variance that an a priori specified team strength measure might explain. Second, the results would provide further justification to use a normal distribution in describing game outcomes, specifically related to point differentials in games.

2.3 Modifying Stern's Law – Win/Loss Records

In an attempt to move away from point spread data, which is available only for games actually being played, research was undertaken to test whether Stern's approach could be applied to college football data when the difference of team win/loss records and home field information was used to assess the likelihood of a team winning a football game (Wilson 2002). It was posited that if team win/loss record differences described college football game outcomes similar to point spreads, then this information could be used to assess which teams played more difficult schedules and could be an objective way of determining a ranking at the end of the season.

College football game data for the 1999 through 2002 regular seasons were studied. Game outcomes and locations were noted. Games played against teams outside of Division I-A football were excluded in the analysis. The winning percentage of each team was calculated from the remaining games. The difference in winning percentage was calculated and tracked with whether the team with the better win-loss record was playing at home or away.

The game outcome information was tallied in bins. Each bin represents a range of win/loss percentage differences of the two teams playing. The Chi-Squared goodness-of-fit test was utilized and showed favorable evidence that game outcomes could be described using a random normal variable defined by the percentage difference of team records and a standard deviation of approximately 0.25.

This percentage result reinforced the use of a normal distribution in game outcomes and may also provide some additional validation of season performances related to team win/loss records for the developed simulation model.

3 SIMULATION MODEL DESCRIPTION

This section describes the implementation details of the simulation model.

3.1 College Football Team Structure

The simulation model was designed to mimic, as closely as feasible, certain aspects of today's Division I-A college football environment. The simulation used 120 teams very similar to the 118 teams that competed at the Division I-A level in 2004. Each year, more teams gain Division I-A status; therefore, 120 was a reasonable approximation.

An even number of teams was used, facilitating an easily repeatable schedule for the simulations. The schedule was designed to closely approximate existing BCS schedules. For instance, the 120 teams were divided into conferences of varying sizes - eight, 10 and 12 teams. Each team played at least seven conference games (obviously, some played eight), and the remaining were non-conference opponents. Each team in the simulation had exactly 11 scheduled games. In recent actual college seasons, the NCAA has allowed a 12th game for teams, as well as the inclusion of special pre-season games and post-season conference championship games. This leads to, in practice, teams playing an unequal number of games. Nonetheless, for simplicity, this simulation system has each team playing the same number of games.

Finally, each team was assigned either five or six home games. During most college football seasons, there is more variance among teams regarding home and away games, but this was not deemed an important difference.

3.2 Simulating a Season

The following process was used to simulate an entire college football season. First, team strengths were assigned randomly to the 120 teams. Second, each of the 660 games were 'played', based upon the fixed schedule. Since team strengths were randomly assigned each season, the fixed schedule was not problematic. Finally, results were tallied and the process was repeated for additional seasons.

One important input in the simulation model was the distribution and magnitude of team strengths. Many options for distribution specification exist. For this study, a normal distribution was utilized. There is presently no empirical evidence that suggests or favors one choice as opposed to another. Therefore, a uniform distribution of team strengths could be a viable implementation alternative.

The normal distribution was chosen because of the conventional wisdom that there are few really strong teams each year and that most teams are fairly equal (so-called parity) given today's scholarship limits. A normal distribution arguably implements this concept. Obviously, this simulation model assumption should be further evaluated and tested in the future.

Team strengths ranged from a high of 115 to a low of five. Discrete fixed values were used from season to season as an additional control. Mean team strength was set at 60 with a standard deviation of 21.5. Descriptive statistics on the fixed team strength values indicated that, even as discrete values, they closely approximated a normal distribution.

The range of team strengths was chosen purposefully, as it is used in determining expected game outcomes. The difference in team strengths was calculated, then divided by two (fractions were truncated). This gave the expected margin of victory (in points) of one team over another. An additional four points was added to the team strength of the home team, simulating the home field advantage, which is consistent with some college football computer ranking models. Then, using a normal distribution based upon past point spread data analysis, the actual outcome (actual point differential) was generated. The win/loss game outcome was determined. This process was repeated for all 660 season games.

3.3 Simulation Runs

The simulation was executed 500 times, resulting in a pooled set of 330,000 games. Data was collected for each season, as ranking methods will ultimately be applied to the simulation data to explore their efficacy. Individual season data was also used in the validation process (see next section). Aggregate data was summarized for the validation process and will be discussed in the next section.

4 RESULTS

Results of the 500 simulation runs are summarized below.

4.1 Face Validity Measures

The difficulty in validating simulation models of college football games/seasons was discussed earlier in this paper. The first evaluation uses the face validity dimension; in other words, are the results of the simulation logical? If not, there is no need to investigate additional detailed, deeper statistical relationships.

The first measure examined in the 500 seasons was the correlation between team strength and the number of wins. One would expect a fairly high positive correlation, although not perfect because there exists an inherent variability of game outcomes and teams play different strength opponents. Teams of similar strength might play very different schedules, which could lead to one team winning significantly more games than the other, even though their team strengths are the same.

The average rank correlation value was .82326 (Spearman's Rho). The range of the correlation values for each season throughout the 500 scenarios was between .7236 and .8907. The standard deviation was .02745. This seems to indicate reasonable results.

Next, the correlation between team strength and the average number of wins was aggregately calculated for the entire 500 seasons. For each season, the explicit number of wins for a given team strength was tallied then totaled for all 500 seasons. Ultimately, if 500 simulations reached a form of 'steady-state,' a correlation near 1.0 would be expected. As an infinite number of seasons approaches, one would expect that strong teams will, on average, win more games than lesser teams. The results did show this, with a 0.99523 correlation based upon raw team strength score value and a 0.99480 correlation using ranks.

Of additional interest was the variation in the number of wins of teams (on the basis of team strength) during the 500 seasons. Not unexpectedly, the teams on either end of the spectrum (either really strong teams or really poor teams) had lower variance in the number of games won than the teams 'in the middle'. For space purposes, the 120 row table is not reproduced here. However, the average standard deviation for the number of wins was approximately 1.7 for teams 'in the middle' of the strength distribution, but as low as 0.9 for the strong or poor teams.

4.2 Actual vs. Expected Point Spreads

Using past market efficiency studies as validation guidelines, the correlation between the expected game outcomes (in points) and the actual game outcomes for each of the 500 seasons was captured individually. The mean correlation was .6971, representing approximately 49% of variance explained in actual outcomes by the expected outcomes (expected outcome was described in Section 3.2). The range of the correlation spanned from .6364 to .7537 with a standard deviation of .0192.

Past studies had shown a 48-55% explanation rate, so the simulation system seems to provide reasonable results along this dimension.

4.3 Detailed Game Comparisons

Finally, the simulation results were compared to the study described in Section 2.3. This is the most complete and potentially insightful dimension of validation.

Table 1 shows the results of analyzing game observed data for the 1999-2002 football seasons. The data is segmented according to win/loss record differential and home team designation; additionally, the corresponding wins and losses are shown.

The first column shows the bins by which the data is tallied. (Recall the use of goodness-of-fit tests previously discussed). The number represents the maximum difference in win/loss percentage for that bin. These bins for actual data are not of equal size, as we attempt to 'match up' with the simulated data (shown in Table 2). As all teams play 11 games in the simulation, percentage differences will be a multiple of 1/11. This is not true in the real data sets (since teams play an unequal number of games), which complicates data comparisons.

The next two columns represent the wins and losses of the 'favored' home team - the situation when the home team has a better record (by the bin amount or range) than the visiting team. The win/loss percentage is shown next. In similar fashion, columns five, six and seven show the wins and losses when away teams have a better win/loss record, and then the corresponding win/loss percentage. The final two columns illustrate the cumulative number (and the percentage) of games for each successive bin.

Table 2 shows comparable data for the 500 simulation runs. The 'DIFF' column represents the difference in losses in records between the two teams involved in a game. Recall that all teams in the simulation play 11 games. Raw numbers of wins and losses are not repeated for the simulation. The home percentage, away percentage and total game percentage figures shown correspond to the fourth, seventh and ninth columns of Table 1.

Table 1: 1999-2002 Season Results

			HOME			AWAY		GAMES
DIFF	W	L	PCT	W	L	PCT	TOT	PCT
0.00	74		62.2%	45		37.8%	119	4.7%
0.12	190	101	65.3%	148	93	61.4%	651	26.0%
0.20	148	43	77.5%	145	62	70.0%	1049	41.8%
0.28	177	22	88.9%	144	26	84.7%	1418	56.6%
0.40	229	23	90.9%	156	22	87.6%	1848	73.7%
0.48	116	7	94.3%	109	3	97.3%	2083	83.1%
0.56	94	4	95.9%	66	4	94.3%	2251	89.8%
0.64	70	0	100.0%	63	1	98.4%	2385	95.1%
0.76	42	0	100.0%	38	0	100.0%	2465	98.3%
0.84	18	0	100.0%	15	0	100.0%	2498	99.6%
1.00	4	0	100.0%	5	0	100.0%	2507	100.0%

Table 2: Simulated Season Data

			TOTAL
	HOME	AWAY	GAME
DIFF	%	%	%
0	55.61%	44.39%	8.91%
1	68.18%	56.93%	26.36%
2	78.92%	69.56%	43.09%
3	86.74%	79.42%	58.27%
4	91.65%	87.44%	71.38%
5	95.27%	92.31%	81.88%
6	97.47%	95.64%	89.61%
7	98.76%	98.02%	94.79%
8	99.43%	99.47%	97.82%
9	99.83%	99.85%	99.33%
10/11	100.00%	100.00%	100.00%

The results between the two tables are similar. Caution must be used to interpret these results; but, with the exception of the first two bins, the winning percentages of the different combinations of win/loss differences and home/away designations are very similar between the real data and the simulated data.

Note that in the simulated data, there is a definitive difference between home team and away team performance (when win/loss percentage is held constant). This illustrates internal validity of the simulation model, as a four-point home field advantage has been added in each game. The real data shows a similar, though not as strong, relationship. An exception occurs at the '0.48' bin, but this can likely be explained by the relatively small number of actual games during the four seasons that fall into this bin.

Another area in which the tables are somewhat dissimilar is the first bin, where teams playing have the same win/loss record. Note that there are twice as many games in the simulated data than the real data (in terms of percentage) and that the winning percentages of the simulation are lower than the real data. The difference in the number of games in this bin is due to the uneven number of games played by teams in college football. For example, consider two teams - each with one loss, but one with 11 games and the other with 12. A game between these teams would not appear in the '0' value bin but in the '<0.12' bin. This circumstance would not occur in the simulated data. Ultimately, the total cumulative percentage games that fall collectively in the first two bins are similar between the real seasons and the simulated seasons.

The simulated data also shows a higher percentage of 'upsets' occurring when teams of highly different win/loss records play one another. Again, this might be explained by the fact that actual season data encompasses only four seasons of games.

To summarize these tables, the results of the simulation look very similar to that of the 1999-2002 actual game data when viewed through the win/loss, home/away result lens. If the simulation data was used as the population win/loss percentages, then any and all proportion tests performed on the 1999-2002 results would reject the null hypothesis, indicating there is no evidence that actual season results differ significantly from the simulated results. Thus, from a number of different validation perspectives, the simulation system appears to be representative of college football game outcomes.

5 CONCLUSION

In this paper, the initial results for validation of a simulation system of college football game outcomes (thus, college football seasons) have been presented. Because of the ill-defined nature of the environment under study, it is difficult to conclusively state that the system's results have been validated beyond doubt.

Nonetheless, it appears that along many dimensions, the simulation system and the utilized parameters do reflect an accurate depiction of the dynamics of college football outcomes. For the future, this system will hopefully provide a suitable test bed for Step 2 of this research stream – analyzing the efficacy of different mathematical ranking models.

Using this work as a starting point, perhaps the BCS system can be modified again to incorporate a criteria-based, rational approach to determining the two teams that play for the national championship. Regardless, this simulation system may offer a way of creating a predictable test bed of football game outcomes for the refinement of ranking methodologies in the future.

REFERENCES

Carlin, B. and H. Stern. 1999. Designing a college football playoff system. *Chance* 12(3): 21-26.

Gandar, J. et al. 1988. Testing market rationality in the point spread betting market. *Journal of Finance* 43: 995-1007.

Golec, J. and M. Tamarkin. 1991. The degree of inefficiency in the football betting market. *Journal of Financial Economics* 31: 311-323.

Stern, H. 1991. On the probability of winning a football game. *The American Statistician* 45: 116-123.

Stern, H. 1995. Who's number 1 in college football? … And how might we decide? *Chance* 8(3): 7-14.

Wilson, R. 1995. Ranking college football teams: A neural network approach. *Interfaces* 25(4): 44-59.

Wilson, R. 1998. "Brand and recency effects in the college football betting market." Proceedings of the annual meeting of the Decision Sciences Institute, Las Vegas, Nevada, 1998.

Wilson, R. 2002. "An improved measure of schedule strength for college football." Paper presented at the annual meeting of the Decision Sciences Institute, San Diego, California, 2002.

AUTHOR BIOGRAPHY

RICK L. WILSON, Ph.D., is the W. Paul Miller Professor of Business Administration and Professor/Head of the Management Science and Information Systems Department in the Spears School of Business at Oklahoma State University. His research interests include sports applications of operations research/management science and data mining. He is a member of INFORMS, Decision Sciences Institute, AIS and IACIS. His e-mail address is <rlwilsn@okstate.edu>.