A Logistic Regression/Markov Chain Model for NCAA Basketball

Paul Kvam, Joel S. Sokol

School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332-0205

Received 16 May 2005; accepted 14 May 2006
DOI 10.1002/nav.20170
Published online 14 July 2006 in Wiley InterScience (www.interscience.wiley.com).

Abstract: Each year, more than $3 billion is wagered on the NCAA Division 1 men's basketball tournament. Most of that money is wagered in pools where the object is to correctly predict winners of each game, with emphasis on the last four teams remaining (the Final Four). In this paper, we present a combined logistic regression/Markov chain model for predicting the outcome of NCAA tournament games given only basic input data. Over the past 6 years, our model has been significantly more successful than the other common methods such as tournament seedings, the AP and ESPN/USA Today polls, the RPI, and the Sagarin and Massey ratings. © 2006 Wiley Periodicals, Inc. Naval Research Logistics 53: 788–803, 2006.

1. INTRODUCTION

More money is bet on the National Collegiate Athletic Association (NCAA) Division I men's basketball tournament than on any other sporting event in the United States. The FBI estimates that every year, more than $3 billion is wagered (legally and illegally) on the tournament's outcome [1]. With so much money on the line, a model that predicts outcomes more effectively than standard ranking and rating systems can be useful, especially if it requires only basic input data. In this paper, we present such a model.

Before describing the model, we provide a short introduction to the NCAA tournament for readers to whom it might not be familiar. At the conclusion of each college basketball season, the NCAA holds a 64-team tournament. The participants are the champions of the 31 conferences in Division I, plus the best remaining teams (as judged by the tournament selection committee). In addition to choosing the teams, the selection committee also seeds them into four regions, each with seeds 1–16. The four teams judged best by the committee are given the No. 1 seeds in each region, the next four are given the No. 2 seeds, etc. Within each region, the 16 teams play a four-round single-elimination tournament with matchups determined by seed (1 vs. 16, 2 vs. 15, etc.); the winner of each region goes to the Final Four. The Final Four teams play a two-round single-elimination tournament to decide the national championship. Throughout all six rounds of the tournament, each game is played at a neutral site rather than on the home court of one team or the other.

In most NCAA tournament pools (the primary outlet for tournament wagers), participants predict the winner of each game. All predictions are made before the tournament starts, so it is possible that the predicted winner of a late-round game might not even be a participant, if that team lost in an earlier round.

Pool participants have several sources that can help them make their predictions. The most common such ranking systems are the Associated Press poll of sportswriters, the ESPN/USA Today poll of coaches, the Ratings Percentage Index (a combination of a team's winning percentage and that of the team's opponents), the Sagarin ratings published in USA Today [18], the Massey ratings [15], Las Vegas betting odds, and the tournament selection committee's seedings. Other sports ranking and rating systems have been developed (see Wilson [23] for an extensive bibliography), but to our knowledge none has been shown to dominate those listed above.

A separate question is, once a ranking or rating system has been selected, how is the information used in a pool setting? Kaplan and Garstka [13] describe a dynamic programming model that, given estimated probabilities of each team beating each other team head-to-head in a tournament game and given point values for each game in the pool, suggests a prediction strategy that can be used to maximize one's pool score. Breiter and Carlin [4] obtain similar (at though slower) results via a brute force algorithm. Of course, the quality of the dynamic programming and brute

Correspondence to: Joel Sokol, 765 Ferst Drive, NW, Groseclose 0205, Room 418, Atlanta, GA. (jsokol@isye.gatech.edu)

force solutions is dependent on having good probability estimates. Schwertman et al. [10,11] suggest methods for estimating win probabilities based on teams' seedings in the tournament. Carlin [7] suggests methods for estimating win probabilities based on the Sagarin ratings and Las Vegas point spreads; Breiter and Carlin [4] use those methods to illustrate their algorithm. Boulier and Stekler [2] fit a probit model to estimate win probabilities based on seedings in order to maximize the number of games predicted correctly. Caudill [8] uses a maximum score estimator model that is also based on seedings and also tries to maximize the number of correct predictions. Caudill and Godwin [9] use a heterogeneously skewed logit model for the same purpose. Kaplan and Garstka [13] propose methods for estimating win probabilities from scoring rates, Sagarin ratings, and Las Vegas point spreads.

Metrick [17] and Clair and Letscher [10] discuss a third relevant question: should one's prediction strategy change based on the number and relative skill of other competing predictors? They observe that sometimes differentiating one's predictions from the competition yields a higher change of having the best predictions.

In this paper, we focus on the first question—how to accurately rank (and/or rate) teams using only basic input data. We present a new model for ranking college basketball teams and estimating win probabilities. Our model uses a logistic regression to populate transition probabilities of a Markov chain. We describe the underlying Markov chain model in Section 2, and in Section 3 we describe the logistic regression model. Section 4 presents our computational results, and in Section 5 we make a conjecture as to why our model is significantly more successful than both standard ranking systems and the NCAA tournament selection committee's seeds when used alone and in the dynamic programming framework. Section 6 summarizes the paper.

2. A MARKOV CHAIN MODEL

In this section, we describe a Markov chain model for ranking teams. We begin with a model used to construct NCAA football rankings by Callaghan, Porter, and Mucha [5,6]; Massey [16] also discusses a similar, but slightly less flexible, model. The underlying model is a Markov chain with one state for each team. The intuition is that state transitions are like the behavior of a hypothetical voter in one of the two major polls. The current state of the voter corresponds to the team that the voter now believes to be the best. At each time step, the voter reevaluates his judgement in the following way: given that he currently believes team i to be the best, he picks (at random) a game played by team i against some opponent j. With probability p, the voter moves to the state corresponding to the game's winner; with probability $(1 - p)$, the voter moves to the losing team's state.

Suppose team i has played a total of N_i games, with the kth game ($k \leq N_i$) being played against opponent O_k. Let I_{ik} be an indicator equal to 1 if team i won its kth game and 0 if team i lost its kth game. Then the transition probabilities t_{ij} from state i in the Markov chain are defined as

$$t_{ij} = \frac{1}{N_i} \sum_{k:0_{ik}=j} [I_{ik}(1 - p) + (1 - I_{ik})p], \qquad \text{for all } j \neq i, \tag{1a}$$

$$t_{ii} = \frac{1}{N_i} \sum_{k=1}^{N_i} [I_{ik}\, p + (1 - I_{ik})(1 - p)]. \tag{1b}$$

If we let W_i and L_i be the number of games that team i has won and lost and w_{ij} and l_{ij} be the number of games that team i has won and lost against team j specifically, then these transition probabilities can be rewritten in a more intuitive form:

$$t_{ij} = \frac{1}{N_i} [w_{ij}(1 - p) + l_{ij}\, p], \qquad \text{for all } j \neq i, \tag{2a}$$

$$t_{ii} = \frac{1}{N_i} [W_i\, p + L_i(1 - p)]. \tag{2b}$$

As Eqs. (1) and (2) imply, state transitions can be defined as the toss of a fair N_i-sided die to select a game, followed by the toss of a weighted coin to determine whether the next state will correspond to the selected game's winner (with probability p) or loser (with probability $1 - p$).

Given the state transition probabilities $T = [t_{ij}]$ defined in (2a) and (2b), Callaghan, Porter, and Mucha use the standard equations $\pi T = \pi$, $\Sigma_i\, \pi_i = 1$ to calculate the steady-state probabilities of each team's node. The teams are ranked in order of their steady-state probability—the team with the highest steady-state probability is ranked first, etc.

A nice characteristic of Callaghan, Porter, and Mucha's Markov chain model is that it can be implemented simply, without much data. Specifically, daily on-line scoreboards such as [24] provide all the necessary data for the model; no additional team or individual statistics are required. When extending their model to college basketball, one of our goals was to preserve this basic simplicity. Therefore, our model also requires no more data than daily scoreboards provide.

2.1. Alternative Transition Probabilities

The transition parameter p can be interpreted in a very intuitive way: the value of p is the model's answer to the question "Given that team A beat team B, what is the

probability that A is a better team than B?" Daily scoreboards give additional useful information that can refine these probability estimates. It is well known in many of the major team sports, including baseball, basketball, soccer, football, and ice hockey, that teams playing at home have an advantage. Another factor that is often considered when evaluating teams is "margin of victory," defined as the difference between the winning and losing teams' scores. A team that wins its games by wide margins is generally thought to be better than a team that wins its games by narrow margins.

In the context of this model, we would like to find transition probabilities that answer the question, "Given that team A beat team B by x points at home (or on the road), what is the probability that A is a better team than B?"

Let $x(g)$ be the difference between the home team's score and the visiting (road) team's score in game g. We define r_x^H to be the probability that a team that outscores its opponent by x points at home is better than its opponent and $r_x^R = 1 - r_x^H$ to be the probability that a team that is outscored on the road by x points is better than its opponent. (Note that x can be negative to indicate that the home team lost the game.) If we denote each game by an ordered pair (i,j) of teams with the visiting team listed first, then we can write the state transition probabilities for each team i as

$$t_{ij} = \frac{1}{N_i}\left[\sum_{g=(i,j)} (1 - r^R_{x(g)}) + \sum_{g=(j,i)} (1 - r^H_{x(g)})\right], \quad \text{for all } j \neq i, \tag{3a}$$

$$t_{ii} = \frac{1}{N_i}\left[\sum_j \sum_{g=(i,j)} r^R_{x(g)} + \sum_j \sum_{g=(j,i)} r^H_{x(g)}\right]. \tag{3b}$$

Wins, losses, locations, and margins of victory are easy to observe; the difficulty with using this model is in estimating values of r_x^H and r_x^R for each x. In Section 3, we present a logistic regression model that exploits the basketball schedule's structure to answer this question.

2.2. Relation to Standard Methods

Standard methods that are used to evaluate college basketball teams (RPI, Sagarin, Massey, etc.) take into account a team's record of winning games and its strength of schedule (i.e., the quality of the teams it played against when compiling that record). In fact, the pre-2005 RPI formula considered these factors explicitly: it took the weighted average of a team's winning percentage and its opponents' winning percentage. In this section, we show how the Markov chain steady-state probabilities can be viewed as a combination of these same two factors.

The steady-state probability π_i of being in the state of team i can be expressed as the product of two terms, the expected time to leave a state and the expected number of entries to that state, divided by an appropriate time constant. The expected time T_i to leave state i satisfies the equation $T_i = 1 + t_{ii}T_i$, so

$$T_i = \frac{1}{1 - \frac{1}{N_i}\left[\sum_j \sum_{g=(i,j)} r^R_{x(g)} + \sum_j \sum_{g=(j,i)} r^H_{x(g)}\right]} \tag{4}$$

Similarly, the expected number of entries E_i to state i is

$$E_i = \sum_j \pi_j t_{ji} = \sum_j \pi_j \frac{1}{N_j}\left[\sum_{g=(j,i)} (1 - r^R_{x(g)}) + \sum_{g=(i,j)} (1 - r^H_{x(g)})\right]. \tag{5}$$

Note that in (4), T_i is a function only of team i's performance in the games it played. In (5), E_i is a function of team i's performance against each team j, weighted by π_j, which is our measure of team j's strength. Therefore, we can see that our method is fundamentally similar to the RPI and to Sagarin and Massey's methods (as well as most others) in that it combines a team's performance with the strength of its opponents. The team's performance dictates how long the system remains in the team's state each time it enters, and the team's strength-of-schedule (and its performance against that schedule) dictates how often the system enters the team's state.

3. A LOGISTIC REGRESSION MODEL FOR CALCULATING TRANSITION PROBABILITIES

In this section, we describe a method for estimating the values of r_x^H, the probability that a team with a margin of victory of x points at home is better than its opponent. (Note that we need only determine r_x^H since $r_x^R = 1 - r_x^H$.) Estimating r_x^H is difficult because, while the input (margin of victory x) is easily observable, the response—whether one team is better than another—is hard to determine. (In fact, if we knew that information *a priori* or were able to directly observe it, there would be no need for the predictive model presented in this paper.)

To estimate r_x^H, we exploit the structure of NCAA basketball schedules. Almost every one of the approximately 330 Division I teams is a member of a basketball conference. Conferences each play a home-and-home round robin schedule in which members i and j of a conference play each other twice each year, once on i's home court and once on j's home court. Smaller conferences play full home-and-

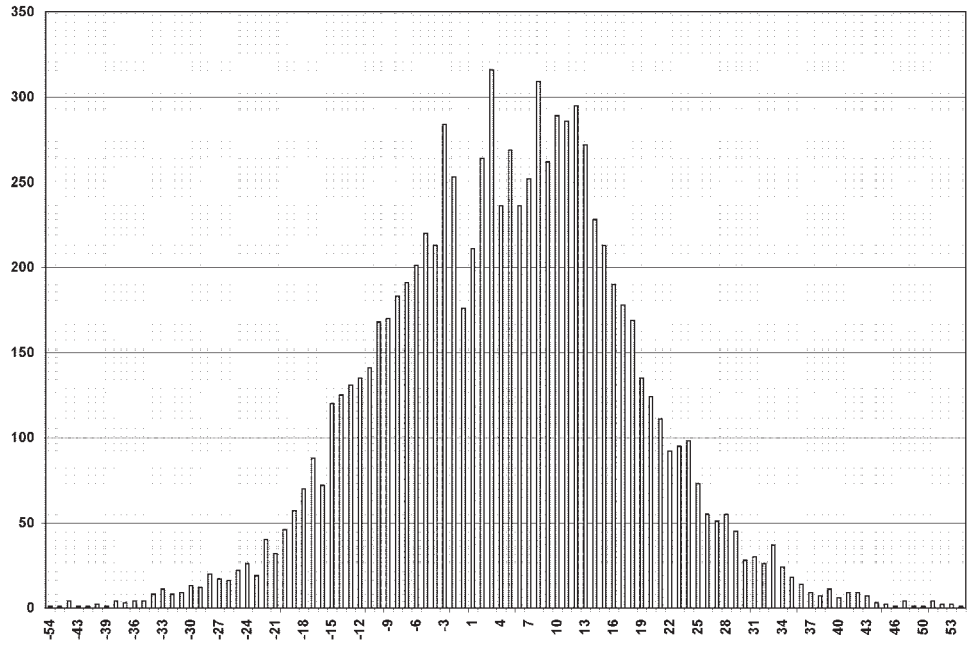

Figure 1. Number of home-and-home games by home team victory margin.

home round robin schedules, where each pair of teams in the conference plays twice. However, some conferences are too large; scheduling restrictions make it impossible for them to play enough games to fulfill the full home-and-home requirements. These conferences play a partial home-and-home round robin, in which most teams play each other twice while a few pairs of teams play each other only once.

We focus on pairs of teams (from both smaller and larger conferences) that play each other twice per season as part of either a full or a partial home-and-home round robin schedule. Our method consists of two steps:

1. Using home-and-home conference data, estimate an answer to the following question: "Given that team A had a margin of victory of x points at home against team B, what is the probability that team A beat team B in their other game, on the road?"
2. Given these road-win probabilities, deduce r_x^H, the probability that the home team is the better team, i.e., "Given that team A had a margin of victory of x points at home against team B, what is the probability r_x^H that team A is better than team B, i.e., that team A would beat team B on a neutral court?"

We used four years of NCAA data (the 1999–2000 through 2002–2003 seasons) to estimate these probabilities. In those four seasons, we found all matched pairs of teams that had played once on each of their home courts. For each of those games we recorded the home team, the visiting team, and the point differential at the end of regulation time.[1] We note that on very rare occasions, conference games might be played on a neutral court; these data were not available, and we do not believe their inclusion would significantly impact our results.

Figure 1 displays the number of games in which the home team won by various margins. As one might expect, the frequency decreases as the absolute value of the point spread increases; there are very few games decided by 50 or 60 points.

Figure 2 shows the observed win probabilities S_x^H, where S_x^H answers the following question: "Given that team A beat team B by x points on A's home court, what is the probability that A beat B on B's home court?" (Note that this is not quite the question we would like to answer; in Section 3.1 we discuss how to deduce r_x^H from S_x^H.)

For each margin of victory (or loss) by a home team i against an opponent j, Figure 2 shows the fraction of times that team i beat the same opponent j on j's home court. For example, 50% of teams that lost by 36 points on their home court beat that same opponent on the opponent's home court. Although this seems improbable, Figure 1 shows the reason: the sample size is only two games. Similar improbable results caused by small sample sizes can be found at the

[1] We treat overtime games as having a zero-point differential; because overtime periods are relatively short, they are played with different strategies that might give a different distribution of point spreads.

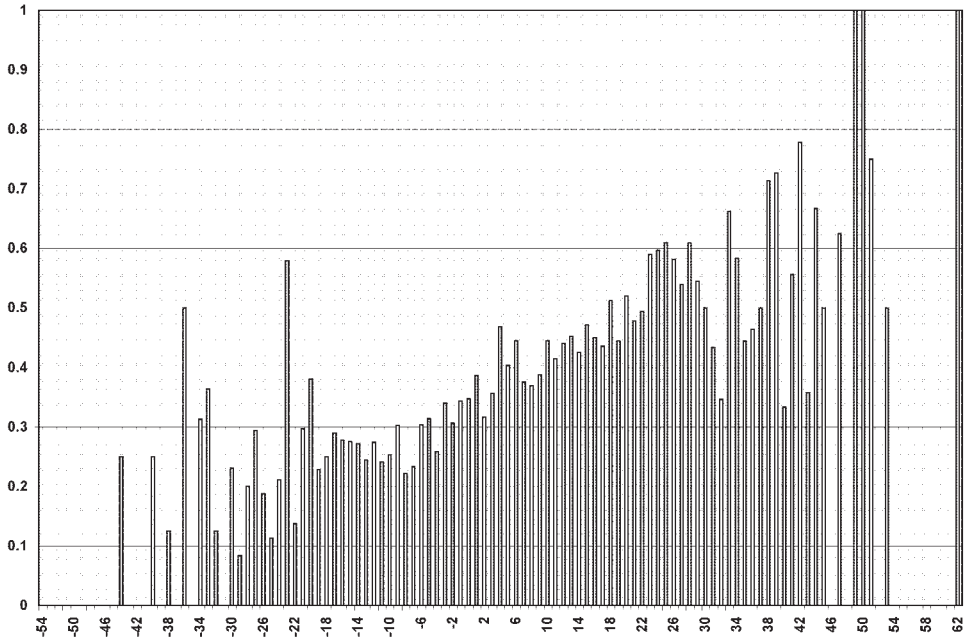

Figure 2. Observed probability of a home team winning its road game against the same opponent, given margin of victory in the home game.

other extreme. For example, 0% of teams that won by 54 points at home also won the road matchup; in this case, there was only one observation of a 54-point win.

To obtain a better, smoother estimate of win probability, we use a logistic regression model to find a good fit. The logistic regression helps linearize the nonlinear function by estimating parameters a and b to fit $(lnS_x^H(1-S_x^H)) = ax + b$. Rearranging terms yields an expression for the probability that a team with an x-point margin at home will win the road matchup: $S_x^H = e^{(ax+b)}/(1 + e^{(ax+b)})$.

The best-fit parameters using the matched-pair games from the 1999–2000 through 2002–2003 seasons are $(a,b) = (0.0292,-0.6228)$ with standard errors (0.0017,0.0231). Figure 3 shows the logistic regression estimate of S_x^H superimposed on the observed probability chart.

3.1. Deducing Neutral-Court Probabilities

The logistic regression model presented in Section 3 estimates S_x^H, the probability that team A will beat team B on B's court given that A beat B by x points on A's court. However, in order to populate the transition matrix for our Markov chain model, we need an estimate of r_x^H, the probability that team A will beat team B on a neutral site given that A beat B by x points on A's court. In this section, we describe how we deduce r_x^H from S_x^H.

The key to finding r_x^H is to consider the case in which the game on B's home court is an even matchup ($S_x^H = 0.5$). We make one significant assumption, that the effect of home-court advantage is additive. In other words, we assume that playing at home increases a team's expected point spread by $h > 0$; in such a model (also implicitly used by [18] and others), h is called the home-court advantage.

Given that home teams have some expected advantage h, we also assume that a game between A and B on B's home court is an even matchup when the expected point spread between the two teams is zero.[2] If the expected point spread on B's home court is zero, then the home-court advantage h must exactly cancel A's inherent advantage over B; the two have equal magnitude. Therefore, we expect that the game between A and B on A's home court would be decided by $2h$, since A would have both its inherent advantage and the home-court advantage.

In the case of a neutral-court game, a team that wins by x points at home would be expected to win by $x - h$ at the neutral site (due to losing their home-court advantage). Since S_x^H denotes the probability of winning when the expected point spread is $x - 2h$, we can deduce that the

[2] One can imagine scenarios where this is not true, e.g., where one team has a 90% chance of winning by a single point, while the other team has a 10% chance of winning by nine points; however, the distribution of observed point spreads shown in Figure 1 suggests that our simple model is a reasonable approximation of reality.

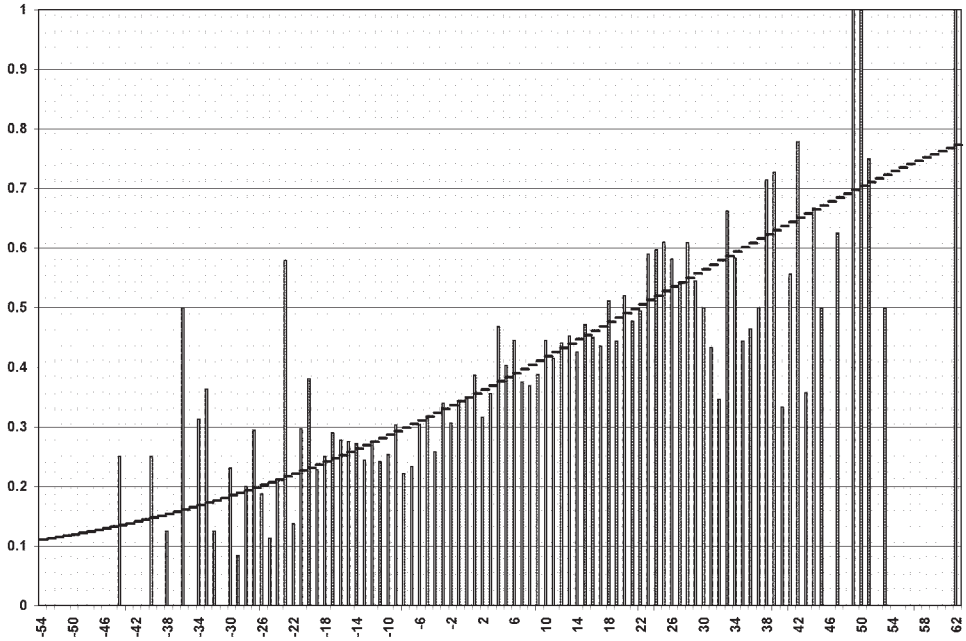

Figure 3. Observed values and logistic regression estimates for S_x^H.

probability of winning when the expected point spread is $x - h$ must be $r_x^H = S_{x+h}^H$.

3.2. Team vs. Team Win Probabilities

The probabilities r_x^H can be used to seed the Markov chain transition matrix, as described in Section 2. The resulting steady-state probabilities π give a natural ranking of teams: the team with the highest steady-state probability is highest-ranked, the team with the second-highest steady-state probability is ranked second, etc. These rankings can be used to predict tournament outcomes, under the assumption that it is best to always pick the higher-ranked team to win a game.

As Breiter and Carlin [4] and Kaplan and Garstka [13] pointed out, picking the highest-ranked available team might not always be the best strategy. Their models require estimates of team-vs.-team win probabilities in order to find a pool strategy. Therefore, we would like to use our logistic regression/Markov chain model to determine estimates for these team-vs.-team win probabilities.

Carlin [7], Breiter and Carlin [4], and Kaplan and Garstka [13] use a simple method for determining team-vs.-team win probabilities. Given an estimated point difference x between the two teams (i.e., given that team i is expected to score x more points than team j in a head-to-head matchup) and a standard error σ of the difference in score, they estimate the probability of i beating j as $p_{ij} = \Phi(x/\sigma)$. This probability estimate can be used with any model that predicts a head-to-head scoring difference, such as Sagarin ratings [4,7,13], Poisson scoring-rate models [13], and Las Vegas betting lines [7,13]; therefore, all we require is a way to estimate scoring difference from our steady-state probabilities π.

Surprisingly, the scoring difference between two teams appears to be fairly well-estimated by a simple linear function of the difference in steady-state probabilities. Specifically, using 1999–2003 regular-season data we find a good estimate to be

$$x_{ij} = 9180(\pi_i - \pi_j). \tag{6}$$

Adding nonlinear (including logarithmic) factors does not improve the fit of the model; even simply allowing for different coefficients of π_i and π_j does not yield an improvement — the two coefficients are nearly identical (modulo their signs), and each is well within the standard error of the other. Overall, the simple one-parameter model in Eq. (6) has a standard error of 10.9 points; by comparison, Breiter and Carlin [4] use a standard error of 11 points when dealing with Las Vegas betting lines and Sagarin ratings. The standard error of the coefficient 9180 in our model is 71.5.

We also attempt to predict team-vs.-team win probabilities directly from steady-state probabilities (i.e., without using scoring difference as an intermediate step). A logistic regression model is appropriate, since the outcomes are binary (wins and losses). Figure 4 shows the relative frequency of steady-state-probability differences between teams that played each other.

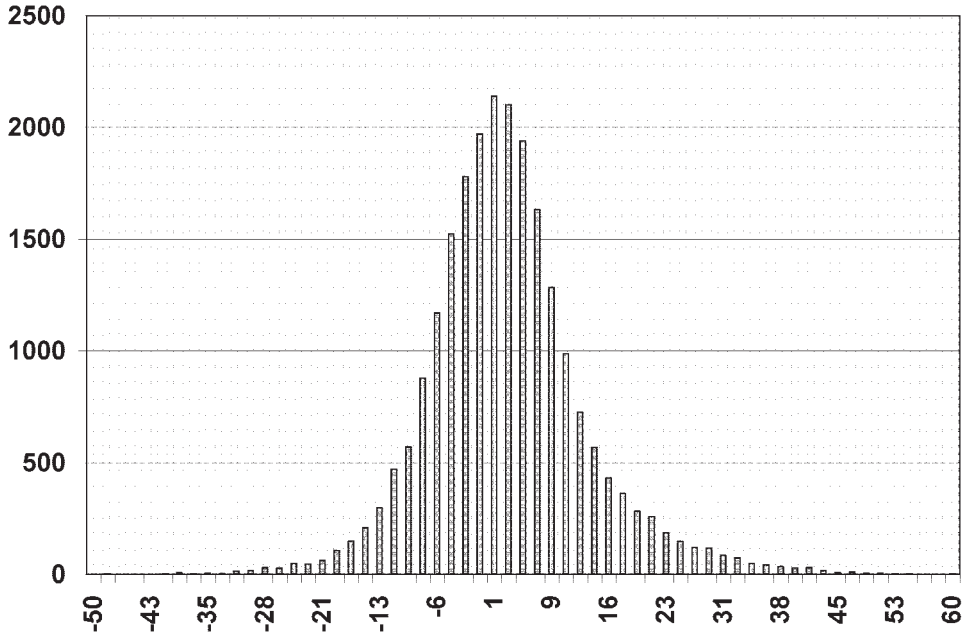

Figure 4. Number of games by steady-state probability differences $\times 10^{-4}$.

Figure 5 shows the observed probability that the home team wins a game with a certain steady-state probability difference.

The best-fit logistic regression model (obtained using Minitab) is

$$\hat{p}_{ij} = 1 - \frac{e^{-1834.72(\pi_i - \pi_j) - 0.6716}}{1 + e^{-1834.72(\pi_i - \pi_j) - 0.6716}}. \tag{7}$$

However, these data implicitly include a home-court advantage. The constant term 0.6716 in the exponential can be thought of as the home-court effect; on a neutral court, the probabilities translate to

$$p_{ij} = 1 - \frac{e^{-1834.72(\pi_i - \pi_j)}}{1 + e^{-1834.72(\pi_i - \pi_j)}}. \tag{8}$$

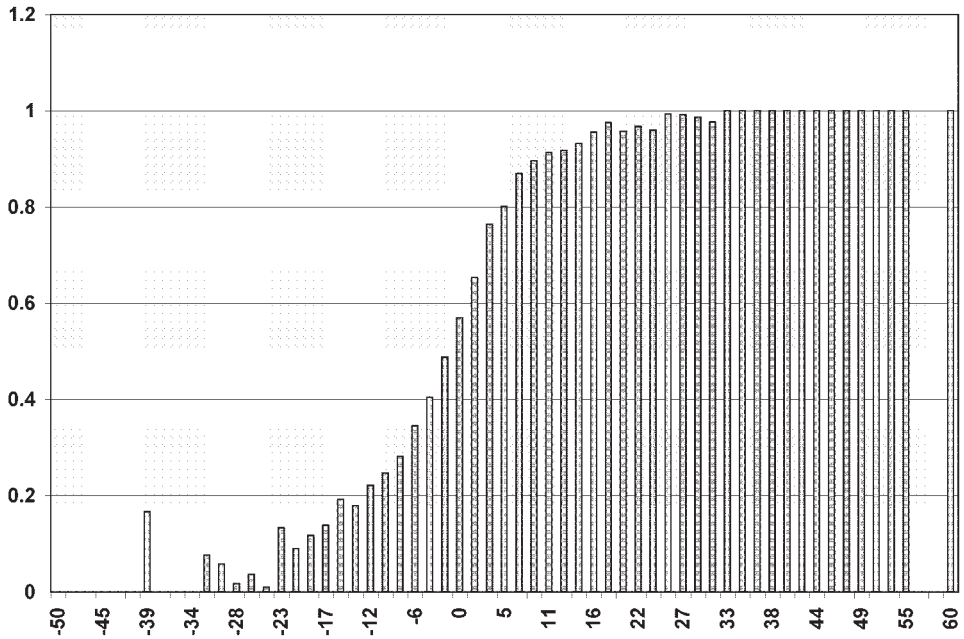

Figure 5. Observed win probability by steady state difference $\times 10^{-4}$.

4. COMPUTATIONAL RESULTS

To test the models developed in Section 3, we analyze their predictions of NCAA Tournament games. The NCAA Tournament schedules all of its games on neutral courts, thus eliminating home-court advantage. Consequently, our transition probabilities, each of which estimates the probability of one team being better than another, are valid for attempting to predict the winners of neutral-court NCAA Tournament games.

We test our model in two different ways: as a stand-alone predictor (where the better team is always predicted to win) and as the source of probability estimates for Kaplan and Garstka's [13] dynamic programming model. By comparing our model's success with the success of other ranking and rating systems, we hope to determine both how good a ranking system it is compared to other common methods and its compatibility (again, compared to other methods) with the dynamic programming framework.

4.1. Predictive Models Used

We compare our logistic regression/Markov chain (LRMC) model to the most commonly-used NCAA basketball ranking, rating, and measurement systems. These include expert polls (the Associated Press poll of sportswriters (AP) and the ESPN/USA Today poll of coaches (ESPN)), mathematical models (the Ratings Percentage Index (RPI), the Sagarin ratings (Sagarin) [18], the Massey ratings (Massey) [14], and Massey's win probabilities (MasseyProb) [14]), predictions from Las Vegas oddsmakers (Sheridan's odds against winning (Sheridan), game-by-game Vegas point spreads from Statfox (Vegas), and Kaplan and Garstka's [13] method (KG) for deriving team ratings from first-round point spreads and over-unders), and the actual NCAA tournament seeds (Seed) assigned by the tournament selection committee.

For each of the models except for game-by-game Vegas point spreads, we use the final pre-tournament ranking of teams. We obtained ranking data from the pre-tournament polls,[3] RPI, seedings, Sheridan's odds against winning, and Sagarin's overall ratings from USA Today [22]. Massey's ratings and win probabilities were taken from his web pages [15].

First-round point spreads and over-unders were compiled from [12] and converted to ratings using the method of Kaplan and Garstka [13]. Given the Las Vegas line that team i is a f_{ij}-point favorite over team j and the over-under (the expected number of points scored in the game) is g_{ij}, Kaplan and Garstka [13] deduce an implicit rating λ_i and λ_j for each of the two teams: $(\lambda_i + \lambda_j) = g_{ij}$ and $(\lambda_i - \lambda_j) = f_{ij}$, so $\lambda_i = (f_{ij} + g_{ij})/2$ and $\lambda_j = (g_{ij} - f_{ij})/2$.

Game-by-game Vegas point spreads were compiled from Statfox [21] game logs. For consistency with Vegas pre-tournament rankings, we tried to use Sheridan's better known predictions for game-by-game point spreads. However, because USA Today does not publish on weekends (and because Sheridan's spreads are not archived) the spreads on many games were unavailable. In fact, Statfox was the only readily available source of archived Vegas point spreads for each NCAA tournament game from 1999–2000 through 2004–2005. Although different Las Vegas oddsmakers will publish slightly different point spreads, the team favored to win rarely differs. Moreover, we found that Statfox made slightly more correct predictions than Sheridan in a sample of over 100 tournament games where data from both were available.

For the 2004–2005 season, the NCAA changed the mathematical formula for RPI; the 2004–2005 predictions using the new formula were almost exactly the same as using the old formula, so we report only the old RPI formula here. We obtained these RPI data from [11].

For the LRMC model, we used all of the game data (home team, visiting team, margin of victory) from the beginning of the season until just before the start of the tournament; we obtained these data, as well as tournament results, on line from Yahoo! daily scoreboards [24]. We note that neutral-site non-tournament games were unknown in our data set; the team listed as "home" on the scoreboard was considered the home team in our data.

In all cases, we naively deduced rankings from ratings-based systems (Sagarin, Massey, RPI, KG, Sheridan, and LRMC) simply by assigning higher-rated teams a higher ranking. In cases where two teams had the same rating, they were assigned identical (tied) rankings.

4.2. Best-Team-Wins Results

Based on the pre-tournament rankings from each source, we evaluated each based on its ability to predict outcomes of tournament games. We first tested naive "best team wins" predictions and counted (1) the number of games for which the method's pre-tournament rankings predicted the correct winner and (2) the number of games in which the higher-ranked team won. These two counts were different because of the possibility of multiple upsets. For example, if two second-round opponents were both upset winners in the first round, then metric (1) would give a score of zero but metric (2) could give a score of 1 if the second-round winner was the higher-ranked of the two upset winners. We report the results of both metrics because they both give information

[3] Only the top 30–40 teams generally get votes in the polls, so all unranked teams were given equal pre-tournament poll rankings. It is rare that two such teams lasted long enough in the tournament to face each other; those instances were considered tossups when evaluating the polls' predictions.

Table 1. Performance of models on two metrics of prediction quality, 1999–2000 through 2004–2005 seasons (378 total games).

	Polls			Mathematical models			Las Vegas predictions			
	AP	ESPN	Seed	RPI	Massey	Sagarin	KG	Sheridan	Vegas	LRMC
Games won by predicted winner	236	235½	235¼	229	242	229	231½	244½	n/a	248
Games won by higher ranked team	266	266½	265	262	268	264	260	268½	273½	277

regarding the quality of the ranking system. Note that because Vegas point spreads are generated game-by-game, they do not give full pre-tournament predictions (metric (1)), but would be expected to have a predictive advantage under metric (2) because they have more games of input data.

Table 1 shows the performance of the various prediction methods according to the two metrics, based on their performance over six seasons: 1999–2000 through 2004–2005. The LRMC model was more successful at picking the winners of tournament games than any of the other rankings. Note that fractional results indicate the presence of ties in the ranking (for example, when two No. 1 seeds play each other each is considered ½ of the predicted seeding-method winner).

Table 2 shows the one-tailed significance test results when comparing the LRMC model with each of the others with regard to game-by-game predictions (row 2 of Table 1). We used a one-tailed version of McNemar's test (essentially a binomial test that takes into account only those games where the two methods' predictions differ). The tests indicate that LRMC is better than AP, ESPN, RPI, Seed, Sagarin, KG, and Sheridan at the 0.05 level of significance or better. Only Massey (0.13) and Vegas (0.31) had significance levels poorer than 0.05. Note that the difference between the number of games predicted correctly by LRMC and by the other methods is, in some cases, slightly different between Tables 1 and 2. The reason is that games predicted by one method or the other as "even" (i.e., a zero-point spread, equal team rankings, or equal team ratings) are given a value of ½ in Table 1 (i.e., ½ of a correct prediction) whereas those games are thrown out of the comparison in McNemar's test.

Table 2. Significance of LRMC's performance against other game-by-game prediction methods, 1999–2000 through 2004–2005 seasons (378 total games).

Prediction method (x)	LRMC correct, x not (No. of games)	x correct, LRMC not (No. of games)	One-tailed significance
AP	30	17	0.04
ESPN	29	16	0.04
Seed	32	19	0.05
RPI	37	22	0.03
Massey	30	21	0.13
Sagarin	24	11	0.02
KG	32	16	0.01
Sheridan	27	15	0.04
Vegas	20	16	0.31

It is not surprising that our method is not significantly better than game-by-game Las Vegas odds. Oddsmakers use additional information: player unavailability (due to injury, suspension, ineligibility, and illness), matchups (either player-vs.-player or team-vs.-team), motivation, recent level of play, and performance in earlier-round tournament games. Any method that effectively utilizes such information can, by definition, be at least as good as a similar method that does not.

In addition to counting statistics, we also tracked the rankings of the eventual Final Four teams. The six seasons' Final Four teams had very different characteristics. There were three "surprise" teams in 1999–2000, including two (Wisconsin and North Carolina) that did not get even one vote in the coaches' poll. On the other hand, in 2000–2001 and 2001–2002, three of the four teams were considered favorites to reach the Final Four.

Table 3 shows each method's rankings of each season's Final Four teams. In five of the six test seasons, the LRMC model had the Final Four teams collectively ranked higher than any of the other ranking systems. Of the 24 teams, 16 were ranked in the top five by the LRMC model, and 21 were ranked in the top 10. Collectively, the 24 Final Four teams had a LRMC total ranking of 152, much better than the total for Sagarin (198), Sheridan (203–224), Massey (217), AP (≥236), ESPN (≥242), and RPI (264). The Seeding total ranged from 192 to 264, spanning the range of Sagarin, Sheridan, Massey, AP, ESPN, and RPI, but still clearly worse than LRMC. The KG model finished last, with a total of 296 (though it did have one notable success, ranking North Carolina as No. 8 in 2000). The Vegas method is not applicable here, as it does not provide pre-tournament predictions.

Table 4 shows the rankings of the same Final Four teams, but within each team's respective region. (For example, a ranking of "1" in Table 4 indicates that the team was that method's highest-ranked team within its tournament region, but not necessarily the top-ranked team overall.) Again, the

Table 3. Final Four teams' rankings.

	1999–2000				2000–2001				2001–2002				2002–2003				2003–2004				2004–2005				
	Michigan State	Florida	Wisconsin	North Carolina	Duke	Michigan State	Arizona	Maryland	Kansas	Oklahoma	Maryland	Indiana	Texas	Kansas	Marquette	Syracuse	Oklahoma State	Duke	Connecticut	Georgia Tech	Illinois	Louisville	North Carolina	Michigan State	Totals
AP	2	13	37	*	1	3	5	11	2	3	4	26	5	6	9	13	4	6	7	14	1	4	2	15	≥236
ESPN	2	11	*	*	1	3	4	11	2	3	4	27	5	6	11	12	3	6	7	15	1	4	3	15	≥242
Seed[1]	1–4	17–20	29–32	29–32	1–4	1–4	5–8	9–12	1–4	5–8	1–4	17–20	1–4	5–8	9–12	9–12	5–8	1–4	5–8	9–12	1–4	13–16	1–4	17–20	192–264
RPI	13	18	32	41	1	3	8	22	1	5	3	20	4	6	10	9	6	1	5	16	2	11	5	22	264
Massey	4	17	26	31	1	3	6	14	3	5	4	26	6	5	13	7	6	1	9	10	1	4	2	10	217
Sagarin	4	10	25	31	1	3	4	10	3	4	5	21	5	4	14	12	5	1	7	8	1	7	2	11	198
KG	11	19	38	8	1	7	2	10	3	11	2	49	1	16	25	14	15	2	13	16	5	6	1	21	296
Sheridan	1–2	6–9	31	27	1	3	5–6	7	2–3	5–6	2–3	18–20	6–7	4–5	17	15–16	6–7	1–2	3–4	10–12	1–2	13–14	1–2	18	203–224
LRMC	3	5	19	26	1	5	4	3	3	5	2	10	7	1	19	10	5	1	2	4	2	6	1	8	152

*Team was unranked; both polls had 42 ranked teams just before the 1999–2000 NCAA Tournament, so these teams were ranked no higher than 43rd.
[1]Four teams (one from each tournament region) are assigned each seed. Therefore, the four No. 1 seeds are ranked 1–4, the four No. 2 seeds are ranked 5–8, etc., without specification.

Table 4. Final Four teams' rankings within their respective regions.

	1999–2000				2000–2001				2001–2002				2002–2003				2003–2004				2004–2005				
	Michigan State	Florida	Wisconsin	North Carolina	Duke	Michigan State	Arizona	Maryland	Kansas	Oklahoma	Maryland	Indiana	Texas	Kansas	Marquette	Syracuse	Oklahoma State	Duke	Connecticut	Georgia Tech	Illinois	Louisville	North Carolina	Michigan State	Totals
AP	1	3	9	10+	1	1	2	3	1	1	1	6	1	2	3	3	1	1	2	3	1	1	1	4	62
ESPN	1	3	11+	12+	1	1	1	3	1	1	1	6	1	2	3	3	1	1	2	4	1	1	1	4	≥66
Seed	1	5	8	8	1	1	2	3	1	2	1	5	1	2	3	3	2	1	2	3	1	4	1	5	66
RPI	3	4	9	10	1	1	3	5	1	2	1	5	1	2	3	3	2	1	1	4	1	3	2	6	74
Massey	1	4	7	9	1	1	2	3	1	2	1	7	1	2	3	2	2	1	2	3	1	1	1	2	60
Sagarin	1	4	7	7	1	1	1	2	1	2	1	5	1	2	3	4	2	1	2	3	1	3	1	4	60
KG	3–4	4	9	4–5	1	1–2	1	3	1	2	1	10–11	1	7	4	4	3	1	3–4	5	1	3	1	6	79–84
Sheridan	1	3	8	7	1	1–2	1–2	2	1	2–3	1	5	2	2	3	4–5	2	1	1–2	2–3	1	3	1	5	60–66
LRMC	1	2	6	6	1	1	1	2	1	2	1	2	2	1	4	2	2	1	1	1	1	2	1	2	46

Note. $n+$ denotes that a team was unranked in the poll and that $n-1$ teams in the region were ranked; therefore, the team ranks nth at best.

Table 5. Significance of LRMC's performance against other pre-tournament within-region prediction methods, 1999–2000 through 2004–2005 seasons (24 total regions).

Prediction method (x)	LRMC rank higher than x (No. of teams)	x rank higher than LRMC (No. of teams)	One-tailed significance
AP	12	4	0.04
ESPN	10	5	0.15
Seed	12	2	0.01
RPI	13	2	0.004
Massey	9	3	0.07
Sagarin	10	2	0.02
KG	12	2	0.01
Sheridan	9	1	0.01

LRMC model has a better total than any of the other methods. Table 5 shows the results of one-tailed McNemar's significance tests on the contents of Table 4. The tests indicate that LRMC is better than AP, Seed, RPI, Massey, Sagarin, KG, and Sheridan at a significance level of 0.07 or better; all but AP and Massey are at the 0.02 or better significance level. Only the comparison with ESPN (0.15) has a poorer significance level than 0.07. Most striking is the comparison with the NCAA tournament seeding committee; LRMC outperforms the seeds at a 0.01 significance level.

Conventional wisdom is that there have been surprise teams in the Final Four every year, but there is no standard definition of surprise. If we define a surprise team to be one outside the top two teams in the region, Table 4 demonstrates that there have been only three teams to surprise our model in the past 6 years, including two in one season. By contrast, Massey has been surprised 7 times, Sheridan 8–10 times, Sagarin, AP, and ESPN 9 times each, the tournament selection committee (Seed) 10 times, RPI 12 times, and KG 14 times.

4.3. Dynamic-Programming-Based Ranking Results

In addition to testing the ranking systems assuming that the best team would always win, we also tested the effectiveness of the ranking systems in the framework of Kaplan and Garstka's [13] dynamic programming model.

Kaplan and Garstka's [13] dynamic program requires team-vs.-team win probabilities for each possible tournament matchup. There are several methods for translating rankings to probabilities suggested in the literature. As before, let p_{ij} be the probability that team i will beat team j on a neutral court. Schwertman, McCready, and Howard [19] suggest the ratio of rankings $p_{ij} = y_j/(y_i + y_j)$, where y_i and y_j are the rankings of teams i and j. (They initially suggested this calculation for use with tournament seeds, but it is easily extendable to other ranking systems.) Bouler and Stekler [2] suggest a probit model that they fit for tournament seeds only.

Schwertman, McCready, and Howard [19] and Schwertman, Schenk, and Holbrook [20] suggest probabilities based on the assumption that teams' strength is normally distributed. Both sets of researchers propose probabilities of the form $p_{ij} = \alpha_0 + \alpha_1(S(y_i) - S(y_j))$, where $S(y_i)$ is the inverse normal cumulative distribution function of team i's ranking relative to the total. For example, if there were 325 teams and team i was ranked $y_i = 21$st, then $S(y_i)$ would be the inverse normal CDF of (325-21)/325. The parameters α_0 and α_1 are fit based on regular-season data. In [19], they define $\alpha_0 = 0.5$ (so that teams of equal strength are assigned a 50% chance of beating each other) and fit α_1; in [20] they also consider fitting both α_0 and α_1. Since this second model might yield $p_{ij} + p_{ji} \neq 1$, we define $p_{ji} = 1 - p_{ij}$ whenever i is a higher-ranked team than j. We also truncate meaningless values of p_{ij}, negative values are assigned 0, and values greater than 1 are reduced to 1. Schwertman, Schenk, and Holbrook [20] suggest similar one- and two-parameter fits based on an exponential probability function:

$$p_{ij} = \frac{1}{1 + e^{\beta_0 + \beta_1(S(y_i) - S(y_j))}}.$$

In the one-parameter fit, $\beta_0 = 0$ ensures that teams of equal strength are assigned a 50% chance of beating each other; in the two-parameter fit, we handle out-of-range probabilities and $p_{ij} + p_{ji} \neq 1$ in the same way as before.

The final ranking-based probability system we test is from Carlin [7]. He suggests using a two-parameter fit to calculate an expected point difference $\hat{x}_{ij} = \gamma_0 + \gamma_1 (y_1 - y_j)^2$ and then estimating the probability p_{ij} from the cumulative normal distribution, i.e., $p_{ij} = \Phi(\hat{x}_{ij}/\sigma)$. We use $\sigma = 11$, as suggested by [7].

Kaplan and Garstka's [13] dynamic programming model is designed for use with tournament pools. There are many different pool scoring systems; we tested ours on three common systems, each of which emphasizes additional solution features.

The first type of pool we tested awards one point per correctly predicted game, regardless of which round the game is in. This type of pool makes the first- and second-round prediction quality more important than later rounds, simply because more than ¾ of all tournament games occur in those first two rounds.

The second type of pool we tested awards an exponentially increasing number of points per correct prediction, based on the round that the game occurs. Specifically, we tested a system where each correct prediction earns $2^{\text{round-1}}$ points (i.e., 1 point per first-round game, 2 points per second-round game, 4 points per third-round game, etc.).

Table 6. Total pool score of models using best-team-wins prediction method, 1999–2000 through 2004–2005 seasons.

Pool type	AP	ESPN	Seed	RPI	Massey	Sagarin	KG	Sheridan	LRMC
One point per game	236	235½	235¼	229	242	229	231½	244½	248
$2^{\text{round-1}}$ points	541	531½	495	465	534	520	519	565	632
seed$\times 2^{\text{round-1}}$ points	1194½	1192½	1115	1104	1232	1155	1188½	1247	1454

This type of pool makes later-round predictions more important than early predictions; however, later-round predictions are more difficult, because they require the predicted winner to win not just that game, but all of its previous tournament games as well.

The third type of pool we tested follows the exponential increase system, but rewards correct upset predictions. Specifically, the base number of points for each game remains the same ($2^{\text{round-1}}$), but is multiplied by the seed of the predicted winning team. For example, in a first-round game between a No. 2 seed and a No. 15 seed, the base point value of the game is $2^{\text{I-1}} = 1$ point. Correctly predicting the No. 2 seed to win earns $2\times 1 = 2$ points, while correctly predicting an upset (the No. 15 seed winning) earns $15\times 1 = 15$ points. This method rewards not only correct long-term predictions, but also better insight into which lower-seeded teams will be successful.

Before testing any of the dynamic-programming-based predictions, we tested the best-team-wins method on each of the three pool types. The previous section's section's results, which suggested that the LRMC method picked more winners and was significantly superior at selecting later-round winners (especially those that might be lower-seeded), led us to expect that LRMC would increase its superiority in exponentially weighted pools and upset-bonus pools.

In fact, as Table 6 shows, this is exactly what happened. For the one-point-per-game pool, LRMC was 1.5% better than the second-best ranking method and 5% better than the average of the other eight methods. For the exponentially weighted pool, LRMC's advantage increased to 11% over the second-best method and 18% over the average of the other eight methods. When the upset bonus was included, LRMC's advantage was even greater, 14% over the second-best method and 19% over the average.

In addition to the results reported in Table 6, we also tested the maximum score estimator model of Caudill [8], another deterministic method. Based on seedings, it uses historical data to predict outcomes (so, for example, if 14th seeds beat 3rd seeds more often than vice versa, it will make this prediction instead). Its performance was worse than that of just selecting the higher seeds (Seed).

Tables 7, 8,, and 9 compare the dynamic programming-based predictions using each ranking method and each ranking-to-probability formula; the final row of each table show the best-team-wins method for purposes of comparison. Again, the results are clear. Regardless of which method is used to derive probabilities from the rankings, the LRMC results are superior. In fact, in every case, even the worst LRMC result (without dynamic programming) is superior to the best result obtained from any of the other rankings, using any of the probability models, with or without dynamic programming. Thus, we can conclude that for the scoring systems we tested, although selecting a good probability model and using dynamic programming both can improve the results, it is more important to begin with a good ranking system. The dynamic program added to LRMC is especially effective in the most complex pool scoring model we tested, where upset incentives are more likely to make picking the better team a suboptimal strategy.

4.4 Dynamic-Programming-Based Rating Results

Four of the ranking methods we have tested, Massey, Sagarin, KG, and LRMC, actually give more data than just the relative ranks of teams. All three assign a rating to a

Table 7. Total one-point-per-game pool score of models using ranking-based dynamic programming prediction methods and best-team-wins, 1999–2000 through 2004–2005 seasons.

	AP	ESPN	Seed	RPI	Massey	Sagarin	KG	Sheridan	LRMC
Ratio of rankings [19]	242	240	241	235	242	238	236	247	250
Linear, $\alpha_0 = 0.5$ [19]	243	238	241	235	238	240	236	246	251
Linear, α_0 and α_1 fit [20]	242	238	241	236	241	240	235	246	249
Exponential, $\beta_0 = 0$ [20]	243	240	241	236	231	241	237	244	254
Exponential, β_0 and β_1 fit [20]	241	238	241	236	241	240	237	247	251
Normal CDF [7]	239	241	241	232	240	237	238	245	253
Seed probit [2]	—	—	240	—	—	—	—	—	—
Best-team-wins	236	235½	235¼	229	242	229	231½	244½	248

Table 8. Total $2^{\text{round-1}}$-points-per-game pool score of models using ranking-based dynamic programming prediction methods and best-team-wins, 1999–2000 through 2004–2005 seasons.

	AP	ESPN	Seed	RPI	Massey	Sagarin	KG	Sheridan	LRMC
Ratio of rankings [19]	546	534	520	470	534	523	531	579	625
Linear, $\alpha_0 = 0.5$ [19]	548	528	520	470	518	543	531	583	633
Linear, α_0 and α_1 fit [20]	546	528	520	472	530	541	529	583	623
Exponential, $\beta_0 = 0$ [20]	541	530	520	472	524	545	535	582	651
Exponential, β_0 and β_1 fit [20]	538	528	520	472	530	543	535	587	633
Normal CDF [7]	538	580	520	420	518	521	557	575	661
Seed probit [2]	—	—	519	—	—	—	—	—	—
Best-team-wins	541	531½	495	465	534	520	519	565	632

team; the Sagarin and KG ratings are meant to be directly translated to point differentials between teams, while in Section 3.2 we have described how to translate LRMC ratings to estimated point differentials. Carlin [7] and Kaplan and Garstka [13] have discussed ways of using estimated point differentials $\lambda_i - \lambda_j$ to estimate team-vs.-team win probabilities p_{ij}. Specifically, Carlin [7] suggests a Normal model using $p_{ij} = \Phi((\lambda_i - \lambda_j)/\sigma)$, where σ is conservatively estimated to be approximately 11. Kaplan and Garstka [13] use a Poisson model to refine the estimate of σ, suggesting

$$p_{ij} = \Phi\left(\frac{\lambda_i - \lambda_j}{\sqrt{\lambda_i + \lambda_j}}\right).$$

These models can be used either with Sagarin ratings or with Kaplan and Garstka's [13] Vegas-based ratings. Carlin [7] also gives a refined probability estimate for Sagarin ratings, noting that teams' observed point difference tends to be slightly underestimated by the Sagarin method. He fits a linear model and obtains the estimate $p_{ij} = \Phi(1.165(\lambda_i - \lambda_j)/\sigma)$ for use with Sagarin ratings. In Section 3.2, we describe two possible methods for translating LRMC ratings to win probabilities, one based on point differences and one directly fitting a logistic regression model. Massey's rating pages [15] provide probability estimates based on his ratings; his estimates are directly calculated using his formulas.

Tables 10, 11, and 12 show the performance of each model in the three pool scoring systems. Again, just as with ranking-based methods, LRMC even without probability models or dynamic programming outscores all of the other methods in any form on the 1999–2005 data. We note, though, that while using ratings instead of rankings helps the Sagarin, KG, and Massey methods, the best LRMC results are obtained from rankings. This suggests that, although the LRMC model appears to give better predictions, we do not yet have a good method for deriving probability estimates from LRMC ratings. For now, even using slightly unreliable probability estimates is sufficient to outperform the other methods; however, we also point out that this opportunity for future research might yield a method that gives even better results.

4.5. Progressive Predictions

In the preceding computational results, we have used all of the pre-NCAA-Tournament games to construct our transition matrix. In this section, we consider the question "how much data is enough?". Specifically, we test the quality of LRMC's tournament predictions given only the first d days of each season's data, for various values of d.

The NCAA basketball season usually begins in mid-November, and the NCAA Tournament does not begin until mid-March; therefore, 4 months of data are available. However, early in the season we cannot use the LRMC predic-

Table 9. Total seed$\times 2^{\text{round-1}}$-points-per-game pool score of models using ranking-based dynamic programming prediction methods and best-team-wins, 1999–2000 through 2004–2005 seasons.

	AP	ESPN	Seed	RPI	Massey	Sagarin	KG	Sheridan	LRMC
Ratio of rankings [19]	1300	1262	1139	1072	1228	1213	1205	1342	1654
Linear, $\alpha_0 = 0.5$ [19]	1294	1280	1094	990	1339	1210	1147	1219	1806
Linear, α_0 and α_1 fit [20]	1208	1249	1120	1031	1369	1175	1203	1351	1832
Exponential, $\beta_0 = 0$ [20]	1226	1186	1026	1069	1344	1146	1180	1352	1838
Exponential, β_0 and β_1 fit [20]	1242	1229	1006	1053	1292	1192	1208	1400	1843
Normal CDF [7]	1210	1260	1139	1023	1161	1224	1149	1277	1611
Seed probit [2]	—	—	1174	—	—	—	—	—	—
Best-team-wins	1194½	1192½	1115	1104	1232	1155	1188½	1247	1454

Table 10. Total one-point-per-game pool score of models using rating-based dynamic programming prediction methods and best-team-wins, 1999–2000 through 2004–2005 seasons.

	Sagarin	KG	MasseyProb	LRMC
Normal [7]	240	239	—	—
Poisson [13]	240	239	—	—
Sagarin fit [7]	240	—	—	—
Massey probabilities [15]	—	—	235	—
LRMC points	—	—	—	252
LRMC direct	—	—	—	252
Best-team-wins	229	231½	242	248

Table 12. Total seed$\times 2^{\text{round-1}}$-points-per-game pool score of models using rating-based dynamic programming prediction methods and best-team-wins, 1999–2000 through 2004–2005 seasons.

	Sagarin	KG	MasseyProb	LRMC
Normal [7]	1362	1366	—	—
Poisson [13]	1247	1322	—	—
Sagarin fit [7]	1292	—	—	—
Massey probabilities [15]	—	—	1347	—
LRMC points	—	—	—	1600
LRMC direct	—	—	—	1705
Best-team-wins	1155	1188½	1232	1454

tion method; the method is only viable once full connectivity is reached. If, on some day d, there are two sets of teams S and S' such that no team in S has yet played a team in S', then the Markov chain equations will not have a unique solution.

For the 1999–2000 through 2004–2005 seasons, our limiting factor was the first season, in which full connectivity was not reached until January 2. At the other extreme, the earliest final date of the pre-tournament season was March 13. Therefore, we tested the quality of LRMC's predictions based on data from the beginning of the season until day d, for $d \in \{\text{January 2}, \ldots, \text{March 13}\}$.

Figure 6 shows the progression of the LRMC predictions. We tested two measures: the number of correct bracket predictions and the number of games in which the higher-ranked team won. Both measures are scaled relative to the full-season's data set; a value of 1 indicates that the number of correct predictions was equal to that using the full season's data, values less than 1 are worse than predictions made from the full season's data, etc.

As Figure 6 demonstrates, the quality of predictions increased from early January (87% for bracket predictions, 94% for higher-ranked teams) through mid-February, but additional data after mid-February did not seem to improve the prediction quality (and possibly made predictions a bit worse, though not significantly). We hypothesize that by mid-February, enough data have been collected to reduce the effects of noise. It is also possible that games in March might be slightly less representative of some teams' quality due to varying motivation. Some teams are clearly "out of the running" or have locked up an NCAA tournament bid based on the quality of their November–February performance, while other teams (known as "bubble teams") will have their bid status decided by their final few games; such bubble teams might have more motivation and consequently might temporarily play better relative to less-motivated opponents.

Interestingly, using the previous year's final rankings to predict the next year's NCAA tournament results achieved a ratio of 81 (for bracket predictions) and 92% (for higher-ranked teams). Although both ratios are lower than even the January 2 results, it does still indicate a reasonable measure of consistency in relative team quality from one year to the next.

Table 11. Total $2^{\text{round-1}}$-points-per-game pool score of models using rating-based dynamic programming prediction methods and best-team-wins, 1999–2000 through 2004–2005 seasons.

	Sagarin	KG	MasseyProb	LRMC
Normal [7]	560	548	—	—
Poisson [13]	560	548	—	—
Sagarin fit [7]	560	—	—	—
Massey probabilities [15]	—	—	564	—
LRMC points	—	—	—	635
LRMC direct	—	—	—	635
Best-team-wins	520	519	534	632

5. CLOSE GAMES AND RANKING SYSTEMS

The logistic regression model described in Section 3.1 and the analysis done in Section 3.2 give rise to an interesting observation about close games. Conventional wisdom, repeated by sportscasters, sportswriters, and fans alike, is that "good teams find a way to win close games." In other words, better teams are frequently able to find some physical or psychological reserves when the outcome of the game is on the line.

Boynton [3] has already shown this idea to be untrue in Major League Baseball. A baseball team's record in close games has less correlation with its overall winning percentage; a more accurate statement is that good teams are more likely to win games that are not close. This result is not surprising, given that opposing teams' run-scoring processes are almost entirely independent. If one team is better than another (its offense is likely to score more runs against the other's pitching/defense than vice versa), then it is more likely to win a non-close game than a close one.

On the other hand, opposing teams' point-scoring processes are less independent in basketball. A good defense can stimulate offensive production by providing turnover-induced fast-break opportunities in which the probability of scoring is much higher than on a normal possession. On the

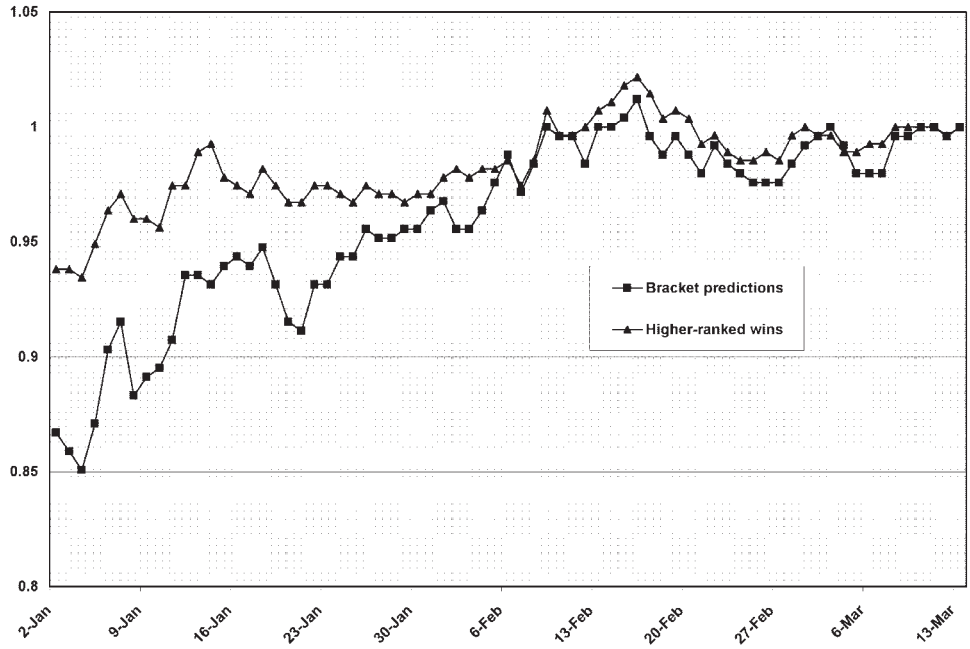

Figure 6. Progression of daily prediction quality relative to final prediction quality.

other hand, a good offense can help defensively as well, especially when the team plays a pressing style of defense that is much easier to implement after the team has just scored. Therefore, one might wonder whether the adage "good teams find a way to win close games" could hold true in basketball even though Boynton [3] has shown it to be untrue in baseball.

However, our results do not support the validity of the conventional wisdom in college basketball. The data from 1999–2003 show that of all 791 teams that won a close home matchup (defined as a spread between 1 and 3 points, or at most one basket), approximately 35% won the road matchup against the same opponent. Of the 713 that lost a close home matchup, approximately 33% won the road matchup. If the better team really is able to win close games more frequently, one would expect the difference in road success to be much larger. Better teams (the ones who, presumably, had won the close games) would be expected to have a comparatively higher road win rate compared to worse teams (who, presumably, had lost the close games). The logistic regression estimate gives similar results; it predicts road win rates of 36 and 33%.

Therefore, rather than good teams winning close games, teams that win several close games (perhaps due more to luck than other factors) might tend to be overrated by fans and the sports media, and teams that lose several close games tend to be underrated. This occurs because an event that might really be a 50/50 (or 35/65) coin flip is translated into a binary win/loss result. In fact, this might explain why our combined logistic regression/Markov chain model is more successful than others in selecting potential Final Four participants. Very good teams that lost a few "extra" close games will tend to be ranked lower than they deserve by the polls, RPI, and other methods that treat wins and losses as binary events; our more-continuous model tends to evaluate those teams more accurately.

6. SUMMARY

The annual NCAA Division I basketball tournament is the largest sports gambling event in the United States. With over $3 billion wagered each year on the outcome of the tournament, bettors turn to expert rankings of teams for help with predictions. The most prevalent ranking systems are the two major polls (the Associated Press poll of sportswriters and the ESPN/USA Today poll of coaches), the Ratings Percentage Index, the Sagarin ratings, the Massey ratings, and the tournament selection committee's seedings; we also tested rankings and ratings derived from Las Vegas odds and betting lines.

In this paper, we describe a LRMC model for predicting the outcome of NCAA Division I basketball tournament games. It uses only basic input data and appears to able to predict individual game outcomes more accurately than the standard ranking systems. Moreover, it is better than other rankings at predicting potential Final Four teams. It also appears to be superior to the NCAA Tournament selection committee's seedings. When tested on three common but diverse NCAA Tournament pool scoring systems, even the

simplest LRMC approach (selecting the higher-ranked team to win each game) outscores the other methods over the past 6 years, even when those methods are supplemented by other researchers' probability and dynamic programming models. When those models are also used with LRMC, the performance of LRMC is even better, especially for more complex pool scoring systems.

We conjecture that part of the reason for the comparative success of our model is that the other models (and perhaps the minds of the NCAA tournament selection committee) treat the outcome of games as binary events, wins and losses. In contrast, our model estimates the probability of the winning team being better than the losing team based on the location of the game and the margin of victory and is therefore able to more accurately assess the outcome of a close game.

The success of our model in predicting the outcome of NCAA tournament games suggests that it gives good rankings of teams and therefore that those rankings might be valid for predicting the outcome of regular-season games as well. For example, one might use data from all games prior to a certain data to predict the outcome of that date's games.

ACKNOWLEDGMENTS

The authors thank our three anonymous reviewers for providing some excellent and helpful suggestions. We also thank Georgia Tech undergraduates Kristine Johnson, Pete Kriengsiri, Dara Thach, Holly Matera, Jared Norton, Katie Whitehead, and Blake Pierce for helping with data collection, coding, and analysis.

REFERENCES

[1] As March Madness Unfolds, the Real Action's in Vegas, USA Today, March 23, 2004.
[2] B.L. Bouler and H.O. Stekler, Are sports seedings good predictors? Int J Forecast 15 (1999), 83–91.
[3] B. Boynton, Are one-run games special? Baseball Res J (1997), 26.
[4] D. Breiter, and B. Carlin, How to play office pools if you must, Chance 10 (1), (1997), 5–11.
[5] T. Callaghan, P.J. Mucha, and M.A. Porter, The bowl championship series: A mathematical review, Not Am Math Soc 51 (2004), 887–893.
[6] T. Callaghan, M.A. Porter, and P.J. Mucha, Random Walker Ranking for NCAA Division 1-A Football. arxiv.org: physics/0310148, 2003.
[7] B. Carlin, Improved NCAA basketball tournament modeling via point spread and team strength information, Am Statist 50 (1996), 39–43.
[8] S.B. Caudill, Predicting discrete outcomes with the maximum score estimator: The case of the NCAA men's basketball tournament, Int J Forecast 19, (2003), 313–317.
[9] S.B. Caudill and N.H. Godwin. Heterogeneous skewness in binary choice models: Predicting outcomes in the men's NCAA basketball tournament, J Appl Statist 29, (2002) 991–1001.
[10] B. Clair and D. Letscher, Optimal Strategies for Sports Betting Pools. Working paper, Department of Mathematics and Computer Science, Saint Louis University, 2005.
[11] CollegeRPI.com web site, http://collegerpi.com, 2005.
[12] ESPN.com daily listing of Sportsinteraction.com betting lines, http://sports.espn.go.com/sports/gen/dailyline, 2000–2005.
[13] E.H. Kaplan and S.J. Garstka, March Madness and the office pool, Manage Sci 47 (2001), 369–382.
[14] K. Massey, Description of Massey rating system. http://www.masseyratings.com/theory/massey.htm, 2005.
[15] K. Massey, Massey Ratings and Massey NCAA tournament win probabilities. http://www.masseyratings.com/archive/cb/ncaa2000.htm through http://www.masseyratings.com/archive/cb/ncaa2005.htm, 2000–2005.
[16] K. Massey, Statistical Models Applied to the Rating of Sports Teams. Honors project in mathematics, Bluefield College. http://www.masseyratings.com/theory/massey97.pdf, 1997.
[17] A. Metrick, March Madness? Strategic behavior in NCAA basketball tournament betting pools, J Econ Behav Org 30, (1996), 159–172.
[18] J. Sagarin, Jeff Sagarin Computer Rankings. Updated weekly; archive of season-end rankings available. http://www.usatoday.com/sports/sagarin.htm, 2005.
[19] N.C. Schwertman, T.A. McCready, and L. Howard, Probability models for the NCAA regional basketball tournaments, Am Statist 45 (1991), 35–38.
[20] N.C. Schwertman, K.L. Schenk, and B.C. Holbrook, More probability models for the NCAA regional basketball tournaments, Am Statist 50, (1996), 34–38.
[21] Statfox, NCAA team and season game logs. http://www.statfox.com/cbb/logs/default.htm, 2000–2005.
[22] USA Today, Daily newspaper. Gannett: McLean, Virginia, 2000–2005.
[23] D.L. Wilson, Bibliography on College Football Ranking Systems. http://homepages.cae.wisc.edu/~dwilson/rsfc/rate/biblio.html, 2005.
[24] Yahoo Sports NCAA Men's Basketball Scores & Schedule. Updated daily, http://.com/ncaab/scor, 1999–2005.