

Improved NCAA Basketball Tournament Modeling via Point Spread and Team Strength Information

BRADLEY P. CARLIN
University of Minnesota

October 25, 1994

Abstract

Several models for estimating the probability that a given team in an NCAA basketball tournament emerges as the regional champion were presented by Schwertman et al. (1991). In this paper we improve these probability models by taking advantage of external information concerning the relative strengths of the teams and the point spreads available at the start of the tournament for the first round games. The result is a collection of regional championship probabilities that are specific to a given region and tournament year. The approach is illustrated using data from the 1994 NCAA basketball tournament.

1 Introduction

For many years, the analysis of sports statistics was limited to huge tables of sums, means, and the occasional descriptive display. More recently, however, statisticians have begun to undertake more sophisticated analyses of these data. In teaching, they can provide interesting and more easily grasped illustrations of important concepts (see e.g. Albert, 1993). In research, they often enable testing of new approaches for handling difficult modeling scenarios, such as nonrandomly missing data (Casella and Berger, 1994) and time-dependent selection and ranking (Barry and Hartigan, 1993). And of course, in many cases the data are interesting in and of themselves; recent papers have investigated the existence of the “hot hand” in basketball (Tversky and Gilovich, 1989; Larkey et al., 1989) and the likelihood of “Shoeless” Joe Jackson’s complicity in the famous “Black

Sox” scandal (Bennett, 1993). The recent creation of a new section of the American Statistical Association devoted to sports statistics provides further testimony to their increasing popularity.

Perhaps the oldest inferential problem related to sports statistics is that of predicting the ultimate winner of some event, based on whatever information is available concerning the various competitors. In the realm of college basketball, the most talked-about such event is the NCAA men’s tournament, held every year in March and early April. In this tournament, 64 teams (some invited by a selection committee, others receiving automatic bids thanks to their having won their own conference tournaments) are divided into four regional tournaments (West, Midwest, East, and Southeast) of 16 teams each. Some effort is made by the committee to balance the overall team strength in each region, while at the same time place teams in the appropriate geographical region. The teams in each region are then “seeded” (ranked) based on their relative strengths as perceived by the committee. In a given region, the tournament begins by having the strongest team (seed 1) play the weakest team (seed 16), the second-strongest (seed 2) play the second-weakest (seed 15), and so on. The winners of these eight first round games then play off in a predetermined order (e.g., the 1–16 winner plays the 8–9 winner) in four second round games, and so on until a single regional champion is determined after four rounds of play. Finally, the four regional champions face each other in fifth and sixth round games to determine a single national champion. For the time being, we focus only on prediction of the regional champions (the “Final Four”).

In a recent paper, Schwertman, McCready and Howard (1991) consider three alternatives for specifying a 16 by 16 matrix P of regional win probabilities. That is, $P(i, j)$ is the probability that seed i defeats seed j in a contest between the two on a neutral court, where of course $i \neq j$ and $P(j, i) = 1 - P(i, j)$. Together with the assumption that the games are independent, they derive the probability that seed i wins the region for $i = 1, \dots, 16$ using elementary (though fairly tedious) calculations implemented in a Fortran program. Their models for $P(i, j)$ are somewhat

ad-hoc, though the most sophisticated (and best fitting) plausibly assumes a normal distribution of national team-strengths, with the 64 tournament teams comprising the upper tail of this distribution. Subsequent work by Schwertman, Schenk, and Holbrook (1993) refines the approach by using past NCAA tournament data to fit linear and logistic regression models for $P(i, j)$ as a function of the difference in either team seeds or normal scores of the seeds.

In this paper, we extend this approach by taking advantage of valuable external information available at the tournament’s outset. Specifically, we may employ any of the various computer rankings of the teams, such as RPI index, Sagarin ratings, and so on, which typically arise as a linear function of several variables (team record, opponents’ records, strength of conference, etc.) monitored over the course of the season preceding the tournament. These rankings provide more refined information concerning relative team strengths than is captured by the regional seedings. Such rankings also enable differentiation between identically seeded teams in different regions.

A second source of information for the first round games is the collection of point spreads offered by casinos and sports wagering services in states which allow gambling on college basketball. A point spread is a predicted amount by which one team (the “favorite”) will defeat the other (the “underdog”); gamblers may bet on whether the favorite’s actual margin of victory will exceed the point spread (“cover the spread”) or not. Point spreads are potentially even more valuable as pregame data than computer rankings, since besides team strengths they account for game- and time-specific information, such as injuries to key players. Previous work by Harville (1980) and Stern (1992) shows that point spread information is the “gold standard” against which all other pregame information as to outcome must be judged.

Unfortunately, point spreads for potential games in rounds 2 through 4 will be unavailable at the tournament’s outset. To remedy this, in Section 2 we describe an approach for imputing point spreads for these later games, and subsequently converting the resulting point spread matrix into

the win probability matrix P . Section 3 applies our approach to data from the 1994 NCAA men's basketball tournament. Finally, Section 4 summarizes our findings and comments briefly on the prediction of the ultimate NCAA basketball national champion.

2 Determination of Win Probabilities

Working with data from three seasons of professional football, Stern (1991) showed that the favored team's actual margin of victory, R , was reasonably approximated by a normal distribution with mean equal to the point spread, Y , and standard deviation $\sigma = 13.86$. That is,

$$Pr(\text{favorite defeats underdog}) = Pr(R > 0) \approx \Phi(Y/\sigma), \quad (1)$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution. This rather surprising result indicates that the group of bettors who determine the point spreads are, on the average, correct in their predictions of game outcome. (Note that it would be wrong to give too much credit for this accuracy of the point spread to the bookies, who merely set an initial spread and subsequently raise or lower it so that roughly the same total amount is bet on both the favorite and the underdog.)

Subsequent unpublished analysis by Stern of two seasons of professional basketball data indicates that (1) again holds, this time with $\sigma = 11.5$. Intuition suggests that this σ value may be a bit large for our purposes, since college basketball is generally a lower scoring game and we would expect the variability in the victory margin to increase with the total points scored. Indeed, data from the first four rounds of the 1994 NCAA tournament produce a value of $\hat{\sigma} = 8.83$, though using this precise value in our analysis would of course be unfair since it was unavailable at the tournament's outset. In what follows, we take $\sigma = 10$ as a reasonable and somewhat conservative compromise.

| region | $j - i$ | $S(i) - S(j)$ | Y_{ij} | R_{ij} | region | $j - i$ | $S(i) - S(j)$ | Y_{ij} | R_{ij} |
|---------|---------|---------------|----------|----------|-----------|---------|---------------|----------|----------|
| West | 15 | 20.28 | 20 | 23 | East | 15 | 20.37 | 25 | 20 |
| West | 13 | 18.43 | 24 | 26 | East | 13 | 15.78 | 18.5 | 18 |
| West | 11 | 11.95 | 18 | 9 | East | 11 | 10.98 | 10.5 | 2 |
| West | 9 | 9.81 | 11 | 14 | East | 9 | 9.31 | 11.5 | 22 |
| West | 7 | 4.78 | 6 | -4 | East | 7 | 2.00 | 4.5 | 12 |
| West | 5 | 6.39 | 8.5 | 14 | East | 5 | 5.53 | 5 | -10 |
| West | 3 | 1.20 | 4 | 3 | East | 3 | 2.53 | 4 | -5 |
| West | 1 | 1.64 | 4 | -8 | East | 1 | -0.19 | -3.5 | -3 |
| Midwest | 15 | 23.59 | 28 | 15 | Southeast | 15 | 20.87 | 23 | 31 |
| Midwest | 13 | 13.32 | 16 | 18 | Southeast | 13 | 18.32 | 20.5 | 12 |
| Midwest | 11 | 8.97 | 11.5 | 4 | Southeast | 11 | 17.31 | 18 | 13 |
| Midwest | 9 | 8.58 | 12 | 10 | Southeast | 9 | 9.90 | 10.5 | 29 |
| Midwest | 7 | 4.45 | 4 | -10 | Southeast | 7 | 5.71 | 9 | 10 |
| Midwest | 5 | 5.71 | 7 | 14 | Southeast | 5 | 4.96 | 7 | 22 |
| Midwest | 3 | 1.04 | -1.5 | -8 | Southeast | 3 | 2.60 | 2 | 11 |
| Midwest | 1 | 1.43 | 2 | -7 | Southeast | 1 | 5.63 | 5 | -6 |

Table 1: Data from Round 1 of the 1994 NCAA Tournament

Hence for a given region, we may use equation (1) with the point spreads for the first round games to determine $P(i, 17 - i)$, $i = 1, \dots, 16$. But this fills in only the antidiagonal of P ; how should we determine the remaining entries? A natural solution would be to obtain a general prediction equation for the point spread y in terms of the difference in team seeding (perhaps after some suitable transformation), and then again use (1) to complete the P matrix. Relevant data for this calculation from the 32 first round games in the 1994 tournament are displayed in Table 1. Besides the seeding differences $(j - i)$ and point spreads Y_{ij} for each game matching seeds i and j where $i < j$, the table also shows the actual victory margin R_{ij} for comparison. Our Y_{ij} values were obtained immediately prior to the beginning of tournament play from the Thursday morning edition of a local newspaper, which in turn got them from a prominent Las Vegas oddsmaker. A negative value of Y_{ij} implies that the team with the poorer seeding (team j) was favored by the bettors to win the game; a negative value of R_{ij} indicates that the result was in fact a victory by the poorer seed (an “upset”).

The first column of plots in Figure 1 shows the results of regressing point spread on squared seeding difference. The fitted regression line obtained is $\hat{Y}_{ij} = 2.312 + .100(j - i)^2$, where $i < j$. The extremely close agreement between this fitted regression line (solid line in upper panel) and the lowess smoothing line (dotted line) indicates a high degree of linearity on this scale, and the standardized residual plot in the lower panel does not indicate any failure in the usual regression assumptions. Further, the R^2 value of .883 indicates reasonably good fit. Note that the data line up in 8 vertical columns, since there are four games pitting seed i against seed $(17 - i)$ for $i = 1, \dots, 8$.

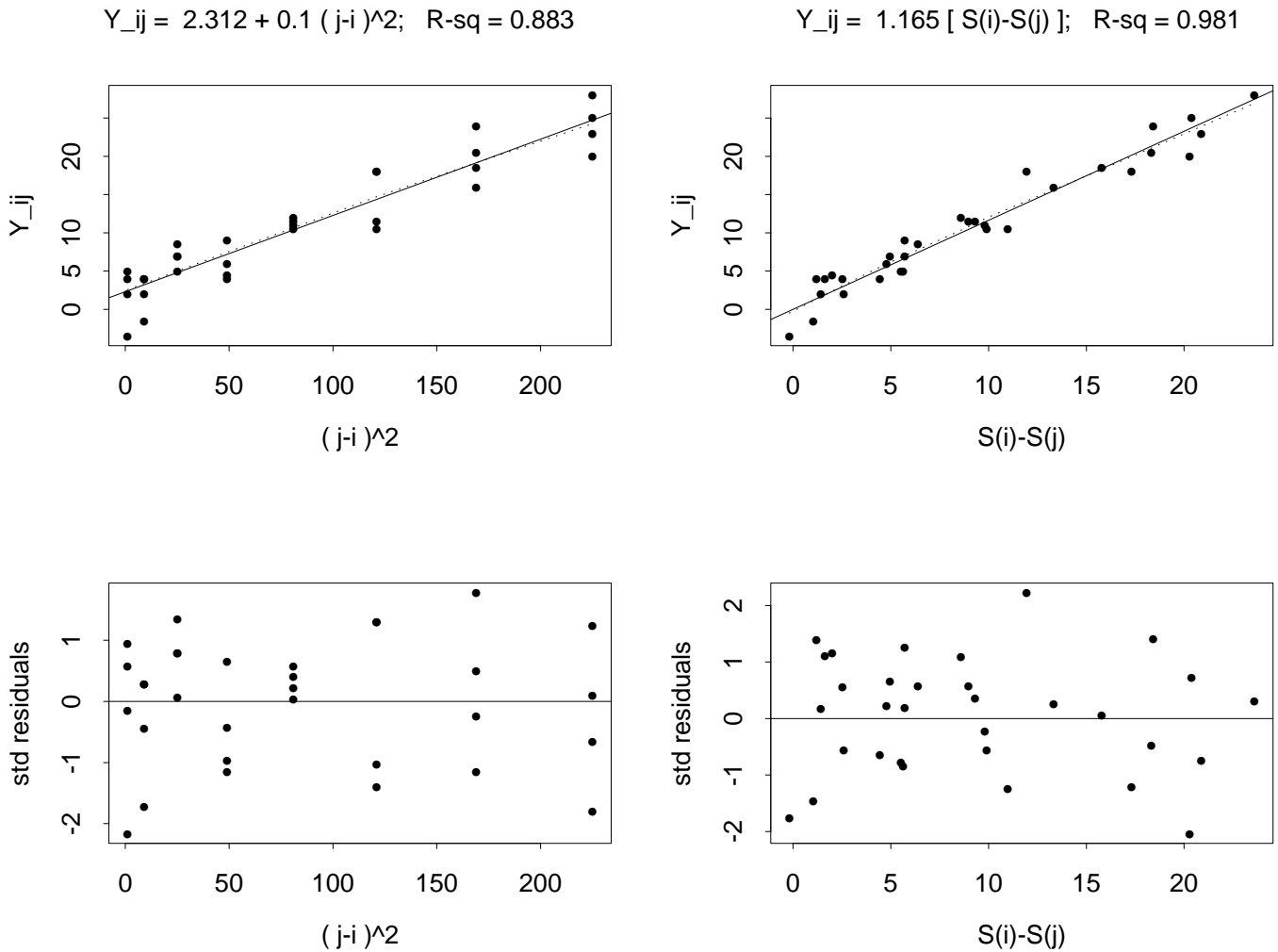


Figure 1: Regression of round 1 point spreads on differences in seeding and Sagarin rating

In the second column of plots in Figure 1, we replace squared seed difference as the predictor variable with the difference in Sagarin rating, a numerical measure of team strength which we denote by $S(k)$ for seed k . These ratings, which account for won-lost record in Division I games and strength of schedule, are published every Monday during the season in the newspaper *USA Today*. Table 1 gives the relevant differences from the collection of rankings published on the Monday immediately prior to the 1994 first round games. The ratings are designed to produce hypothetical point spreads when differenced, and our results bear this out: the data seem to require no transformation to achieve linearity, and the intercept in the full regression model is not significantly different from zero. Forcing the line through the origin, we obtain the fitted model $\hat{Y}_{ij} = 1.165[S(i) - S(j)]$, where $i < j$. The fitted slope coefficient suggests that the rating difference tends to slightly underestimate the point spread in matches between opponents of widely differing strengths. The improved R^2 value of .981 confirms the visual impression from the figure that Sagarin rating difference is superior to seed difference as a predictor of point spread. We remark that while the best 39 of the 301 Division I teams (as measured by Sagarin rating) were included in the 1994 tournament, the four #16 seeds had Sagarin rankings 166, 173, 196, and 216, calling into question the assumption of Schwertman et al. (1991) that the 64 tournament teams may be safely thought of as the best 64 teams in the country.

3 Application to the 1994 NCAA Tournament

We begin by comparing the results of several approaches suggested by Figure 1 using the 1994 Southeast regional tournament data, since differences amongst the approaches are most apparent using data from this region. Table 2 gives the estimated probability of emerging as the regional champion for each of the 16 teams. The first method listed is the one recommended by Schwertman

et al. (1991). This method assigns nearly 50% of the mass to the #1 seed, while giving only 5% to the entire lower half of the bracket (seeds 9–16). The next column provides results obtained by using the regression of point spread on squared seed difference to obtain the win probabilities. This method gives slightly more mass to the upper seeds other than #1, but even less mass to the lower division teams (total probability less than 2%). Like the Schwertman method, it uses only seed information to determine win probabilities, so the results in these columns would apply to any of the four regional tournaments. The seed regression results *are* specific to this *year*, however, since 1994 spread data were used to fit the model.

| seed | team | Schwertman method | Seed Regression | Sagarin Differences | Sagarin Regression | Sagarin Regr w/R1 Spreads |
|------|-----------------|----------------------|--------------------|------------------------|-----------------------|------------------------------|
| 1 | Purdue | 0.459 | 0.326 | 0.316 | 0.343 | 0.349 |
| 2 | Duke | 0.188 | 0.235 | 0.151 | 0.148 | 0.150 |
| 3 | Kentucky | 0.110 | 0.155 | 0.245 | 0.260 | 0.255 |
| 4 | Kansas | 0.068 | 0.110 | 0.111 | 0.108 | 0.103 |
| 5 | Wake Forest | 0.047 | 0.064 | 0.032 | 0.024 | 0.027 |
| 6 | Marquette | 0.036 | 0.045 | 0.026 | 0.019 | 0.021 |
| 7 | Michigan State | 0.026 | 0.028 | 0.033 | 0.026 | 0.025 |
| 8 | Providence | 0.015 | 0.018 | 0.067 | 0.061 | 0.058 |
| 9 | Alabama | 0.011 | 0.008 | 0.005 | 0.003 | 0.003 |
| 10 | Seton Hall | 0.011 | 0.005 | 0.010 | 0.006 | 0.007 |
| 11 | SW Louisiana | 0.009 | 0.003 | 0.002 | 0.001 | 0.001 |
| 12 | Charleston | 0.006 | 0.001 | 0.002 | 0.001 | 0.000 |
| 13 | TN–Chattanooga | 0.005 | 0.001 | 0.001 | 0.000 | 0.000 |
| 14 | Tennessee State | 0.004 | 0.000 | 0.000 | 0.000 | 0.000 |
| 15 | Texas Southern | 0.003 | 0.000 | 0.000 | 0.000 | 0.000 |
| 16 | Central Florida | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 |
| sum | | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |

Table 2: Comparison of $\Pr(\text{seed wins Southeast region})$ across models

Listed next are results obtained by simply taking the unadjusted difference of the Sagarin ratings as the hypothetical point spread between the two teams. Note that the probability of a triumph by the #1 seed has dropped again, and total support for the lower division teams remains a mere 2%. Note also that these probabilities are specific to *both* year and region, since teams with

identical seeds in different regions need not have identical Sagarin ratings. Indeed, the probabilities are no longer strictly decreasing from seed 1 down to seed 16, due to ordering conflicts between the seedings and the Sagarin ratings (e.g., #2 Duke rated 89.90, #3 Kentucky rated 91.59). The next column in the table adjusts the Sagarin differences using the regression model obtained in the previous section before converting them to win probabilities via equation (1). This adjustment amounts to giving a boost to the two most highly rated teams in the region (Purdue and Kentucky) at the expense of the others; notice that the total probability allocated to the lower half of the bracket is now barely 1%. Finally, the last column replaces the imputed first round point spreads used in the previous method with the actual point spreads. This results in subtle changes to only 16 entries in the P matrix, but since they are the entries corresponding to the first games played, the effect on the regional championship probabilities is apparent, occasionally visible in the second decimal place.

Table 3 applies this final method (using the Sagarin regression model plus actual first round spreads) to data from each of the four 1994 regional tournaments. There are several reversals of the seeding order, the most interesting of which is the prediction of #2 Arizona, not #1 Missouri, as the team most likely to win the West regional. (To the model's credit, this was indeed the outcome in this region.) Support for seeds in the lower half of each region is again quite low. The largest single probability in the table is .379 (Arizona), substantially lower than that given to any #1 seed by the Schwertman model, and perhaps indicative of the increased "parity" in college basketball often mentioned by sportswriters and coaches during the 1994 season. Again, the tournament results bear this out somewhat, as the actual regional champions were seeded 2 (Arizona), 1 (Arkansas), 3 (Florida), and 2 (Duke).

Schwertman et al. (1991) explore model fit more formally by comparing observed and expected numbers of teams with a given seeding to become regional champions over the first six years of

| seed | West | Midwest | East | Southeast |
|------|-------|---------|-------|-----------|
| 1 | 0.182 | 0.310 | 0.349 | 0.349 |
| 2 | 0.379 | 0.232 | 0.306 | 0.150 |
| 3 | 0.147 | 0.176 | 0.089 | 0.255 |
| 4 | 0.103 | 0.093 | 0.109 | 0.103 |
| 5 | 0.053 | 0.046 | 0.043 | 0.027 |
| 6 | 0.072 | 0.047 | 0.041 | 0.021 |
| 7 | 0.009 | 0.013 | 0.020 | 0.025 |
| 8 | 0.037 | 0.044 | 0.009 | 0.058 |
| 9 | 0.011 | 0.020 | 0.017 | 0.003 |
| 10 | 0.003 | 0.011 | 0.004 | 0.007 |
| 11 | 0.002 | 0.002 | 0.002 | 0.001 |
| 12 | 0.003 | 0.005 | 0.010 | 0.000 |
| 13 | 0.000 | 0.000 | 0.000 | 0.000 |
| 14 | 0.000 | 0.001 | 0.000 | 0.000 |
| 15 | 0.000 | 0.000 | 0.000 | 0.000 |
| 16 | 0.000 | 0.000 | 0.000 | 0.000 |
| sum | 1.000 | 1.000 | 1.000 | 1.000 |

Table 3: Comparison of $\Pr(\text{seed wins region})$ across regions

64-team NCAA tournament play (i.e., 24 regional champions). This method of model validation is unavailable to us, however, since the probabilities in Table 3 are specific to both team and year, not just seeding. Instead, we judge a method that produces the win probability matrix P using the scoring rule

$$I(P, Z) = \frac{1}{K} \sum_{k=1}^K \log[p_k Z_k + (1 - p_k)(1 - Z_k)] , \quad (2)$$

where $Z = \{Z_1, \dots, Z_K\}$, $Z_k = 1$ if the team with the more favorable seeding wins game k , and $Z_k = 0$ otherwise. Equation (2) is simply the average log-win probability predicted by the model for those teams that actually won, so that models are rewarded for assigning high probability to these teams. $I(P, Z)$ can also be thought of more formally as the average Shannon information in the likelihood (see e.g. Lindley, 1956).

Table 4 gives the information scores for each of the five methods compared in Table 2. The bottom line sums over all $K = 60$ games in Rounds 1-4 of the 1994 tournament. To provide a

| region | Schwertman method | Seed Regression | Sagarin Differences | Sagarin Regression | Sagarin Regr w/R1 Spreads |
|-----------|----------------------|--------------------|------------------------|-----------------------|------------------------------|
| West | -0.116 | -0.111 | -0.106 | -0.101 | -0.102 |
| Midwest | -0.134 | -0.147 | -0.134 | -0.134 | -0.127 |
| East | -0.154 | -0.148 | -0.149 | -0.152 | -0.145 |
| Southeast | -0.114 | -0.103 | -0.116 | -0.114 | -0.111 |
| All | -0.517 | -0.508 | -0.505 | -0.502 | -0.485 |

Table 4: Information scores for the five tournament probability estimation methods

reference point for the scores on this line, the completely naive method that assumes every game is an even toss-up (all $P(i, j) = 0.5$) would have a score of $\log(0.5) = -.693$. Notice the monotonic improvement in score as we move from left to right, with our final method (the one used in Table 3) emerging as noticably better than the rest. From the component scores within each region, we see that this final method scores higher than the Schwertman method in every region, though there is little difference in the Southeast region, where Kentucky's early exit and Duke's ultimate win contradicted the Sagarin ratings. But the methods based on these ratings perform quite well in the West region, where the ratings correctly predicted that #2 Arizona would beat #1 Missouri.

4 Conclusion

In this article we have developed a method for improved probability modeling of NCAA regional basketball tournaments. The method requires only elementary ideas in probability theory, statistical graphics, and linear regression analysis, and as such should provide an interesting and instructive exercise for students. Implementation for a given year requires only the Sagarin ratings for the appropriate 64 teams, perhaps the collection of first round point spreads (if refitting the regression model is desired), and the Fortran program for reducing the P matrix to the collection of regional championship probabilities (available from the author upon request via electronic mail).

We have argued on behalf of the use of (actual or imputed) point spreads in determining win

probabilities, on the grounds that true point spreads are superior to computer rankings, which are in turn superior to the crude summaries provided by tournament seedings. Our regression analyses in Section 2 and the information scores in Table 4 support this belief, as does other, more anecdotal evidence from the 1994 tournament. For example, Table 1 shows that in two first round games, the lower seed was actually favored by the bettors: East #9 (Boston College) 3.5 points over East #8 (Washington State), and Midwest #10 (Maryland) 1.5 points over Midwest #7 (St. Louis). The Sagarin ratings corrected the seeding error in the former case, but not in the latter (Maryland rating 83.43, St. Louis rating 84.47, a difference of 1.04). Perhaps the bettors were influenced in this case by their additional knowledge that St. Louis had lost 3 of their last 4 games prior to the tournament, or that the team's tallest player was injured, suggesting that they would have a hard time guarding Maryland's 6'10" center, Joe Smith. As it turned out, Smith had 29 points and 15 rebounds in the game, and Maryland won by 8 points.

As a final comment, we note that in some cases interest may focus not on the prediction of the Final Four, but on the prediction of the ultimate national champion. While our ideas could of course be extended to the case of a single 64 by 64 P matrix, the programming involved in reducing this matrix to the vector of championship probabilities would be almost unbearably tedious. As a simple alternative, we might assume that the Final Four teams are more or less evenly matched, and thus select as our national champion the team most likely to make it this far in the tournament (in our case, W#2 Arizona). It is worth pointing out, however, that this logic along with Table 3 suggests that no single team would be likely to have even a 10% chance at the tournament's outset of winning the required six consecutive games, so that any such prediction is almost certain to be incorrect.

Acknowledgements

The author thanks Prof. Neil Schwertman for supplying the Fortran code to convert the P matrix into the regional championship probabilities, Prof. Jim Albert for supplying the 1994 pre-tournament Sagarin ratings, Prof. Hal Stern for invaluable advice and for suggesting the scoring rule used in Table 4, and Profs. Lance Waller and Alan Gelfand for helpful discussions.

References

- ALBERT, J.H. (1993), “Teaching Bayesian Statistics Using Sampling Methods and MINITAB,” *The American Statistician*, **47**, 182–191.
- BARRY, D. and HARTIGAN, J.A. (1993), “Choice Models for Predicting Divisional Winners in Major League Baseball,” *Journal of the American Statistical Association*, **88**, 766–774.
- BENNETT, J. (1993), “Did Shoeless Joe Jackson Throw the 1919 World Series?” *The American Statistician*, **47**, 241–250.
- CASELLA, G. and BERGER, R.L. (1994), “Estimation with Selected Binomial Information or Do You Really Believe that Dave Winfield is Batting .471?” *Journal of the American Statistical Association*, **89**, 1080–1090.
- HARVILLE, D. (1980), “Predictions for National Football League Games via Linear-Model Methodology,” *Journal of the American Statistical Association*, **75**, 516–524.
- LARKEY, P.D., SMITH, R.A. and KADANE, J.B. (1989), “It’s Okay to Believe in the ‘Hot Hand’,” *Chance*, **2**(4), 22–30.
- LINDLEY, D.V. (1956), “On the Measure of Information Provided by an Experiment,” *Annals of Statistics*, **27**, 986–1005.

- SCHWERTMAN, N.C., MCCREADY, T.A. and HOWARD, L. (1991), “Probability Models for the NCAA Regional Basketball Tournaments,” *The American Statistician*, **45**, 35–38.
- SCHWERTMAN, N.C., SCHENK, K.L. and HOLBROOK, B.C. (1993), “More Probability Models for the NCAA Regional Basketball Tournaments,” technical report, Department of Mathematics and Statistics, California State University – Chico.
- STERN, H. (1991), “On the Probability of Winning a Football Game,” *The American Statistician*, **45**, 179–183.
- STERN, H. (1992), “Who’s Number One? – Rating Football Teams,” In *Proc. Section on Sports Statistics*, **1**, Alexandria, VA: American Statistical Association, pp. 1–6.
- TVERSKY, A. and GILOVICH, T. (1989), “The Cold Facts about the ‘Hot Hand’ in Basketball,” *Chance*, **2(1)**, 16–21.