

Introduction to RUSH

- **What is RUSH?** RUSH (Ratings Using Score Histories) is a technique for rating college football teams based not on just the final score margin of the games, but on the games' entire history of the score margin.

Fans interested in so-called "computer rankings" are familiar with the distinction between ratings like those used in the later years of the BCS and those that are based on the margin of victory (MOV). BCS-type ratings use no information about a game's results other than which team won. Ratings based on MOV use more information and hence provide more accurate estimates of teams' true abilities. However, using MOV bothers some because it misses important information about how the teams arrived at the final score: in particular, meaningless late scoring plays can be as influential as scores that played a key role in determining the winner.

RUSH goes beyond using MOV to utilize the entire score history, and by doing so it is able to deemphasize scoring plays that cannot realistically affect which team wins the game. For example, winning teams are not rewarded for "running up the score" by scoring late touchdowns after the game has long been decided, and they are also not penalized if their opponents score late to make the final score deceptively close. RUSH also treats games that are quickly put out of reach differently than games that are close for a long time before one team pulls away. For example, two evenly matched teams can realistically play a 28-0 game where the winners score one touchdown in each quarter, but a game with all four touchdowns in the first quarter is stronger evidence that the winners are substantially better.

RUSH's approach for deemphasizing unimportant scoring plays is closely related to the concept of Leverage Index, which looks at changes in the probability of a particular team winning a game as a function of specific plays. Leverage Index is a creation of Tom Tango, and he and others have extensively applied the idea in baseball and other sports. In football, the concept is an ingredient in Total Quarterback Rating (QBR), for example.

RUSH is a Bayesian statistical model. As such, it can provide answers to all sorts of probability questions, such as "What is the probability that Team A is better than Team B?", "What is the probability that Team A is the best team in the country?", "What is the probability that Team A will beat Team B on a neutral field?" These answers depend on the RUSH assumptions and are only based on the information used to build the RUSH model. (The word "Bayesian" has been used in college football ratings discussion to mean that ratings are based on prior information about how strong teams have been in past seasons, but we are using it in the more general probability theory sense, in which all uncertainties are modeled using probability distributions.)

- **What information is used by RUSH?** RUSH uses the entire history of each game's score as it evolves through time. It uses all games from a single season. It does not use information on team abilities from past seasons: in particular, team abilities will be estimated rather poorly early in the season. (It is possible to add information from past seasons into the RUSH calculation, and occasionally we will do this in order to provide early-season ratings, but if a given set of RUSH ratings includes past seasons' results, it will be clearly stated.) RUSH assumes a home field advantage common to all teams. Each team has a parameter called tau which governs how many scoring plays are likely to take place in a game involving that team. We rate all Division I teams, and exclude games involving teams not in Division I.

Since a team's performance in a given game is evaluated based on how strong their opponent was, schedule strength is intimately involved in RUSH ratings, although we do not compute a simple strength-of-schedule measure. If two teams have the same record, the team who has

played the stronger schedule will not necessarily be rated more highly, if the team with the weaker schedule was more impressive relative to the strength of their competition.

If two teams have played each other, the head-to-head winner will not necessarily be rated more highly, since RUSH considers all games when comparing teams. (The more decisive the head-to-head result, the more likely the relative rankings will be consistent with the result). A team can certainly drop in the RUSH ratings from one week to the next even in a week where they win, if their win was unimpressive and/or their previous opponents have bad weeks themselves. Games played recently are not treated as being more influential than early games, and in particular conference championship games are treated like any other game.

Some other advanced computer ratings use more detail about what happens in a game than RUSH does. For example, Brian Fremeau's Fremeau Efficiency Index looks at all *drives* in a game, whether or not they result in points, and Bill Connelly's S&P+ ratings look at all *plays* and whether or not they are successful. These are both available at footballoutsiders.com and are both enormously promising approaches. RUSH will miss some information that these ratings will benefit from, but I think it's of interest to see what happens when one ignores drives and plays that do not actually lead to important changes on the scoreboard.

- **What are the goals of RUSH?** The ideal use for RUSH is as a tool for a selection committee tasked with choosing teams for postseason playoffs. For this purpose, RUSH has the important property that all teams are on equal footing at the start of a year, but it tries to use as much of the information available in the game results of a given year without benefiting teams that score meaningless points toward the ends of games.

However, it is not appropriate to use RUSH results automatically in a formula as the BCS computer ratings have been. (In fact, I believe that "computer ratings" should never be used automatically in this way, but should be taken seriously by human decisionmakers who understand their strengths and weaknesses.) RUSH is complex, slow to evaluate, and based on data not necessarily available in standardized form, so practical issues could prevent it from being relied upon in official use.

- **Does RUSH rank the teams from strongest to weakest, or from best resumé to worst resumé?**

This is a tricky but important question. These two philosophies of rating teams can lead to quite different rankings, and people may disagree strongly with rankings because they are expecting a different philosophy.

RUSH is based on a statistical model, so it tries to estimate the true abilities of the teams. However, it only rates a team based on their games in one season, so in that sense it is resumé-based. The term "resumé" in college football has come to be associated with the BCS-compliant ratings, and most often refers to who a team has beaten and lost to without regard to how convincingly they did so.

I think it's fair to say that compared to other ratings based only on a given season's results, RUSH ratings are closer to being strongest-to-weakest than they are to best-to-worst resumé. This may not be what you want in a particular situation. To use a recent example, at the time of BCS selection in 2012, Notre Dame was rated #7 in RUSH, indicating that RUSH believed Notre Dame not to be one of the 2 or even 4 strongest teams in the country (RUSH was far from alone in having this opinion). However, it was completely reasonable to believe that Notre Dame should be on top of the list of teams most deserving of a spot in a playoff, since their resumé of being undefeated against a respectable schedule was arguably superior to anyone else's. (Indeed, they were rated #1 in many ratings systems based only on who won the games).

- **Should I use RUSH to gamble?** No. We don't recommend gambling on sports in the first place, but even if you insist on doing so, there are a couple of reasons why RUSH is not an ideal tool to use. Most importantly, it uses only the current season's results. RUSH predictions early in the season could certainly be improved by using information from the previous season(s), and this information is likely to be valuable even late in the year. Also, RUSH ignores scoring plays that occur after the result is decided, while there is often useful information in these scoring plays for prediction purposes. In particular, RUSH should be very bad at guessing whether a team will cover a very large spread such as 35 points, since in general RUSH does not care whether a team wins by 28 or 42, for example.
- **What are all these things in the RUSH report?**

The first element in the report is a list of teams in the order in which RUSH ranks them. The columns in this table are as follows. First there is a column of ranks ranging from 1 to the number of teams. Next is the "expected rank" from the RUSH posterior distribution; RUSH ranks teams in order of this column. The expected ranks can range from 1 to the number of teams, and an expected rank of 1.00 means that we are certain that that is the best team. The expected ranks become more spread out as the season progresses. After the team's name, the next column, "rating", is the "triple-goal" estimate of a team's parameter, followed by its posterior mean, followed by its posterior standard deviation (sd). (For more on triple-goal estimates, see papers by W. Shen and T.A. Louis, including in the Journal of the Royal Statistical Society Series B, 2002.) A low value of sd means that RUSH is relatively confident of its estimate of the team's ability: in general, playing more games gives you a lower sd, and playing close games against both strong and weak teams will also reduce a team's sd. Finally we give each team's estimated (posterior mean) pace (τ) parameter: a high value of τ means that there tend to be many scoring plays in that team's games.

After the table of teams, we review each team's season from the RUSH point of view. We list each team's games in order of how well that game went for them compared to what the RUSH model would expect. The columns are a game identifier, visiting team, visitor score, home team, home score, an estimate of the team's rating for that game alone, a standard deviation associated with that rating, a ranking that indicates where the team would rank if they had played as well in all their games as they did in that game, and a sort of two-standard-deviation confidence interval for the team's ranking based on that game alone. These game-specific ratings are a work in progress, but we hope this is a useful tool in helping people understand how RUSH works by giving a rough idea of which games benefit a team's rating and which games hurt it.

By definition, the best team in the nation has a rating of 10. (The #1-rated team in RUSH will have a rating lower than 10, because RUSH will not be 100% certain who the best team is.) Another surprising feature is that one team may have a higher posterior mean than another but be ranked lower according to expected rank. Toward the top of the rankings, this happens because there is more uncertainty about the strength of the higher-rated, lower-ranked team.