Predicting College Basketball Results Using Only Statistics

Brian Vernarsky*

(Dated: January 3, 2016)

I. INTRODUCTION

For a few weeks every year college basketball completely captures the interest of the sports fan's world. The NCAA basketball tournament starts with 68 teams (as of 2016), and after six rounds of competition ends with one champion, in a result that is decided entirely on the court. On the Sunday before the tournament begins, the tournament committee releases the complete bracket, outlining the path that each team will have to take to get to the championship; there is no re-seeding after the rounds, so the path is set in stone: each team knows who they could possibly play in each of the rounds. Because of this setup, it has become very popular and fun to make predictions about the tournament by filling in a bracket, starting from the one the committee has set, and then progressing through the rounds. The bracket challenge is one of the most prevalent and institutionalized forms of gambling and sports prognostication in the United States, with office pools, ESPN, CBS, and other sports websites setting up competitions, and one year H&R Block even offered a billion dollars to anyone who could pick a perfect bracket (a near impossibility).

Based off of this competition, I decided to come up with any way that I could to improve my picks, and so I decided to write a computer program to make my picks for me. What started off as a very basic program based off of probabilities, random-number generation, and self-defined categories such as heart and coaching strength, eventually developed into a more sophisticated set of programs, which could not only make predictions for the bracket in the NCAA tournament, but also nearly every single college basketball game over the course of the season, the crux of which is a single equation giving a single number which predicts the likelihood of each team winning the game.

A. The Function

What I want is a fairly simple equation to quickly tell me, for a game involving Team A and Team B, at a set location (home for A, home for B, or at a netural location), on a given date (at least a few weeks into the season), how likely Team A is to win, and how likely Team B is to win.

I want this function to depend solely on the stats to that point in the season, no other information can be input. So there is no information about individual players, injuries or suspensions to players or coaches, the fact that it is homecoming, or a huge rivalry, or anything else: just the stats. The function can use any of the basic or advanced statistics, but I will start by considering only the so-called *four factors*, defined by Dean Oliver, of offensive rebounding percentage, effective field-goal percentage, free-throw-made rate, and turnover percentage. Some variant of points scored and allowed or winning percentage can also be involved: the SRS (or simple rating system) is a likely candidate for points, and RPI could be used for winning percentage.

The function would weight each of the factors individually so that the more important statistics are given more weight. Since we must consider the same stats from Team A and Team B, and Team B's defense will affect Team A's offense, and vice versa, it makes sense to have each term be of the form

$$w_s(P(s_{O,A}|\bar{s}_{O,A,date}, \bar{s}_{D,B,date}, loc_A) - P(s_{O,B}|\bar{s}_{O,B,date}, \bar{s}_{D,A,date}, loc_B)), \tag{1}$$

where w_s is the weight for stat s, $s_{O,X}$ is the predicted offensive stat s for Team X, $\bar{s}_{O/D,X,date}$ is the (weighted) average offensive/defensive stat s for Team X on the date of the game, loc_X is the location of the game for Team X (if Team A is at home, Team B will be away, and vice versa; Team A and B can also be at a neutral location), and the P function is a function that makes a prediction for $s_{O,X}$ based on the average offensive statistic s for Team X and the defensive statistic s for its opponent on the date in question.

* brian.vernarsky@gmail.com (919) 260-6957

B. Weighted Averages and Their Functions

A and B's offensive and defensive averages in the *four factors* seem like relatively simple values to computer, but taking location into account is a bit harder. Certainly, teams win more at home than on the road (66% to 33% in fact), and probably score more points, but does their effective-field- goal percentage actually change? Or any of the other statistics? What I would like is to be able to calculate the averages as if each game were played on a neutral court, and then "deneutralize" the prediction based on location. In a league like the NBA, where each team plays an equal amount of home and away games (and no games on neutral floors), this might not be necessary, but the bigger college basketball teams tend to play more home games, especially early in the season, than road games, while smaller schools have to go on the road a bit more often.

And, as long as we are neutralizing the stats, we should also consider the fact that the level of competetion varies considerably among the 350+ Division-I college basketball teams. Therefore, in order to make our predictions as reliable as possible, we should try to make the averages take the level of competition into account. This means that we should weight the performance in a game by how good the opponent was; and we can do this not just on a team-wide level, but on a stat-by-stat level: if a team is great at rebounding, but not that great at shooting, they will have significantly different weights for those stats. Grabbing 30% of the available offensive rebounds in a game is very impressive against a team that generally only allows their opponents to grab 20%, but would considered underwhelming against a team that gives up 50% on average. However, we also need to consider the level of competition that the opponent has faced, so we want to take into account the opponents' opponents. Ideally, we'd go even further, but it is a time-consuming process and adding another layer multiples the run-time by the approximate number of opponents for a team in a season, which would be a factor of roughly 30.

So, I will develop a method to produce neutralized and opponent-adjusted stats, which we can then input into the prediction function. Then I will need to determine the P functions from Equation 1 to predict Team X's offensive stats based on its offensive weighted averages, and its opponent's defensive weighted averages. Then I will need to determine the weights for each of the four statistics and SRS. Since each of the four statistics and the SRS each have their own characteristic range of values, I will need to calculate standard deviations for each of them so that they contribute roughly equally to the function before attempting to determine their relative weights.

C. Summary

Once the final function is determined, and the weights have been set, it is a relatively simple matter to develop a mapping of the value of the function onto a predicted likelihood of winning for each team using historical data. The method is relatively simple, containing only four or five terms, while taking into account A's effect on B, and B's effect on A, the location of the game, and the strength of each team's schedule. It is possible to determine a ranking for the teams based on their performance against a completely average opponent.

One way to determine how well the function is doing is to make predictions about games in the past (using only statistics from before the game was played), and comparing the predictions to the actual results. Another way is look at the betting point-spreads that were available for the game at game-time. The spreads can be converted into likelihood of winning and vice versa, allowing for direct comparisons.

In this paper, I will first discuss the data that I will be using in the predictions in Section II, as well as explaining each of the statistics and their formulas in Section III. Then I will go through the process determining the factors necessary to "neutralize" the statistics in Section IV. Then I will explain how to develop the weighted averages in Section V. The method used for determining SRS will be detailed in Section IX. In Section VI, I will outline how to determine the P functions to predict a team's offensive output based on its and its opponents weighted averages, as well as how to calculate the necessary standard deviations. Then in Section VII, I will show how I determined the weights necessary for the main function. Finally, in Section VIII, I will show some of the results of the function, including predictions for the NCAA tournament in years past, before concluding in Section X.

II. DATA

The data that I will be using in this analysis is the box-score results from every game played Division-I opponents in the years from 2001-2015 (note: I refer to seasons by the year in which the season ends, thus the 2014-2015 season is referred to as the 2015 season), as well as the on-going 2016 season. I will not be using games by Division-I teams against non-Division-I opponents because I do not have a means of getting the box scores from all of the Division-II and III schools. Games played against non-Division-I opponents will simply be treated as if they did not occur. Since I don't consider win-loss records, I won't have to determine whether or not to include them in those results.

The box scores for the games include the number of points, field-goals made and attempted, two- and three-pointers made and attempted, free throws made and attempted, offensive and defensive rebounds, assists, steals, blocks, turnovers, and personal fouls. Each of these stats are recorded both for the team itself and its opponent.

There are a total of 76,306 games between Division-I schools in the 2001-2015 seasons. I will consider a subset of those games at each step in the process, and will be sure to make note of what the subset is.

III. THE STATISTICS

As mentioned in the Introduction, the *four factors* provide the basis of my analysis, along with the simple rating system or SRS, which accounts for the points-scored for and against. In this section, I will detail each of the four factors and give the definition of each. Note that each stat is determine for both the team in question and for their opponents. I will refer to the stats as offensive and defensive therefore, even though this leads to some stupidly named statistics such as "offensive offensive-rebounding percentage" and "defensive offensive- rebounding percentage". Since the stats are always defined in terms of the team in question, those stats mean the percent of offensive rebounds the team in question got, and the percent of offensive rebounds they gave up on defense (i.e. that their opponents got).

The offensive offensive-rebounding percentage is determined by the forumla

$$oorp = \frac{orb}{orb + drb_{opp}}, \tag{2}$$

where orb is the number of offensive rebounds, and drb_{opp} is the number of defensive rebounds that their opponents grabbed. Since for each available rebound off a shot taken by a team, either they gather the offensive rebound, or their opponent grabs the defensive rebound, this stat basically determines what percentage of available offensive rebounds a team was able to get. The defensive offensive-rebounding percentage is given by

$$dorp = \frac{orb_{opp}}{orb_{opp} + drb}. \tag{3}$$

The offensie effective field-goal percentage is determined by

$$oefgp = \frac{fgm + 0.5 \cdot threem}{fga} \tag{4}$$

where fgm is the number of field-goals made by the team in question, $threem$ is the number of three-point field-goals made, and fga is the number of field-goals attempted. This stat is used in place of the straight field-goal percentage because it takes into account the fact that three-pointers are worth more than two-pointers, and therefore they should be given more credit for making them. The defensive stat is determined with the same formula, but for the opponents.

The offensive turnover percentage is given by

$$otop = \frac{tov}{poss} \tag{5}$$

where tov is the number of turnovers, and $poss$ is the number of possessions, where

$$poss = fga + tov + 0.44 \cdot fta - orb\ , \tag{6}$$

and fta is the number of free-throws attempted. This stat determines what percent of the time the offense turns the ball over. The defensive stat is determined with the same formula, but for the opponents, and determines how often the team in question was able to force a turnover. Since turnovers are a bad thing for an offense, this stat is generally treated as $(1 - otop)$ and $(1 - dtop)$, which is the percentage of time the offense does not turn the ball over, and therefore lower numbers are bad and higher numbers are good as with the other stats.

The free-throw-made rate is defined as the number of free-throws made per the number of field goals attempted,

$$ftmr = \frac{ftm}{fga}, \tag{7}$$

where ftm is the number of free-throws made. This basically determines how likely a team is to get to the line, as well as how good they are at making their free throws. The defensive version of the stat is the same, just using the opponents' stats.

A. The Simple Rating System

The final stat I will use is a derived stat known as the simple rating system stat, or SRS for short. Its name suggests that it is a simple stat, but it does have a slight complication. The stat is defined as

$$SRS = \frac{1}{N_{games}} \sum_{AllGames} (pts - pts_{opp}) + \frac{1}{N_{games}} \sum_{o=Allopponents} (SRS_o), \tag{8}$$

where the first term is the team's average point differential in all games played (only against Division-I opponents) so far this season, and the second term is the average of its opponents' SRS, where the opponent's SRS does get included multiple times if they have played multiple games. The opponents' SRS values are their values on the day the SRS is being calculated for, not the day that the game was being played. Thus, if Team A plays Team B on January 1^{st} and then again on March 1^{st}, and we're calculating their SRS on March 15^{th}, Team B's SRS as of March 15^{th} would be included in the sum twice.

The complication in the calculation is that we don't initially know what everybody's SRS is, and therefore can't properly calculate the second term in each equation. The way to get around this is to solve it iteratively. Thus we start with an initial guess for each team's SRS, which will just be their average point differential, and then iterate, meaning recalculate that second term, which is also known as the *strength of schedule* or SOS, using the value of SRS for each team that was calculated in the initial round. Then do it again, using the value of SRS for each team as calculated in the previous round. After a number of iterations the SRS values should start to settle down. That number is generally between 50-100, so I usually iterate 100 times. Early in the season, the numbers sometimes never settle due to the small number of games that have been played, but later in the season they become quite stable.

This stat basically says in a game against an average opponent for the season in question, on a neutral floor, Team A would likely win by whatever value the SRS is; if the value is negative, then it means they would likely lose to an average opponent. In a game between Team A, with SRS_A, and Team B, with SRS_B, Team A would win by $SRS_A - SRS_B$ points (or lose if that value is negative). If the game is not played at a neutral site, then the SRS will need to be tweaked a bit, and we will get into that in Section IX.

There are many modifications that can be made to the SRS, from placing a cut-off on blowout wins (e.g. any win by more than 20 points is simply treated as a win of 20 points), or giving a bonus for winning the game, but they will not be used in these calculations.

IV. AVERAGES AND NEUTRAL RATIOS

The idea of "neutralizing" statistics is a fairly basic one: see how well home teams do in the stat and compare that to how teams do in games at a neutral site, and then do the same for away teams, those two ratios can then be used to neutralize a stats. In practice, I actually sum up all of the games, regardless of location, to get my "neutral" values, instead of only using games played at neutral locations. The reason that I have chosen to do this is that true neutral-site games tend to come at the very beginning of the season, and the very end of the season, while there are almost none in the middle of the season. This has the possibility of skewing the averages and ratios slightly, and so I have elected to simply sum up all of the values. Since there are an equal number of home and away games total, these sums should still be neutral, and since there are so many games, they will be quite stable as well. However, this also means that my sums over all games will be different from those at just neutral sites, and so there will also be a ratio for neutral-site games.

Ideally, when considering these averages and ratios for a given season, it would be nice to know what the average will be for that season ahead of time, but that is not possible. The average of all the games played by December 1^{st} will likely be significantly different from the average of all games played by March 1^{st}, four months later. It does not make sense to keep recalculating the averages and ratios and then recalculating the weighted averages and so on. So, it is necessary to guess what the averages and ratios will be. Since the league is fairly stable, it is a safe bet to assume that the average offensive rebounding percentage this year will be close to the league-wide average last year. We can assume that we have the statistics for the previous season and thus are able to calculate those averages. However, consider the data in Table I, which shows the average number of points scored in each year from 2010 to 2015, as well as the predictions that would have been made for that year using the previous season's result as a prediction, as well as a prediction using the previous 5 seasons' results. While the 1-year data is faster to respond to changes, like the general downturn in scoring over these six seasons, it allows anamalous seasons, like 2014's sudden jump in scoring, or 2013's dip, to lead to poorer predictions on average. The predictions made using the previous 5 seasons' data on the other hand may be a bit slower to react to certain trends, but is more stable on the whole, and does not over-react to anamalous results.

Year	Average Points	Prediction Using Previous 1 Year	Prediction Using Previous 5 Years
2010	68.6	68.0	68.6
2011	68.4	68.6	68.6
2012	67.4	68.4	68.6
2013	66.9	67.4	68.3
2014	70.2	66.9	67.9
2015	66.9	70.2	68.3

TABLE I. The average number of points scored in a year compared to predictions made using the previous 1 seasons' games, as well as just the previous 5 seasons. The advantage of using a single season is that it can react faster to trends, like the obvious downward trend for scoring over the past six seasons, as in 2013, where the 5-year prediction is not coming down as fast as the 1-year prediction. However, the disadvantage is that single-season anomalies, like 2014's sudden scoring burst, can lead to significant deviations from the best prediction. I have elected to use the 5-year averages to predict each season because of the accuracy and stability it provides.

When generating the averages and ratios, I consider every game between Division-I schools in the 5 years prior to a season. Thus, for the 2016 season, I will use the years 2011, 2012, 2013, 2014, and 2015.

The averages for the simple, counting statistics, like points scored, field goals made and attempted, and the like, are the simple averages, the sum divided by the number of games

$$\bar{s}_{m/a} = \frac{\sum s_{m/a}}{n}, \tag{9}$$

where n is the number of games in the sum. For percentages, the average is

$$\bar{s}_p = \frac{\sum s_m}{\sum s_a}, \tag{10}$$

where $\bar{s}_p$ is the the average percentage of stat s, and $s_{m/a}$ is the number of makes/attempts for stat s, and the sums are over all games.

For the ratios, I have four different sums, one over all home games, one over all road games, one over all neutral-site games, and one over all games regardless of location. The ratios are then defined to be

$$r^s_{m/a/p,loc} = \frac{\bar{s}_{m/a/p,total}}{\bar{s}_{m/a/p,loc}}, \tag{11}$$

where $r^s_{m/a/p,location}$ is the average makes/attempts/percentage for stat s in location loc. To neutralize the stats, it is thus only necessary to muliply a stat by the proper ratio.

Table II shows the averages and neutral ratios for each year from 2003-2016 calculated using the 5 seasons previous to the season in question, with the exception of the 2003, 2004, and 2005 seasons, which were calculated using only 2, 3, and 4 seasons respectively since I only have the box scores for all games going back to the 2001 season.

V. WEIGHTED AVERAGES

The weighted averages that I use in my calculations are supposed to represent the averages for the team as if they had played each game on a neutral floor against a completely average opponent. The normal average for a stat is calculated as in Section IV, where one simply sums up the results in each game and then divides by the number of games played for counting stats, or sums the number of makes and attempts and then divides for percentage stats. But the weighted averages goes through each game and neutralizes the stats and weights each stat by how good their opponent is at that stat.

The basic formula for Team A's weighted average for stat s would read

$$\bar{s}^{w,o/d}_{date,TeamA} = \frac{1}{n_{games}} \sum_{i:\ gamedate<date}^{n_{games}} s_i^{o/d} * r^{s,o/d}_{loc_i} * \frac{\bar{s}}{\bar{s}^{w,d/o}_{date,opp_i}}, \tag{12}$$

where $\bar{s}^{w,o/d}_{date,TeamA}$ is the weighted average for the offensive/defensive stat s on the indicated date, the sum is over all games played by Team A before the indicated date, and n_{games} is the number of those games; $s_i^{o/d}$ is the offensive/defensive stat s for Team A in game i, $r^{s,o/d}_{loc_i}$ is the neutral ratio for the offensive/defensive stat s based on the

Year	$\overline{oor_p}$	$r^{oor}_{p,home}$	$r^{oor}_{p,away}$	$r^{oor}_{p,neut}$	$\overline{oefg_p}$	$r^{oefg}_{p,home}$	$r^{oefg}_{p,away}$	$r^{oefg}_{p,neut}$	$\overline{oftmr_p}$	$r^{oftmr}_{p,home}$	$r^{oftmr}_{p,away}$	$r^{oftmr}_{p,neut}$	$\overline{oto_p}$	$r^{oto}_{p,home}$	$r^{oto}_{p,away}$	$r^{oto}_{p,neut}$
2003	0.341	0.967	1.036	0.995	0.496	0.970	1.028	1.014	0.265	0.923	1.094	0.990	0.186	1.050	0.951	1.020
2004	0.340	0.968	1.035	0.994	0.496	0.970	1.028	1.016	0.262	0.923	1.092	0.999	0.186	1.049	0.950	1.023
2005	0.339	0.968	1.034	0.996	0.496	0.971	1.028	1.017	0.260	0.922	1.094	0.996	0.186	1.050	0.950	1.023
2006	0.338	0.969	1.034	0.996	0.497	0.970	1.028	1.018	0.258	0.922	1.095	0.991	0.186	1.051	0.950	1.022
2007	0.338	0.970	1.032	0.993	0.497	0.970	1.027	1.019	0.254	0.921	1.095	0.998	0.186	1.051	0.950	1.019
2008	0.336	0.969	1.033	0.995	0.499	0.970	1.029	1.017	0.253	0.920	1.097	0.995	0.186	1.050	0.950	1.019
2009	0.334	0.970	1.033	0.993	0.500	0.969	1.030	1.015	0.252	0.920	1.099	0.988	0.185	1.051	0.950	1.019
2010	0.333	0.971	1.033	0.991	0.500	0.969	1.030	1.016	0.251	0.920	1.100	0.981	0.184	1.050	0.951	1.021
2011	0.331	0.971	1.032	0.992	0.499	0.969	1.030	1.013	0.253	0.922	1.097	0.982	0.182	1.049	0.951	1.020
2012	0.328	0.970	1.033	0.992	0.498	0.968	1.031	1.013	0.255	0.924	1.098	0.974	0.179	1.047	0.952	1.024
2013	0.325	0.971	1.032	0.993	0.496	0.969	1.030	1.013	0.255	0.924	1.097	0.975	0.177	1.045	0.953	1.026
2014	0.323	0.971	1.031	0.997	0.494	0.969	1.030	1.012	0.255	0.925	1.097	0.971	0.175	1.044	0.954	1.025
2015	0.321	0.970	1.032	0.997	0.495	0.969	1.030	1.008	0.261	0.926	1.093	0.978	0.171	1.044	0.955	1.020
2016	0.317	0.969	1.033	0.995	0.495	0.969	1.030	1.010	0.260	0.925	1.095	0.979	0.169	1.042	0.957	1.019

TABLE II. Predicted averages for each year from 2003-2016, as well as the *neutral ratios* at home, away, and at neutral sites, for each of the *four factors*. Each of the averages and ratios were generated using the 5 seasons previous to the year in question, with the exception of the 2003, 2004, and 2005 seasons which used only 2, 3, and 4 seasons respectively since I only have statistics starting from the 2001 season. The ratios are calculated with respect to the averages of all the games, not just those at neutral sites, which is why the ratios for neutral games are not exactly 1 for each stat.

location of game i; $\bar{s}$ is the league-wide average for stat s for the season in which the game is being played (as calculated in Section IV; and $\bar{s}^{w,d/o}_{date,opp_i}$ is the weighted average for the defensive/offensive stat s for Team A's opponent in game i. Note that the opponents' stats are flipped with respect to the Team A's, defense goes with offense, and offense goes with defense: this is because we are weighting Team A based on how well it is doing against its opponents abilities to oppose them; thus if we're looking at rebounding, we care about Team A's offensive rebounding abilities compared to what its opponents usually allow, and about what Team A allowed their opponents to do compared to what they usually get.

This looks similar to the SRS equation in that we need to calculate $\bar{s}^{w}_{date}$ for every team on a given date before we can calculate any of them. It might be possible to calculate these iteratively like we do for SRS, but I have found it more reliable to calculate things on the fly by looking at each opponent's season game by game as well. That leads us to the formula that I use, which is

$$\bar{s}^{w,o/d}_{date,TeamA} = \frac{1}{n_{games}} \sum_{i:\ <date}^{n_{games}} s_i^{o/d} * r^{s,o/d}_{loc_i} \frac{\bar{s}}{\frac{1}{n_{opp}} \sum_{j:\ <date}^{n_{opp}} s_j^{d/o} * r^{s,d/o}_{loc_j} \frac{\bar{s}}{\frac{1}{n_{oppopp}} \sum_{k:\ <date}^{n_{oppopp}} s_k^{o/d} * r^{s,o/d}_{loc_k}}}, \tag{13}$$

where I sum over the opponents' games weighted by their opponents (i.e. the opponents' opponents). While I could go deeper and weight the opponents' opponents' games by their opponents, I have left them unweighted, though I do still neutralize the stats. All of the sums go only over games that have been played before the date of the weighted average being calculated. We sum over all games by the opponents and the opponents' opponents, except for games played against Team A; this is done so as not to bias the results.

It is necessary to calculate all eight of these stats (the four factors on offense and defense) for each team on every day of the season in order to make predictions. Practically speaking, I calculate them starting about two weeks into the season, to allow teams to have played more than one opponent, and when a team's opponents have not played anyone else, I treat the weighting factor as 1. We should expect to see the weighting being most important near the beginning of the season, when teams' home/road schedules may be a bit skewed and the strength of schedule may be uneven. Further into the season, the weighted averages should reflect less of the location corrections and more pure strength of schedule.

Figures 1 and 2 show the offensive and defensive weighted averages for the 2015 North Carolina Tar Heels compared to their normal averages. North Carolina plays a particularly difficult schedule (the second most difficult in the league according to SOS) and therefore we should expect to see some significant difference between the two types of average, and in fact we do see that, although note that each of the averages are shifted by different amounts, and that amount

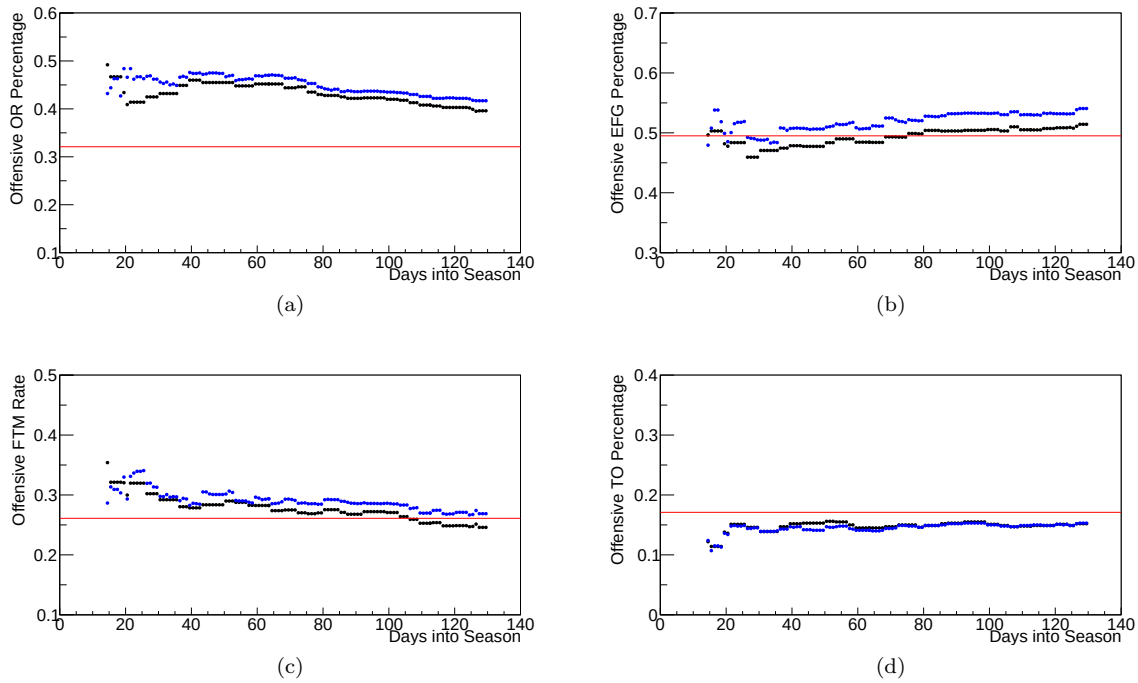


FIG. 1. Normal offensive averages (black) and weighted averages (blue) for the 2015 North Carolina Tar Heels for (a) offensive rebounding percentage, (b) effective field-goal percentage, (c) free-throw- made rate, and (d) turnover percentage. The averages start two weeks into the 2015 season and run through the beginning of the NCAA tournament. The red line is the predicted league average for the 2015 season. North Carolina is in the ACC and plays a difficult non-conference and conference schedule, therefore they tend to play good teams and so their stats are more impressive than they might seem looking only at the normal averages. Note that each stat gets its own weighting, thus they get a big bump in effective field-goal percentage because they play a lot of teams that are good at guarding against that, while their turnover percentage is nearly unchanged with the weighting. Their defensive stats can be found in Figure 2.

can evolve throughout the season. The difference between the two for the offensive effective-field-goal percentage indicates that they have been playing against teams that generally hold their opponents well below the average (which is indicated by the red line on the graphs), while their offensive turnover percentage is nearly unchanged, indicating that the teams they've played during the season have been just average at forcing turnovers. The data for the defensive free-throw-made rate clearly shows how the weights can evolve throughout the season, going from a very significant difference early on to nearly no difference at the end of the season.

Similar plots can be viewed for a wide variety of teams, those with fairly easy schedules, those with mediocre schedules, and those with a schedule that varies drastically throughout the season (many small-conference teams play very difficult early schedules that get significantly easier once conference play starts), and they all accurately reflect the changes throughout the year, though do note that the difference between the normal and weighted averages always represent how good their opponents were at that particular stat, not how good they are in general.

VI. WEIGHTED AVERAGE FUNCTIONS AND STANDARD DEVIATIONS

Now that we have developed the method for calculating the weighted averages it is useful to consider what happens when a team, Team A, with a weighted average in an offensive stat s, call it $\bar{s}_A^{w,off}$ plays a team, Team B, with the corresponding defensive stat $\bar{s}_B^{w,def}$. Obviously, in the real world anything can happen, but what I'm looking for is what would happen on average in this situation.

It makes sense to think of this in terms of the difference between the two weighted averages. If Team A has an offensive weighted average of 0.350 offensive rebounding percentage and Team B has a defensive weighted average of 0.350, and they play a large number of times, we would expect to see a range of values for Team A's offensive rebounding percentage that is centered around 0.350 with some characteristic spread. Likewise, if Team A and B have equal offensive and defensive weighted averages, we'd always expect to see the average of the results be close to

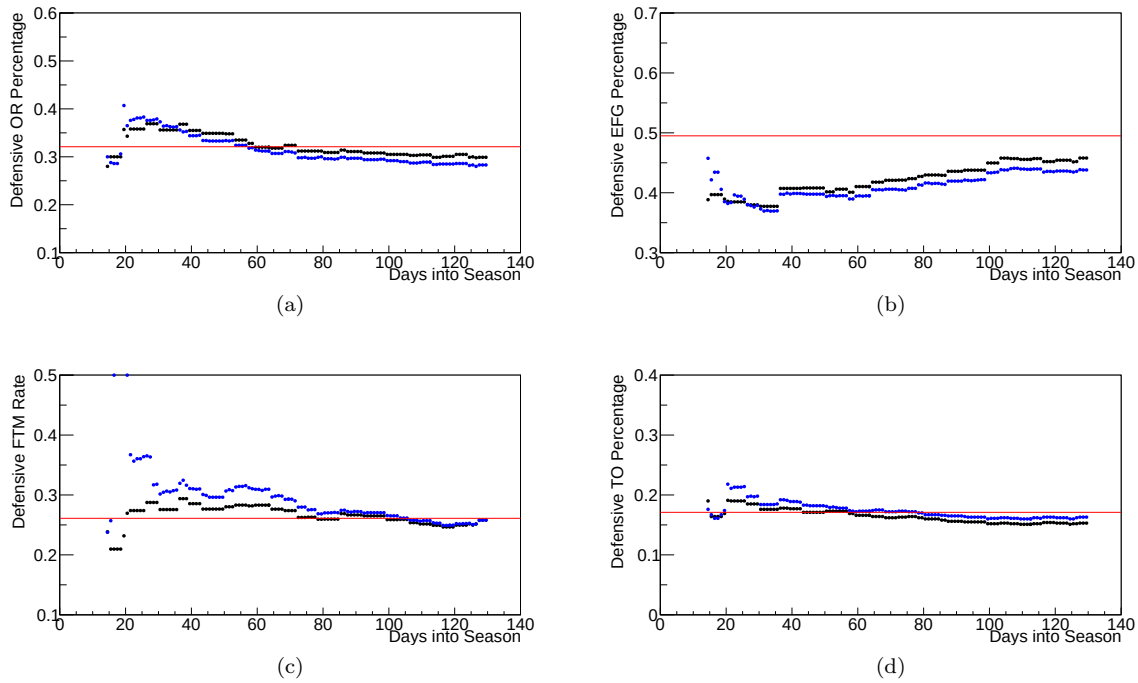


FIG. 2. Normal defensive averages (black) and weighted averages (blue) for the 2015 North Carolina Tar Heels for (a) offensive rebounding percentage, (b) effective field-goal percentage, (c) free-throw- made rate, and (d) turnover percentage. The averages start two weeks into the 2015 season and run through the beginning of the NCAA tournament. The red line is the predicted league average for the 2015 season. North Carolina is in the ACC and plays a difficult non-conference and conference schedule, therefore they tend to play good teams and so their stats are more impressive than they might seem looking only at the normal averages. Here we see that they give up less than the league average in effective field-goal shooting, and they have been doing it against opponents who normally shoot above average, therefore it is even more impressive. The free-throw-made rate shows that the weighting can change drastically over the course of the season, starting out quite significant and dropping down to nearly no weighting. Their offensive stats can be found in Figure 1.

that average. But what happens when the two are different?

To study this problem I've looked at each of the four stats separately (I only predict the offensive stats because predicting the defensive stats would be redundant), trying to determine a function of the difference between Team A's offense and Team B's defense, yielding a multiplicative factor to apply to Team A's offensive weighted average to give Team A's expected offensive performance in that stat. Generally speaking, we should expect this function to be roughly 1 when the difference is 0, greater than 1 when Team B's defense gives up more than Team A usually gets, and less than 1 when Team B's defense gives up less than Team A usually gets.

In practice I calculated these functions for each year individually by looking at every game in the 5 years preceeding it that I had weighted averages calculated for (2003 on), selecting only the games taking place between December 1^{st} of that season and the start of the tournament (to avoid wonky weighted averages early in the season, and because I don't calculate weighted averages throughout the tournament). For each game I determined a bin based on Team A's offensive weighted average for the given stat and Team B's defensive weighted average for that stat. I also recorded Team A's actual neutralized offensive stats in that game with respect to its weighted average going into the game. I then looked at each bin and and calculated the difference between the weighted averages and the average ratio of actual game performance to weighted average. This led to multiple values, with corresponding standard deviation, at nearly every possible value of the difference between the two (the max difference were usually on the order of 10%, with smaller differences being much more common), so I took the weighted average of those values and calculated a corresponding error, following the methods of Taylor [1]. I then graphed these values, as you can see in Figure 3, and saw that all four of the graphs looked like they could be very neatly fit with a linear function, and did so. These four linear functions are then used as my function to predict Team A's offensive performance given Team A's offensive weighted average before the game and Team B's defensive weighted average (i.e. those functions needed in Equation 1.

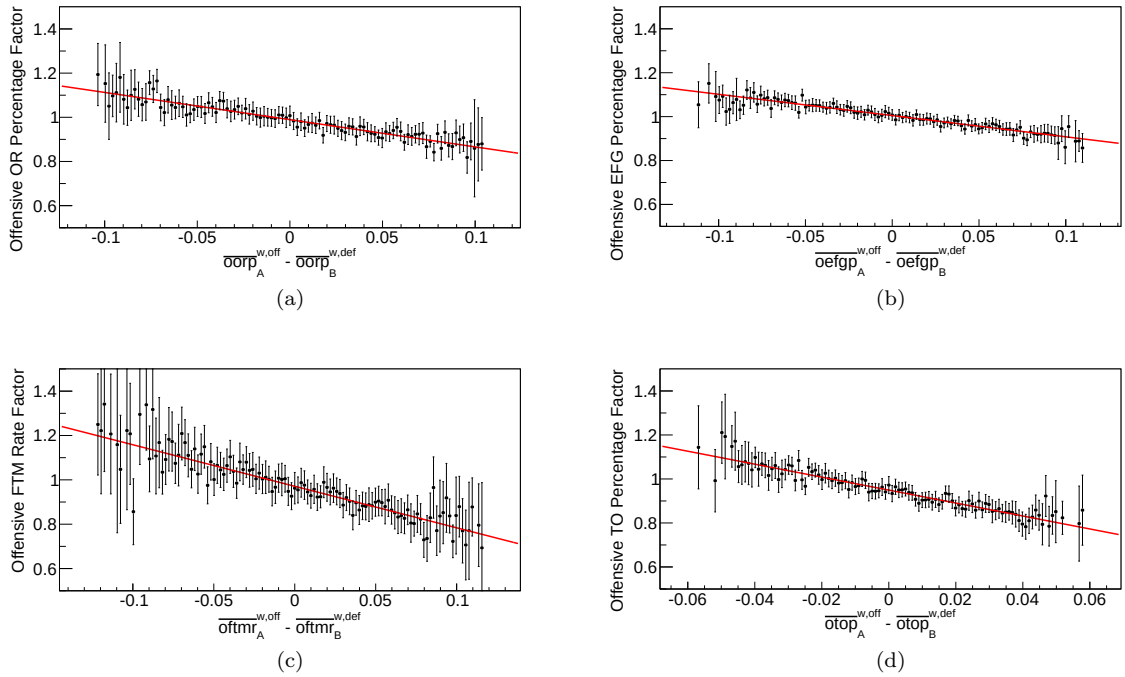


FIG. 3. The weighted average functions for the 2015 season, relating a multiplicative factor for Team A's offense to the difference between the weighted averages for Team A's offensive stat s, and Team B's defensive stat s, for (a) offensive rebounding percentage, (b) effective field-goal percentage, (c) free-throw-made rate, and (d) turnover percentage. Each point on the plots is the average of all of the games played with that difference in the statistics, where the factor on the y-axis relates how well the teams with that difference did in those games with respect to their actual weighted average going into the game. The error bars are related to the standard deviations of the averages of the games.

Thus we can say that

$$P(s_{A,date}|\bar{s}^{w,off}_{A,date}, \bar{s}^{w,def}_{B,date}, loc_A) = \frac{Par(0) + Par(1) * (\bar{s}^{w,off}_{A,date} - \bar{s}^{w,def}_{B,date})}{r^{s,off}_{loc_A}} \tag{14}$$

is our weighted average function for each stat s, where Par(0) is the x-intercept of the calculated function and Par(1) is the slope, and the entire result is divided by the neutral ratio to *deneutralize* it, that is, to put it in the appropriate location for the game for Team A. Each of Team A's four offensive stats for the game can then be predicted once we have these functions, and obviously Team B's stats can be predicted for the game by simply reversing Team A and B in the above functions (so we consider Team B's offense and Team A's defense.

Looking at Figure 3, it is clear that some of the slopes are more pronounced than others, which you may also note by looking at Table III, which gives the weighted average functions for every season from 2005-2016. The free-throw-made rate and the turnover percentage stats are clearly more affected by differences in defense than the offensive rebounding and especially effective field-goal percentages. This indicates that teams that teams that thrive on shooting are not very likely to be affected by a good defense, but that defense is much more important when it comes to forcing turnovers and getting chances from the free-throw line.

VII. THE GAME FUNCTION

Now that we have the weighted averages and the functions to predict game performance for each team based on these weighted averages, all that we have left to do is to determine the weights for the game function. Fully written

Year	ORP function	EFGP function	FTMR function	TOP function
2005	0.988 - 1.235*x	1.005 - 1.021*x	0.965 - 2.293*x	0.946 - 2.971*x
2006	0.996 - 1.299*x	1.010 - 0.993*x	0.981 - 2.244*x	0.947 - 2.118*x
2007	0.992 - 1.255*x	1.010 - 0.996*x	0.984 - 1.973*x	0.948 - 2.091*x
2008	0.993 - 1.228*x	1.009 - 0.987*x	0.993 - 2.027*x	0.942 - 2.006*x
2009	0.994 - 1.308*x	1.008 - 0.938*x	0.995 - 2.060*x	0.942 - 2.390*x
2010	0.993 - 1.278*x	1.007 - 0.935*x	0.989 - 2.119*x	0.945 - 2.588*x
2011	0.989 - 1.121*x	1.006 - 0.953*x	0.982 - 2.185*x	0.944 - 2.603*x
2012	0.988 - 1.158*x	1.007 - 0.934*x	0.984 - 2.063*x	0.945 - 2.564*x
2013	0.988 - 1.250*x	1.007 - 0.934*x	0.976 - 1.914*x	0.947 - 2.908*x
2014	0.989 - 1.184*x	1.008 - 0.911*x	0.974 - 1.951*x	0.948 - 2.824*x
2015	0.989 - 1.233*x	1.005 - 0.968*x	0.971 - 1.873*x	0.949 - 2.945*x
2016	0.990 - 1.223*x	1.008 - 0.994*x	0.963 - 1.959*x	0.948 - 3.178*x

TABLE III. Weighted average functions for each year from 2005 to 2016. The data for each year is the previous 5 seasons' data, with the exception of 2005, 2006, and 2007, which use the previous 2, 3, and 4 seasons' data since I only have weighted averages calculated back to 2003. The x-intercepts for the offensive rebounding percentages and effective field-goal percentages are very close to 1 indicating that equally-matched teams will lead to a factor of 1; but the free-throw-made rate and turnover functions have an intercept below 1, indicating that evenly-matched teams will lead to the team performing worse than usual, this may be an indication that something is wrong, or simply that the defense has an advantage when the teams are evenly matched.

out the game function is

$$\begin{aligned} gameScore(TeamA, TeamB, date) = {} & w_{oorp}(P(oorp_{O,A}|\overline{oorp}_{O,A,date}, \overline{oorp}_{D,B,date}, loc_A) - \\ & P(oorp_{O,B}|\overline{oorp}_{O,B,date}, \overline{oorp}_{D,A,date}, loc_B))/\sigma_{oorp,date} + \\ & w_{oefgp}(P(oefgp_{O,A}|\overline{oefgp}_{O,A,date}, \overline{oefgp}_{D,B,date}, loc_A) - \\ & P(oefgp_{O,B}|\overline{oefgp}_{O,B,date}, \overline{oefgp}_{D,A,date}, loc_B))/\sigma_{oefgp,date} + \\ & w_{oftmr}(P(oftmr_{O,A}|\overline{oftmr}_{O,A,date}, \overline{oftmr}_{D,B,date}, loc_A) - \\ & P(oftmr_{O,B}|\overline{oftmr}_{O,B,date}, \overline{oftmr}_{D,A,date}, loc_B))/\sigma_{oftmr,date} + \\ & w_{otop}(P(otop_{O,A}|\overline{otop}_{O,A,date}, \overline{otop}_{D,B,date}, loc_A) - \\ & P(otop_{O,B}|\overline{otop}_{O,B,date}, \overline{otop}_{D,A,date}, loc_B))/\sigma_{otop,date} + \\ & w_{srs}(srs_{A,date} - srs_{B,date})/\sigma_{srs,date} \end{aligned} \tag{15}$$

where each of the five terms in the equation is given a weighting factor, w_s, and the term itself is the difference between the predicted offensive production for the two teams, based on their weighted averages on the date of the game and the location, and each term is divided by the standard deviation of term as measured in the years which will be used as data, indicated by the $\sigma_{s,date}$, where the date simply implies that we are using the standard deviations for the correct year. Dividing by the standard deviations is done to ensure that each term contributes roughly equally on average before weighting is applied, otherwise the relative strength of the weights could not be assessed and the turnover rate, which is generally around 20% would get dominated by the effective field-goal percentage, which is generally near 50% and so on. I have included a term for the SRS although we will calculate the weights for just the four factors, with w_{SRS} forced to 0, before allowing it to vary freely.

The results of this function are pretty easy to understand: if the value is positive, Team A is favored to win; if the value is negative Team B is favored to win, and the larger the absolute value of those numbers is, the more likely the team is to win. The function is symmetric, meaning that if you plug in Team A and Team B on a given date and get a gameScore of x, you will get -x by reversing the teams, i.e. having Team B play Team A on the same date. Once the weights have been determined, it is possible to develop an approximate mapping of gameScore onto likelihood of winning, which we will do later in this section, but for now just remember that larger values for gameScore indicate a higher likelihood of winning. A gameScore of 0 would mean the game is a complete toss-up: 50/50% likelihood of either team winning.

A. Determining the Best Weights

The process for determining the best weights to use is to create a function using the value of Equation 15 and which also reflects what percentage of games are predicted correctly using this set of weights and then maximize it. The function I have used (and it may not be the best one, but it works well enough for me) is to simply count up the number of correctly picked games and divide the result by the total number of games analyzed. This gives a winning percentage for the games with this particular set of weights, and then you try to maximize that result. In practice, I actually minimize the negative of that winning percentage, since minimization techniques are more common.

I use the MIGRAD fitter from the TMinuit package of the statistical analysis software program called ROOT to run my fits. As with most everything else, I use 5 seasons worth of games as my data for the fits (except when not available). I only include games from the 1^{st} of December until the beginning of the tournament. I also include only one version of each game; that is, I include Team A vs Team B, not Team B vs Team A. Since the game function is symmetric, these would just give me the same information while doubling the number of games to be processed. I have ensured this by requiring that Team A's name be greater than Team B's name lexigraphically. I usually use about 500 completely random starting positions for the 4 (or 5 depending on if I'm including SRS) variables and then choose the set of weights that gave the best overall winning percentage. The fits were completely unbounded, although my starting values were always between 0 and 1, and the final results are always normalized for ease of determining the relative importance of the weights. Negative weights are possible and expected for certain terms (like the turnover percentage); when normalizing, I consider the absolute value of those terms.

B. Weights Without SRS

Table IV displays the best-fit weights, where I force w_{SRS} to be zero, for each year from 2007 to 2016, along with the winning percentage achieved by the fit with those weights, and the actual winning percentage achieved by using those weights on the games in the years they were predicting (which will be discussed later). Notice that the weights for FTMR and TOP are negative, incidating that it is generally bad to have higher values of these stats than your opponent: this makes complete sense for the turnover rate, as turnovers are in fact bad, but I'm not sure exactly why it would be bad to have a higher free-throw-made rate than your opponent, but that's why I run the fits instead of just guessing at the values myself.

Looking at the fit values for the weights, it is interesting to note that the relative ordering of the importance of the stats is very similar in each of them, with effective field-goal percentage being the most important with nearly half the weight in each year, followed by offensive rebounding percentage and turnover percentage, which are relatively similar in importance, and then lastly by the free-throw-made rate. It should be noted that Dean Oliver, who originally identified the *four factors* assigned them the relative importance of EFGP (40%), TOP (25%), ORP (20%), FTMR (15%), which is not that terribly different than some of the weights that I determined, although he did not allow for year-to-year variation [2].

Also note that the fit winning percentage is fairly stable over the years at being able to predict approximately 70% of the games correctly in the fit seasons simply based on their stats.

C. Weights With SRS

The previous results were obtained by looking only at the four factors. It is impressive that we are able to predict 70% of games correctly using only these four statistics and absolutely no reference to a team's winning percentage or even a reference to the number of points they scored. However, there are other ways of making predictions for the victor in a game, and we have already discovered one of them in the SRS. As discussed in Section III A the SRS values can be used to make predictions about a game as well. If Team A's SRS is greater than Team B's, then Team A should be favored to win, and vice versa. Over the period of 2007-2015 this method correctly picked the winners of games at a rate of 0.705, just slightly better than the 0.702 that the gameScore method produced.

As has already been seen in Equation 15, we can include the SRS with its own weighting factor into the gameScore function. Doing so, and then refitting the data using the same method as above, we obtain the weights listed in Table V. The SRS weight generally settles near 50% of the total weight, while the others fall into a similar dominance structure that they had when w_{SRS} was 0. Interestingly, while w_{otop} is still negative, as we would expect it to be, w_{oftmr} is now mostly positive, and in the case of 2015 very positive at near 11% (or 22% of the total non-SRS weight). Why it would be consistently negative when w_{SRS} is 0, and then only sometimes negative when w_{SRS} is non-zero is a mystery to me.

Year	w_{oorp}	w_{oefgp}	w_{oftmr}	w_{otop}	Fit Winning Pct	Actual Winning Pct
2007	0.2241	0.4228	-0.0738	-0.2793	0.706	0.706
2008	0.2223	0.4397	-0.0671	-0.2708	0.705	0.701
2009	0.1621	0.4815	-0.1262	-0.2302	0.705	0.709
2010	0.2004	0.5090	-0.0734	-0.2172	0.705	0.714
2011	0.1473	0.5425	-0.0929	-0.2174	0.706	0.695
2012	0.1783	0.4753	-0.1269	-0.2195	0.707	0.705
2013	0.2349	0.4926	-0.1066	-0.1659	0.708	0.700
2014	0.2252	0.5264	-0.0578	-0.1906	0.708	0.691
2015	0.2588	0.5125	-0.0572	-0.1715	0.705	0.697
2016	0.1565	0.4950	-0.1491	-0.1994	0.700	N/A

TABLE IV. The weights for each of the four factors in each season from 2007-2016, along with the maximum value of the winning percentage that the fit reached. Each year's fit used as data the 5 seasons previous to the year in question, except for 2007, 2008, and 2009, which used 2, 3, and 4 seasons respectively since the earliest season for which I have calculated the weighted average functions is 2005. Notice that effective field-goal percentage is clearly the most important stat, as it is near 50% for all of the years; offensive rebounding and turnover percentage are the next most important, while the free-throw-made rate consistently has the smallest weights. Notice also that turnover percentage and free-throw-made rate's weights are negative, indicating that high values of these are bad: this makes sense with turnover, as having a lot of turnovers is a bad thing and forcing turnover is a good thing, but it makes less sense with the FTMR. Combining all of the years together yields an overall winning percentage of 0.702.

Year	w_{oorp}	w_{oefgp}	w_{oftmr}	w_{otop}	w_{SRS}	Fit Winning Pct	Actual Winning Pct
2007	0.1841	0.1602	0.0072	-0.1469	0.5016	0.732	0.720
2008	0.1715	0.1642	-0.0385	-0.1579	0.4679	0.728	0.724
2009	0.1806	0.1534	-0.0083	-0.1615	0.4962	0.728	0.731
2010	0.1711	0.1509	-0.0028	-0.1638	0.5115	0.729	0.737
2011	0.1220	0.2698	0.0286	-0.1320	0.4477	0.731	0.733
2012	0.0740	0.2487	0.0394	-0.1528	0.4850	0.733	0.749
2013	0.0827	0.2350	0.0314	-0.1580	0.4929	0.738	0.726
2014	0.0547	0.1853	0.0903	-0.1436	0.5261	0.738	0.717
2015	0.0527	0.1385	0.1194	-0.1846	0.5049	0.737	0.718
2016	0.0589	0.2290	0.0692	-0.1301	0.5128	0.731	N/A

TABLE V. The weights for each of the four factors plus SRS in each season from 2007-2016, along with the maximum value of the winning percentage that the fit reached. Each year's fit used as data the 5 seasons previous to the year in question, except for 2007, 2008, and 2009, which used 2, 3, and 4 seasons respectively since the earliest season for which I have calculated the weighted average functions is 2005. Notice that the SRS weight is roughly 0.5 in each of the years, that effective field-goal percentage is still the most important of the four factors, that the turnover percentage weights are all still negative, and that for some reason only some of the free-throw-made rate weights are negative. All of the FTMR weights were negative when w_{SRS} was 0, and now only some of them are, very curious. Combining all of the years together yields an overall winning percentage of 0.728.

D. Converting gameScore To a Likely Winning Percentage

As I mentioned earlier, the greater the absolute value of the gameScore, the more likely the team is to win the game, but we would of course like to have a way to quantify how likely a team is to win. In order to do this I simply ran through every game in the fit years (i.e. 2011-2015 for the weights used in 2016) and calculated their gameScore with the best-fit weights, binned those results into two different histograms, one if the gameScore correctly predicted the winner and one if it was incorrect, and then fit those two histograms with Gaussian functions. Using those two functions I am able to determine the likelihood of getting the game right or wrong at any value of gameScore, as well as an associated error; this allows me to successfully convert a gameScore into a likely winning percentage.

Figure 4 shows this function for the year 2016, derived using the games from the years 2011-2015. The function is in blue, while the function used to determine the associated error at each gameScore is in red. The error is calculated

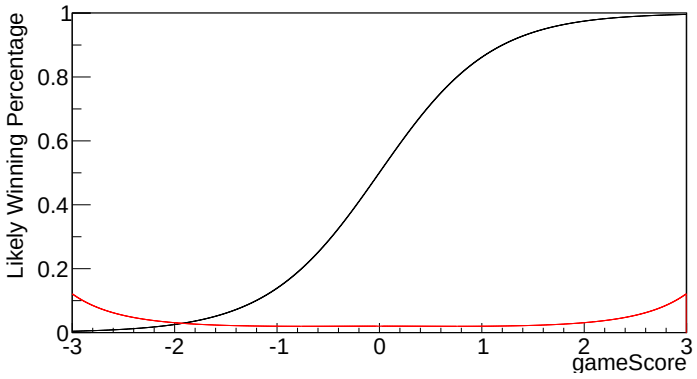


FIG. 4. Likely winning percentage (black) vs gameScore function for the year 2016 along with the associated uncertainty (red). This figure shows the conversion from gameScore to likely winning percentage derived by looking at all of the games (between Dec 1^{st} and the beginning of the NCAA tournament) in the years 2011-2015. The error is derived using binomial statistics.

using binomial statistics and is defined as [1]

$$\sigma = \sqrt{\frac{p(1-p)}{numGames}}, \tag{16}$$

where p is the likely winning percentage and numGames is the value of the function used to fit the histogram of correctly predicted games evaluated at that gameScore plus the value of the function used to fit the histogram of the incorrectly predicted games evaluated at that gameScore. The rise that occurs at the edges occurs because there are very few games that have gameScores with an absolute value greater than 2, so the denominator is getting smaller. These errors represent the average error that I have seen among games with a similar gameScore.

VIII. RESULTS

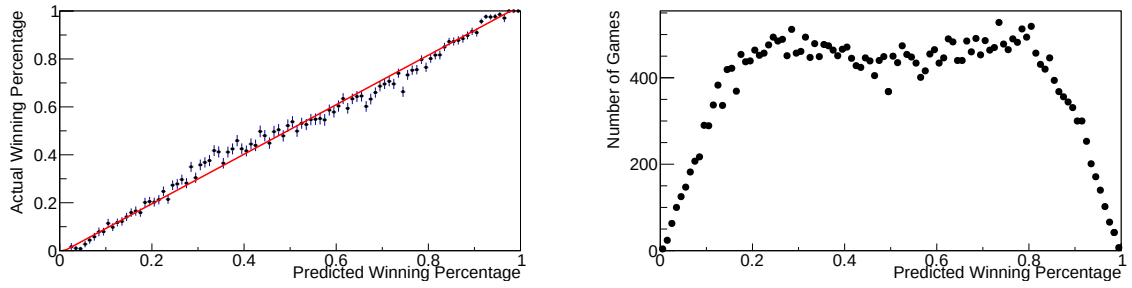
A. Without SRS

The results of using the gameScore predictions with $w_{SRS} = 0$, and their associated predicted winning percentages, can be seen for all games played between December 1^{st} and the beginning of the tournament in the years 2007-2015 in Figure 5, and the actual winning percentages in each year can be seen in Table IV. Overall, this method predicts 70.2% of the games correctly (27041 out of 38530) in the years 2007-2015. The figure also shows that the predicted winning percentage is a very good predictor of the actual winning percentage, which is a very positive sign. A linear function was fit to the measured data points and resulted in an intercept of 0 and a slope of 1.03, just barely off from the ideal value of 1. There is a slight wave to the the values, in that they don't all fall exactly on the line, but it is very close.

It is true that in college basketball the home team wins 66% of their games across the board, and so making predictions solely based on the location of the game will give you a winning percentage of near 66% (although it would be unable to make predictions in a neutral-site game which account for nearly 10% of the total number of games in a season). It is also true that the method I have presented does take into account location, with the neutral ratios. However, the effect of the location corrections are rather small, and the predictions fare only slightly worse when ignoring them altogether; and in any case the extra 4%, plus all of the neutral-site games indicate that this method is performing admirably.

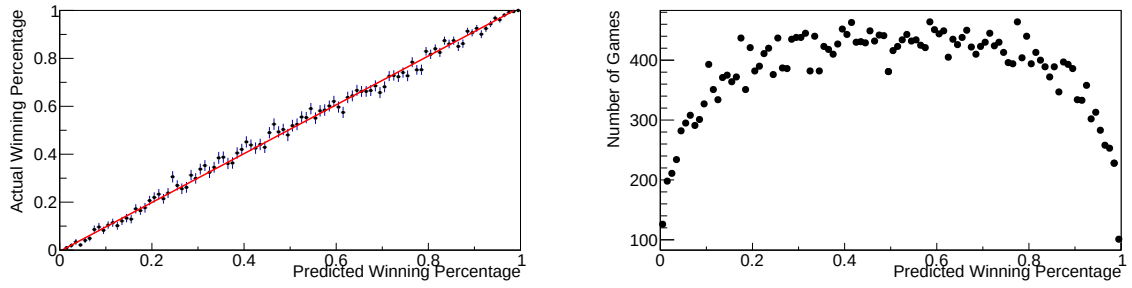
B. With SRS

The results of using the gameScore predictions with $w_{SRS} \neq 0$, and their associated predicted winning percentages, can be seen for all games played between December 1^{st} and the beginning of the tournament in the years 2007-2015 in Figure 6, and the actual winning percentages in each year can be seen in Table V. Overall, this method predicts 72.8% of the games correctly (28061 out of 38530) in the years 2007-2015. Figure 6 shows that these weights also lead to a



(a) (b)

FIG. 5. (a) Actual winning percentage vs predicted winning percentage for all games played between December 1^{st} and the beginning of the tournament in the years 2007-2015 using the weights described in Table IV with $w_{SRS} = 0$ along with (b) the number of games analyzed in those years as a function of predicted winning percentage. The error bars in (a) are the binomial standard deviation for y, and the bin width for x. The red line is a linear fit line to those results; ideally the x-intercept should be 0 and the slope should be 1, the actual values are 0 and 1.03, indicating that the predictions and their associated likely winning percentages are accurate predictors.



(a) (b)

FIG. 6. (a) Actual winning percentage vs predicted winning percentage for all games played between December 1^{st} and the beginning of the tournament in the years 2007-2015 using the weights described in Table V along with (b) the number of games analyzed in those years as a function of predicted winning percentage. The error bars in (a) are the binomial standard deviation for y, and the bin width for x. The red line is a linear fit line to those results; ideally the x-intercept should be 0 and the slope should be 1, the actual values are 0 and 1.02, indicating that the predictions and their associated likely winning percentages are accurate predictors. Note that the number of games is more evenly distributed among the predicted winning percentages than when $w_{SRS} = 0$.

very good predictor, where the predicted winning percentage correlates very well with the actual winning percentage, with an x-intercept of 0 and a slope of 0.01. We also note that the number of games is more evenly distributed among the predicted winning percentages than when $w_{SRS} = 0$ which showed more of a humped structure, with fewer games predicted to be 50/50 than 25/75. The SRS weight generally settles near 50% of the

We can see from Table V that predicted winning percentages are significantly higher, by about 3%, and the actual winning percentages are also 2-3% higher. The winning percentage over all of the seasons from 2007-2015 is 0.728 compared to 0.702, indicating that this is a significant improvement. This set of predictions, with $w_{SRS} \neq 0$, will be my preferred method of making predictions for regular season games.

C. Discussion

The benefit of having both sets of predictions is that they tell two different stories, sometimes with the same end (i.e. predicting the same team will win), but valuing different statistics. There are some games where Team A's SRS may just completely dwarf that of their opponent, and yet they lose. When you look at the prediction made with $w_{SRS} = 0$, you may see that statistically the teams were quite similar, but perhaps Team B played in an easier conference, or Team A played such difficult opponents that simply playing them raised their SRS significantly,

regardless of their actual point differential. If you take two copies of the exact same team and have one play a very easy schedule and the other play a very difficult schedule, they will end up with wildly disparate SRS values, but their weighted averages should be similar because it is designed that way.

The SRS method also doesn't allow Team B to influence Team A in any way. Your SRS is your SRS and it doesn't matter what your oppoent's SRS is, it won't affect yours. The gameScore method allows the two participants to affect one another via the weighted average functions.

The NCAA tournament has a fair number of upsets every year and coming up with a way to predict them is important if you want to do well with your bracket. The seeds in the tournament correlate very strongly with the teams' SRS values, therefore upsets are almost by definitions going contrary to what the SRS would predict. Even if the gameScore method with $w_{SRS} = 0$ only predicts that the game will be closer than predicted by the $w_{SRS} \neq 0$ method, that is valuable information.

IX. THE SIMPLE RATING SYSTEM PREDICTIONS

X. CONCLUSION

---

[1] John R. Taylor, *An Introduction to Error Analysis, Second Edition*, University Science Books (1982).
[2] *http://www.basketball-reference.com/about/factors.html*