# Extending the Colley Method to Generate Predictive Football Rankings

R. Drew Pasteur

**Abstract.** Among the many mathematical ranking systems published in college football, the method used by Wes Colley is notable for its elegance. It involves setting up a matrix system in a relatively simple way, then solving it to determine a ranking. However, the Colley rankings are not particularly strong at predicting the outcomes of future games. We discuss the reasons why ranking college football teams is difficult, namely weak connections (as 120 teams each play 11-14 games) and divergent strengths-of-schedule. Then, we attempt to extend this method to improve the predictive quality, partially by applying margin-of-victory and home-field advantage in a logical manner. Each team's games are weighted unequally, to emphasize the outcome of the most informative games. This extension of the Colley method is developed in detail, and its predictive accuracy during a recent season is assessed.

Many algorithmic ranking systems in collegiate American football publish their results online each season. Kenneth Massey compares the results of over one hundred such systems (see [9]), and David Wilson's site [14] lists many rankings by category. A variety of methods are used, and some are dependent on complex tools from statistics or mathematics. For example, Massey's ratings [10] use maximum likelihood estimation. A few methods, including those of Richard Billingsley [3], are computed recursively, so that each week's ratings are a function of the previous week's ratings and new results. Some high-profile rankings, such as those of USA Today oddsmaker Jeff Sagarin [12], use methods that are not fully disclosed, for proprietary reasons. Despite the different approaches, nearly all ranking methods use the same simple data set, the scores of games played during the current season. A few also use other statistics, such as yardage gained. College football's Bowl Championship Series (BCS), which matches top teams in financially-lucrative postseason games, including an unofficial national championship game, chooses its teams using a hybrid ranking, currently including two human polls and six computer rankings [5]. The use of victory margins in BCS-affiliated computer rankings was prohibited following the 2001 season [4], out of concern that coaches might run up huge margins, violating good sportsmanship.

There are two opposing philosophies in ranking methods, leading to *retrodictive* and *predictive* rankings. Retrodictive rankings aim to reflect most accurately the results of the current season in hindsight (minimizing *violations*, cases of a lower-ranked team defeating a higher-ranked one). Predictive rankings attempt to identify the strongest teams at the present, so as to forecast the winners of upcoming contests. Most predictive rankings use margin-of-victory and also consider home-field advantage, both in previous and future games. To achieve reasonable early-season results, predictive ranking methods typically carry over data from the previous season, perhaps with adjustments made for returning or departing players and coaching changes. Retrodictive rankings start from scratch each season, so they are not published until each team has played several games. Jay Coleman's MinV ranking [6, 7], designed to achieve optimal retrodictive results, is superior to any other ranking in that category [9]. No predictive ranking algorithm has consistently outperformed the Las Vegas oddsmakers [1], who have a strong financial interest in accurately assessing upcoming games. Many ranking systems seek a balance of predictive and retrodictive quality, attempting to give insight into future contests while remaining faithful to past results.

Among the six Bowl Championship Series computer rankings, the Colley Matrix, designed by astrophysicist Wes Colley, is unique. Colley's algorithm [8], rooted in linear algebra, is elegant in its simplicity. His rankings are relatively easy to reproduce, and Colley's method involves neither margin-of-victory nor home-field advantage. This method is not a strong predictor [2], nor is it designed to be, raising the question of whether it could be extended to create a predictive algorithm based on linear algebra. In such an attempt, we must accept losing the elegance of Colley's original method, and will need to use additional data, such as victory margins and home-field advantage.

To extend Colley's method in this way, we must first understand how it works. The method begins by rating each team using a modified winning percentage, then adjusts the ratings according to the quality of opposition a team has faced. Each team effectively starts the season with a win and a loss, to avoid having undefeated teams automatically at the top (and winless teams at the bottom) regardless of schedule strength. All teams are placed in a fixed, arbitrary order, so that each row of a square matrix system $Ax = b$ relates to a particular team. Before any games are played, $A$ is a diagonal matrix with 2's for all diagonal entries, and $b$ is a vector consisting of all ones, as shown in (1). If we solve the system, we find that $x_j = 0.5$ for each $j$, agreeing with a winning percentage of 0.5, from an implied 1-1 initial record. Thus, all teams are considered equal prior to the start of the season.

$$\begin{pmatrix} 2 & & & & \\ & 2 & & & \\ & & 2 & & \\ & & & \ddots & \\ & & & & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \quad (1)$$

To include the result of a game, we add one to the diagonal element associated with each of the teams involved, and subtract one from each of the two off-diagonal elements whose locations are coordinates are given (in either order) by the index numbers of those two teams. Finally, we add one-half to the entry of $b$ associated with the winning team, and subtract one-half from the entry of the losing team. After all games have been included, the nonzero entries in $A$ are as follows:

$a_{jj} = 2 +$ (number of games played by team #$j$)

$a_{ij} = -1 \cdot$ (number of games between team #$i$ and team #$j$), for $i \neq j$

$b_j = ½ \cdot$ [2 + (wins by team #$j$) – (losses by team #$j$)]

For example, if team #3 defeats team #1 in the season's first game, we would obtain the system in (2). The principle is that the rating of team #3 minus that of team #1 should equal one-half, assuming equal schedules, and our changes to each team's row reflect such a condition Solving this system would show that $x_1 = 0.375$ and $x_3 = 0.625$

(with all other $x_j$ still 0.5), so team #3 is currently rated as the strongest team, while team #1 is rated the weakest. This is logical, given that we have no information regarding the strength of any other teams. The strength-of-schedule inequality reduces the difference between the teams' ratings to 0.625 – 0.375 = 0.25, instead of the original 0.5 difference.

$$\begin{pmatrix} 3 & & -1 & & \\ & 2 & & & \\ -1 & & 3 & & \\ & & & \ddots & \\ & & & & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} 0.5 \\ 1 \\ 1.5 \\ \vdots \\ 1 \end{pmatrix} \qquad (2)$$

At any time, we can solve $Ax = b$ to determine the team rating vector $x$, and rank the teams according to their $x_j$ values, in descending order. For future games, we would predict a team with a higher rating (from $x$) to defeat any with a lower rating.

One of the difficulties in algorithmic college football rankings is that there are many teams (120 in the highest classification, and a total of over 700 NCAA teams), yet each team plays relatively few games in a season, typically 10-13. Of these games, 7-8 games are played within a conference, and non-conference opponents are often chosen partially by geographic proximity. Most randomly chosen pairs of teams do not play one another, but are connected only through chains of common opponents. By season's end, any two of the 120 teams can be connected by three or fewer intermediate teams. (In graph theory terms, this is equivalent to a graph diameter of four.) However, the early-season lack of connectedness increases the difficulty of ranking teams. College basketball, despite having more teams, presents a simpler ranking problem, because there are more non-conference games and more interregional contests. By their design, some methods handle such weakly-connected networks better than others, and Colley notes that the performance of his method is dependent on the degree of connectedness among the teams [8]. While this cannot be completely resolved without fundamentally changing the model's structure, we will attempt to compensate for it.

Another ranking issue is strength-of-schedule differential, which largely arises from conference affiliations, since most teams play roughly two-thirds of their games against league foes. Of the eleven conferences in major college football, three (the Sun Belt, Mid-American, and Conference USA) are relatively weak. Over the last five seasons, each of these leagues has won fewer than one in seven games against teams from the BCS-affiliated conferences [13]. It is common for a team in such a conference to go through an entire season without playing any nationally-ranked opponents, so even an undefeated record might not be meaningful. On the other hand, teams in the Big 12 and Southeastern Conference routinely play as many as six or seven regular-season games against consensus top-25 teams. The issue of schedule differences is not difficult to address within the Colley framework.

We hope to extend the Colley model in a way that improves predictive accuracy. We will leave much of the structure intact, but will alter the coefficient changes made with when including each new game result. The changes are as follows:

- Include margin-of-victory
- Weight games unequally, depending both on margin-of-victory and the expected result.
- Weight recent games more heavily.
- Start each season with teams having unequal ratings, based on results from the previous season, to improve early-season predictions, but diminish the effects of the initial inequalities as more games are played.
- Quantify home-field advantage, for use in ratings-based predictions.

While margin-of-victory is no longer used in the BCS rankings, it remains useful as a predictive tool. A team that consistently wins by substantial margins is likely to be superior to one that wins by just a few points, unless there is a substantial difference in strength-of-schedule. A similar argument can be made regarding losing teams that are soundly defeated compared with those that often lose close games. Victory margins are important, but we must choose carefully how best to use them. Scoring a late touchdown to increase a small lead is significant, but once a team is far ahead, an additional score becomes largely irrelevant. Thus, we need a diminishing returns principle applied to victory margins, obtained by using a sigmoidal (S-shaped) curve to determine the margin-based output value for a particular game. We will use a cumulative normal distribution function, translated to pass through the origin, as shown in Figure 1.

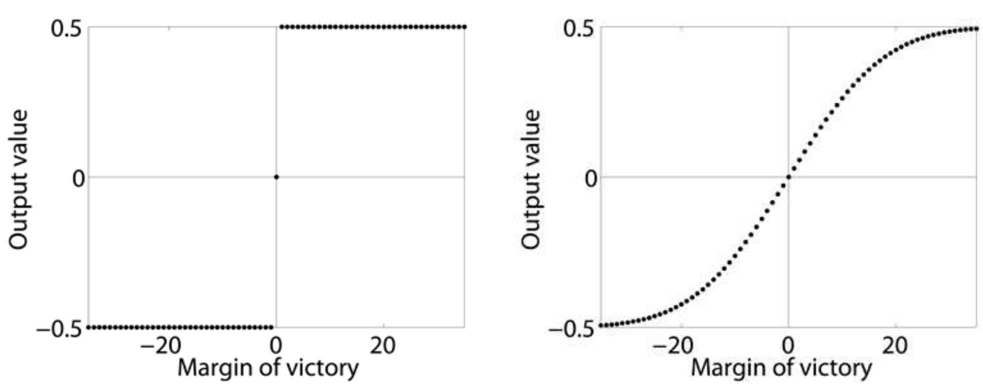


**Figure 1: Output value *m* (to be added to $b_j$), based on the outcome of a game, under (left) the Colley system, and (right) a diminishing-returns principle.**

Some ranking systems that use margin-of-victory impose a cap, often 21 or 28 points, the equivalent of 3 or 4 touchdowns, ignoring any margin in excess of that cap, and we will use a 21-point cap. This results in a function with a range smaller than the desired [-0.5, 0.5] range. One way to resolve this issue would be to multiply the capped function by a constant to stretch it vertically. We will take a different approach, inserting the missing increment between the zero point (representing a tie game) and ±1 point values, as seen in Figure 2. The translations offer a middle ground in between a margin-blind formula and a fully margin-based formula.

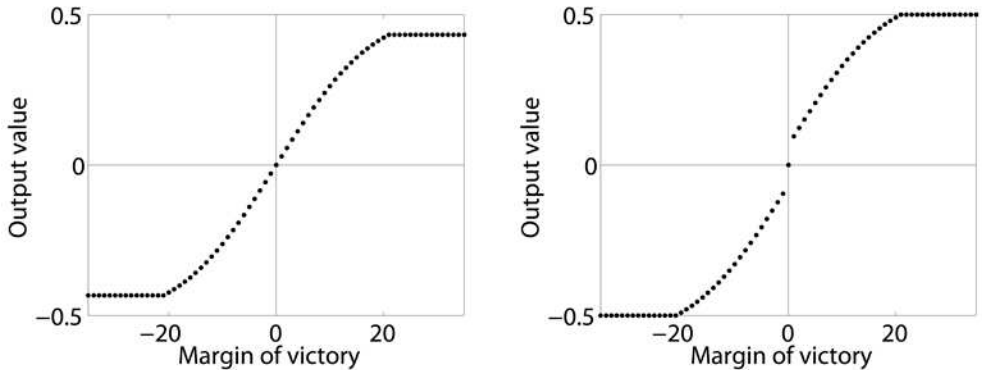


**Figure 2: Output value *m* using a 21-point cap on margin-of-victory, before (left) and after (right) making a range-adjusting translation.**

In college football, games in which one team is very heavily favored (by 35 or more points) are not unusual, due to the large disparity between the best and worst teams. Using the Las Vegas odds, there were fourteen such games in 2008 [11]. These games may produce a no-win situation for the favored team under Colley's system, leading to a rating decrease regardless of the outcome, because the strength-of-schedule decrease can outweigh any benefit gained from adding a victory. On the other side, a low-rated team may be rewarded (with a rating increase) for playing a top opponent, regardless of how badly they are beaten. While this may be reasonable within some types of rating systems, (as they may aim to show which teams have performed the best against good competition) it is a weakness for predictive ratings. Logically, if a heavily-favored strong team wins easily over a weak opponent, then neither team's rating should change significantly as a result, because no new information was gained. Thus, we use differential weighting of games, based on both expected and actual outcomes. After we compute the new ratings, the expected outcome of games may change, so we re-weight every game and re-compute the ratings. This leads to a shift from direct solution (solving a matrix system $Ax = b$ once in Colley's method) to iterative computation, repeating steps until we obtain convergence, when teams' ratings stabilize. To compute the weight given to each game, we will apply the following principles:

- Heavily favored teams that win by substantial margins should not be penalized, so such games will have very small weights.
- Any true upset (a game in which the losing team is favored in hindsight) will receive the highest possible weight.
- A heavily-favored team that wins a close game will be penalized, but not as much as if they had lost the game.
- Barring upsets, games between evenly-matched teams are the most informative, so they will be weighted more heavily.
- The function computing game weights from margins should be smooth, while allowing a discontinuity at the point representing a margin of zero.

To follow these guidelines, we use a weight function that decreases exponentially with the product of the expected and actual margins, as shown in Figure 3, whenever the favored team wins. Both the favored team and the expected margin will be determined in hindsight, because of iterative computation.

For a variety of reasons, a team's performance may improve or decline during a season, i.e. we will weight recent games more heavily. To determine the weight *w* of a game played *n* weeks ago, we use the exponential decay formula $w = \alpha^n$, where $0 < \alpha < 1$. Using $\alpha = 0.95$, we obtain time-based weights shown in Figure 4, which are multiplied by the margin-based weights to compute the weight for a particular game.

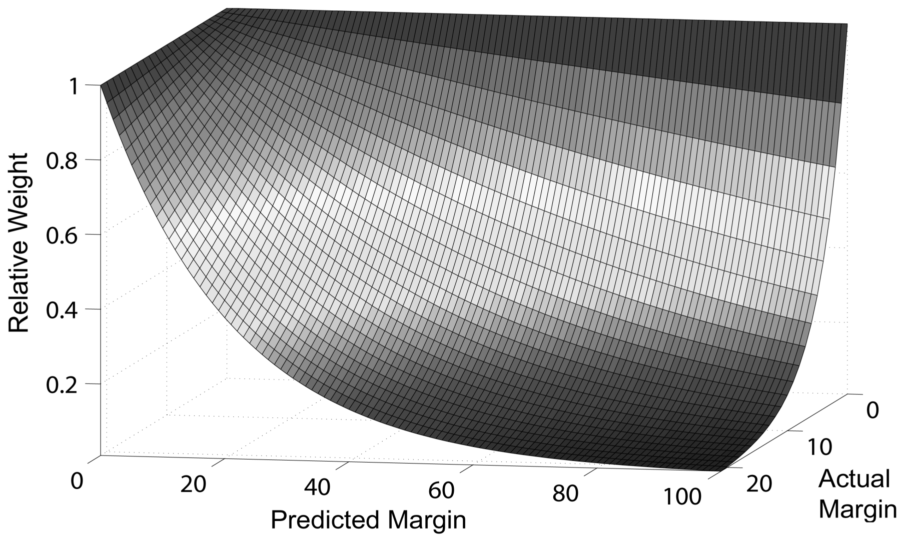


**Figure 3: Relative weight function *r* for games won by the favored team**

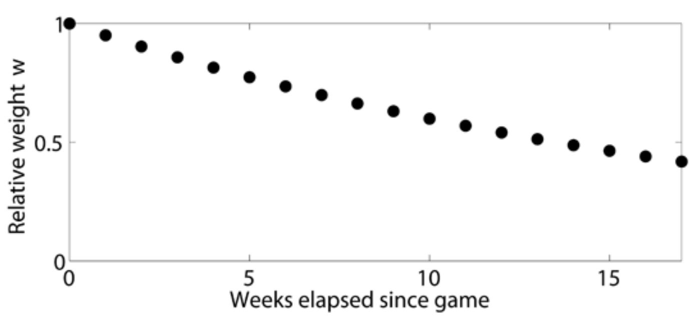


**Figure 4: Time-based weight function *w*, using a 5% weekly decrease.**

Just as every team starts a new season with a record of 0-0, regardless of the previous season's successes or failures, rankings should start fresh each year. In small, well-connected groups such as the National Football League, this is feasible, but in large, poorly-connected pools of teams, it is unrealistic. There may be few interregional games, and those may occur only in the playoffs, too late to be of predictive value. For this reason, many predictive systems use the results of one or more previous years as a starting point for a new season's rankings. As more data become available, the old results are weighted less, and perhaps dropped altogether once all teams are minimally connected. Using previous rankings as a starting point can improve predictive accuracy early in the season. If the starting values are not phased out entirely, they may help in predicting interregional games, particularly when certain regions have been historically strong or weak but scheduling is mostly localized. We will use the final ratings of the previous season as a starting point for a new year, and will consider them equivalent to two fully-weighted games, just as Colley in essence starts each team with a 1-1 record. Just as current-season games are weighted 5% less each week, we apply the same damping to the initial rating weights.

Computation of the value of home-field advantage could easily be the subject of a paper by itself. Some rating systems use a fixed value for the number of points by which a home team is improved, compared to playing on a neutral field. Others (such as Sagarin's ratings [12]) re-compute home-field advantage weekly during the season, either as a universal constant or with a value for each team. We will compute a single value based on previous seasons and use it throughout a given year. While college football teams rarely face an opponent twice during the same season, they routinely play the same team in consecutive years, with each team hosting one game. We will consider the margins of games in such home-and-home series to estimate a home-field advantage constant. We sum the margins, from the home team's perspective, in all such pairs of games over two seasons, then divide by the total number of games considered. By considering only home-and-home series, we eliminate home and away scheduling inequalities, such as the practice of top college teams paying weaker opponents to play at the stronger team's stadium with no return game. The home-field advantage value $h$ was approximately 3.70 points in major-college football over two recent seasons. Since $h$ should not change substantially in a single year, we will consider it constant through a season.

Putting together these ideas, we must compute two values used to change the matrix system when a game is added. First, the final weight $y$ given to the game is a product of the relative weight $r$ (based on expected and actual outcome) and time-based weight $w$. The coefficient adjustment $z$ (applied to $b_j$) is a product of $y$ and the margin-based game output value $m$.

$$y = rw, \; z = ym \qquad (3)$$

We can then make the changes shown in (4) to account for a game in which team #3 defeats team #1. In (4), entries denoted by * are unchanged.

$$\begin{pmatrix} +y & * & -y & & * \\ * & * & * & \cdots & * \\ -y & * & +y & & * \\ & \vdots & & \ddots & \\ * & * & * & & * \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} -y \cdot z \\ * \\ +y \cdot z \\ \vdots \\ * \end{pmatrix} \quad (4)$$

After solving the matrix system to obtain the rating vector **x**, we can determine hypothetical margins of future (and past) games. To do this, we convert the values from $\boldsymbol{x}_j$ into ratings easily used for predictions. The conversion formula is given in (5), where $c$ is a scaling factor (applied so that a one-point difference in ratings will be equivalent to a one-point predicted margin) and $a$ is the desired rating for an average team. Based on data fitting experiments, we will use $c = 60$. While the value of $a$ does not affect the predictive outcomes, we will use $a = 100$, keeping with common practice. Note that in (5), we subtract 0.5, the rating of an average team in the Colley system.

$$R_j = a + c \cdot (x_j - 0.5) \quad (5)$$

To predict the outcome of a game, we consider the teams' ratings, $R_h$ and $R_v$ (for the home and visiting teams) given by (5). Unless the game is contested at a neutral site, we add the home-field constant $h = 3.70$ points to the home team's rating. After this, the team with the higher rating will be predicted to win, and the difference between the modified ratings is the predicted margin-of-victory $M$, as seen in (6).

$$M = |\, R_h + h - R_v \,| \quad (6)$$

For highly mismatched teams, the formula will typically overestimate the margin. A team that is 70 points better than its opponent is unlikely to win by such a huge margin, instead inserting reserve players and playing conservatively with a large lead. If the predicted margin $M$ is greater than 21 points, we reduce it using the fractional-exponent formula given by (7). Based on data fitting, we take $\alpha = 0.8$.

$$M^* = 21 + 1/\alpha \cdot [(\mathrm{M} - 20)^{\alpha} - 1], \text{ if } M > 21 \quad (7)$$

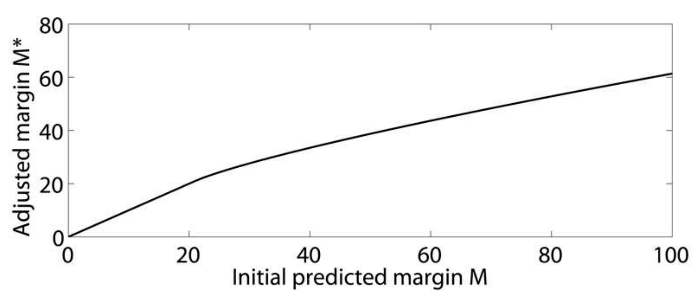


**Figure 5: Adjustment of margin predictions which initially exceed 21 points.**

The margin, possibly adjusted, is then rounded to the nearest integer. Because tie games are not possible in college football, a margin of less than one-half point will be rounded up to one point instead of down to zero. In contrast, some rating systems (such as those of the Las Vegas oddsmakers) allow a predicted margin of zero.

The weight assigned to a game depends partly on its predicted outcome, so we need an initial set of ratings as a starting point. While we could initially rate all teams equally, we will instead begin with ratings carried over from the previous season. As we obtain new ratings, the predicted outcomes of some games will change (and for others the margin is altered without changing the predicted winner). These changes affect the weights assigned to the games, changing the ratings again. So, rating computation is an iterative process, in which ratings from one iteration are used to compute the game weights for the next. We continue until we obtain convergence, when ratings remain unchanged with additional iterations. Convergence usually takes only a few iterations.

The primary measure of a predictive rating system's quality is the percentage of games it correctly determines in advance; accuracy in determining margins is a secondary criterion. The Prediction Tracker [2] compares the results of dozens of rating systems. For the 2008 season, the system we describe correctly predicted the winner of 74.3% of games between two major-college (NCAA Division I FBS) teams. This compares favorably with other systems, placing it in the top 20% of all models. Prediction percentages for the second half of the season are also a good measure, because many ranking systems start from scratch each year, requiring more data to make accurate assessments. While more information is available later in the season, predictions become more difficult because conference games and postseason bowls are more often evenly-matched contests. Not surprisingly, our prediction percentage was lower during the latter half of the season, at 72.4%, but this still placed in the top quarter of all ranking systems. The final top 25 teams for the 2008 season, using our method, are listed as the Appendix.

There are many algorithmic rankings in major college football, probably because it has neither a playoff system nor an official national champion. However, I also have a strong interest in ranking high school football teams, dating back to several years spent teaching mathematics in a public high school. Power ratings and predictions for high school football in North Carolina and Ohio, computed by the same method discussed here, are published weekly during the season at www.fantastic50.net.

Some of the difficulties encountered in ranking college football teams are more apparent in those leagues. Ohio's 700+ teams each play only ten regular-season games, and North Carolina's geography leads to heavily-localized scheduling. These factors substantially reduce connectedness among the teams, so that at season's end, some pairs of teams are connected only through lengthy chains of intermediate teams. Strength-of-schedule differences can be enormous, as some large urban schools fielding three large teams play only against local peers, while small-school conferences may consist entirely of teams that have barely enough players to field one team. Obtaining scores of games involving rural schools with no nearby daily newspapers can be a challenge. However, high school football offers another interesting proving ground for mathematical ranking systems. Many states use a computational system of some kind to select and/or seed playoff teams, but there are few predictive rankings published in a given state. Such rankings offer a measuring stick for championship teams in lower classifications if they are unable to schedule games against the large-school powers.

## Acknowledgements

Historical game scores and Las Vegas odds were obtained from Warren Repole; scores were also provided by Jay Coleman. Thanks to Arnold Solomon of NCpreps.com, Eric Frantz of JJHuddle.com, and Tim Stevens at The News and Observer, for kindly providing venues to publish the author's high school football rankings.


## Sources Cited


1. T. Beck, College Football Ratings PT Awards, *The Prediction Tracker*, accessed on 16 July 2009. <http://tbeck.freeshell.org/fb/awards2008.html>

2. T. Beck, Computer Rating System Prediction Results for College Football, *The Prediction Tracker*, accessed on 16 July 2009. <http://www.thepredictiontracker.com/ncaaresults.php?type=1&year=08>

3. R. Billingsley, Dynamics of the system, Oct. 2008, *College Football Research Center*, accessed on 28 July 2009. <http://www.cfrc.com/Archives/Dynamics_08.htm>

4. ___________, BCS Chronology, 25 June 2009, *Bowl Championship Series*, accessed on 11 August 2009. <http://www.bcsfootball.org/bcsfb/history>

5. ___________, BCS Computer Rankings, 11 Nov. 2008, *Bowl Championship Series*, accessed on 21 July 2009. <http://www.bcsfootball.org/bcsfb/rankings>

6. J. Coleman, Minimum Violations (MinV) College Football Ranking, 10 Jan. 2009, *University of North Florida*, accessed on July 8, 2009. <http://www.unf.edu/~jcoleman/minv.htm>

7. J. Coleman, Minimizing Game Score Violations in College Football Rankings, *Interfaces* **35** (2005) 483-496.

8. W. N. Colley, Colley's Bias Free College Football Ranking Method, 30 May 2002, *Colley Matrix*, accessed on 25 Aug. 2009. <http://www.colleyrankings.com/method.html>

9. K. Massey, College Football Ranking Comparison, 23 Mar. 2009, *Massey Ratings*, accessed on 22 July 2009. <http://www.mratings.com/cf/compare.htm>

10. K. Massey, Massey Ratings Description, 15 Aug. 2000, *Massey Ratings*, accessed on 28 July 2009. <http://www.mratings.com/theory/massey.htm

11. W. Repole, Historical Scores, 31 May 2009, *The Sunshine Forecast*, accessed on 23 June 2009. <http://www.repole.com/sun4cast/scores2009.html>

12. J. Sagarin, Jeff Sagarin NCAA Football Ratings, 9 Jan. 2009, *USA Today*, accessed on 22 July 2009. <http://www.usatoday.com/sports/sagarin/fbt08.htm>

13. C. Stassen, Aggregate Conference Record, *Stassen.com College Football Information*, accessed on 20 Aug. 2009. <http://football.stassen.com/records/multi-conference.html>

14. D. L. Wilson, American College Football – Rankings, *Nutshell Sports Ratings*, accessed on 22 July 2009. <http://www.nutshellsports.com/wilson/popup.html>

**Appendix – Top 25 ranking, at the end of the 2008 season, by this method**
Records shown include only games against FBS (formerly Division I-A) opponents.

1) Florida (12-1)
2) Southern Cal (12-1)
3) Oklahoma (11-2)
4) Texas (12-1)
5) Penn State (10-2)
6) Ohio State (9-3)
7) Utah (12-0)
8) Oregon (10-3)
9) Alabama (12-2)
10) Georgia (9-3)
11) Boise State (11-1)
12) TCU (10-2)
13) Texas Tech (9-2)
14) Virginia Tech (9-4)
15) Mississippi (8-4)
16) Oregon State (9-4)
17) Florida State (7-4)
18) Missouri (9-4)
19) California (9-4)
20) Arizona (8-5)
21) West Virginia (8-4)
22) Iowa (8-4)
23) Oklahoma State (8-4)
24) Boston College (8-5)
25) LSU (7-5)

R. DREW PASTEUR is an Assistant Professor of Mathematics at The College of Wooster. Before pursuing graduate study at North Carolina State University, he was a math teacher and head athletic trainer at Fuquay-Varina High School, in Fuquay-Varina, NC. It was during this time that Drew became interested in mathematical ranking algorithms and developed the first version of his Fantastic 50 high school football rankings. His other research area is in mathematical biology, applying dynamical systems to human physiology. Outside of mathematics, Drew enjoys singing, running, and spending time with his wife and son.

311 Taylor Hall, 1189 Beall Ave., Wooster, OH, 44691. rpasteur@wooster.edu