Computational Bracketology

Harrison Hall, John Sigman
Ph.D. Candidates in Engineering
Dartmouth College
{harrison.k.hall.th, john.b.sigman.th}@dartmouth.edu

February 19, 2013

1 Background

The NCAA Men's Basketball Tournament is a 64-team, single elimination tournament held every year that determine's the nation's national champion. Even with 6.45 million brackets, as in Figure 1, filled out on ESPN.com last year[2] the winner still failed to predict 12 of the games correctly[3]. While it is astronomically unlikely that anyone has or will ever picked a perfect bracket, a chance of 1 in 2^{63}, it is clear that the current augmented human predictions are not perfect. While some machine learning algorithms exists which are competitive with the brackets of professional sports analysts, these algorithms are designed to take into account only team-level statistics. While single-game, individual player statistics are available[4] current, published approaches tend to not evaluate the the importance of individual players or potential player match-ups.

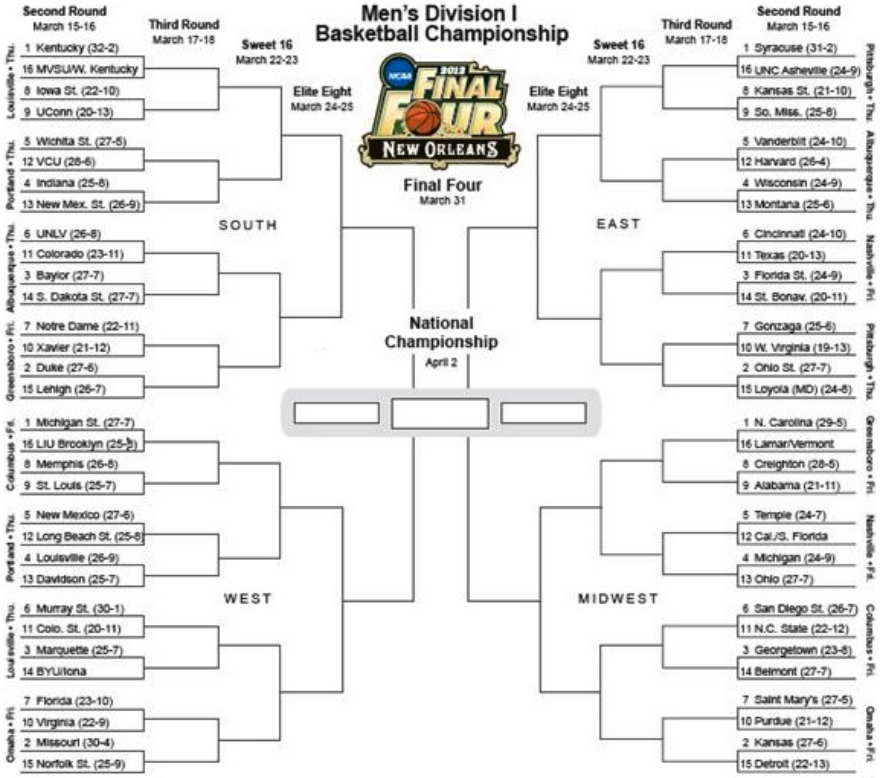

Figure 1: Blank 2012 NCAA Division I Men's Basketball Tournament bracket. Image courtesy of AP.

2 Scope and Goals

Since the proposal the scope of our project, improving the predicted outcome of the NCAA tournament using team and player statistics, has not changed. However the focus of our project has shifted slightly from implementing the additional approach of Logistical Regression/Markov Chains (LRMC)[6][7][1] as another comparison point in favor of attempting to improve RPI[9] as discussed in Section 5 in addition to adding the player-level features as mentioned in our proposal.

3 Dataset

Our dataset was scraped from the ESPN NCAA Men's Basketball website[4]. The set of information provided by ESPN can be seen in Table 1. They provide schedule data for all of the Division I teams that field men's basketball teams and their full schedules going back to the 2001-2002 season. For each game in those years ESPN provides game level box scores. In total we were able to scrape data for the 347 teams, excluding non-D1 opponents, corresponding to 21,366 games.

Table 1: Fields provided by ESPN and their respective datatypes.

(a) Game

Field	Type
id	Integer
Game Time	Datetime
Site Arena	String
Site City	String
Site State	String
Site Arena	String
Home Team	FK(School)
Away Team	FK(School)
Home Team First Half Score	Integer
Away Team First Half Score	Integer
Home Team Second Half Score	Integer
Away Team Second Half Score	Integer
Home Team Final Score	Integer
Away Team Final Score	Integer
Number of Overtimes	Integer
Home Team Overtime Score	Integer
Away Team Overtime Score	Integer
Regular Season	Boolean
NCAA Tournament	Boolean

(b) Player Statline

Field	Type
Game	FK(Game)
Player	FK(Player)
School	FK(School)
Starter	Boolean
Points	Integer
Minutes	Integer
Field Goals	Integer
Field Goal Attempts	Integer
3-Point Goals	Integer
3-Point Goal Attempts	Integer
Free Throws	Integer
Free Throw Attempts	Integer
Offensive Rebounds	Integer
Defensive Rebounds	Integer
Assists	Integer
Steals	Integer
Blocks	Integer
Turn Overs	Integer
Personal Fouls	Integer

(c) School

Field	Type
id	Integer
Name	String
Mascot	String
Conference	FK(Conference)

(d) Conference

Field	Type
id	Integer
Name	String

(e) Player Biography

Field	Type
id	Integer
Name	String
Position	String
Birthday	Date
Hometown	String
Home State	String
Height (Feet)	Integer
Height (Inches)	Integer
Weight (Pounds)	Integer

ESPN also provides player statistics for every game as well as biographic data about each player. From the set of games described above we were able to extract 429,876 player statistics with the fields described in Table 1b. We were also able to pull biographic data as shown in Table 1e. However that data is not on a

timeline so it is not possible to track variance in player weight, height, position, etc. across their college basketball careers.

4 Implementation

4.1 RPI

The Ratings Percentage Index (RPI)[9] is an industry-standard statistic that comes from the following relations:

$$t_i, t_j \in \mathbb{T} = \{\text{Team}_1 \dots \text{Team}_m\} \tag{1}$$
$$n = \text{number of days in the season} \tag{2}$$
$$k \in [1 \dots n] \tag{3}$$
$$\mathbb{G} = \text{The set of all games played in a season} \tag{4}$$
$$\mathbb{O}_{i,k} = \text{The set of all teams } t_i \text{ played in the first } k-1 \text{ days of the season} \tag{5}$$
$$g_{i,j,k} \in \mathbb{G} \; s.t. \text{ the game played between } t_i \text{ and } t_j \text{ is on day } k \tag{6}$$
$$H^w_{i,j,k} = \text{Indicator function if } g_{i,j,k} \text{ exists and } i \text{ won at home} \tag{7}$$
$$H^l_{i,j,k} = \text{Indicator function if } g_{i,j,k} \text{ exists and } i \text{ won at home} \tag{8}$$
$$A^w_{i,j,k} = \text{Indicator function if } g_{i,j,k} \text{ exists and } i \text{ won away} \tag{9}$$
$$A^l_{i,j,k} = \text{Indicator function if } g_{i,j,k} \text{ exists and } i \text{ won away} \tag{10}$$
$$N^w_{i,j,k} = \text{Indicator function if } g_{i,j,k} \text{ exists and } i \text{ won at a neutral site} \tag{11}$$
$$N^l_{i,j,k} = \text{Indicator function if } g_{i,j,k} \text{ exists and } i \text{ won at a neutral site} \tag{12}$$
$$a = 0.25 \tag{13}$$
$$b = 0.5 \tag{14}$$
$$c = 0.25 \tag{15}$$
$$d, e, f \in \mathbb{R}[0 \dots 2] \tag{16}$$

$$RPI_{i,k} = a \cdot WP_{i,k} + b \cdot OWP_{i,k} + c \cdot OOWP_{i,k} \tag{17}$$
$$WP_{i,k} = \frac{\sum_{x=1}^{k-1} \sum_{y=1}^{m} d \cdot H^w_{i,y,x} + e \cdot A^w_{i,y,x} + f \cdot N^w_{i,y,x}}{\sum_{x=1}^{k-1} \sum_{y=1}^{m} d \cdot H^w_{i,y,x} + e \cdot A^w_{i,y,x} + f \cdot N^w_{i,y,x} + (2-d) \cdot H^l_{i,y,x} + (2-e) \cdot A^l_{i,y,x} + (2-f) \cdot N^l_{i,y,x}} \tag{18}$$
$$OWP_{i,k} = \frac{\sum_{o \in \mathbb{O}_{i,k}} WP_{o,k}}{||\mathbb{O}_{i,k}||}[1] \tag{19}$$
$$OOWP_{i,k} = \frac{\sum_{o \in \mathbb{O}_{i,k}} \sum_{p \in \mathbb{O}_{o,k}} WP_{p,k}}{\sum_{o \in \mathbb{O}_{i,k}} ||\mathbb{O}_{o,k}||} \tag{20}$$

RPI is a good way to rank teams and correct for the strength of an individual teams schedule. For NCAA Basketball, the wins and losses are weighted so that a win at home counts as $d = 0.6$, and a win on the road counts for $e = 1.4$ wins. Away losses count as $2 - e = 0.6$ and home losses count as $2 - d = 1.4$ losses. Neutral site games are counted as away games for both opponents, thus $d = f$. These weightings are to compensate for Home-Court Advantage, which we will discuss in our results.

[1] Typically, $OWP_{i,k}$ is calculated by omitting the meetings of team i with all of its opponents, however in this definition it was omitted for succinctness.

4.2 Ordinal Logistic Regression Expectation

Ordinal logistic regression modeling and expectation (OLRE) [9] is generates ratings for the teams in the NCAA tournament as a function of team-level performance features. The algorithm uses the 64 teams selected and the fact that each team in the tournament will win between 0 and 6 games, and we a priori know the number of teams that win each number of games. This allows us to form an expectation on the number of wins a particular team will win in the tournament.

The specification from the paper states that they only use the season winning percentage, the season points differential of points the team scored versus points scored against them, a proprietary strength of schedule metric calculated by Jeff Sararin, the number of wins the team has against top 30 teams as determined by the strength of schedule metric at the end of the regular season, and the total number of wins the team has recorded in the NCAA tournament in all seasons of your dataset. As we do not have the Sagarin ratings, we will use RPI as defined in Section 4.1 to calculate strength of schedule and wins against top 30 teams.

The probability that team i will win j games where $j \in 0, 1, 2, 3, 4, 5$ is given by π_{ij} as follows:

$$\pi_{ij} = \frac{exp(\alpha_j + \mathbf{x_i}\beta)}{1 + exp(\alpha_j + \mathbf{x_i}\beta)} - \sum_{k=0}^{j-1} \pi_{ik}, \tag{21}$$

α_j is the intercept for the j-th outcome, x_i is a vector of values for team i on the team-level predictor variables, β is a vector of coefficients from the training of the model to fit the data using logistic regression. This generates a 64x7 matrix of predicted probabilities that team at row_i will have $column_j$ wins.

Table 2: The number of teams in the NCAA tournament that will achieve j tournament wins in a single season.

j	Number of Teams
0	32
1	16
2	8
3	4
4	2
5	1
6	1

We know a priori the number of teams with a certain number of wins, as shown in Table 2. Thus the sum of the j-th column in the constructed table must equal the number of teams with exactly j wins while simultaneously having each row sum to 1. This is the form of a contingency table. Therefore we can use maximum likelihood estimation to fit a Multinomial-Poisson Homogeneous Models using the a software package written by Joseph Lang [8] which adjusts the probabilities while satisfying the marginal constraints we described.

$$E_i[WINS] = \sum_{j=0}^{6} j \cdot \hat{\pi_{ij}} \tag{22}$$

Finally the expected number of wins of each team i is calculated as shown in Equation 22.

5 Results

Using our compiled data from dataset listed in Section 3, we computed a table of NCAA teams and their RPI scores across every date of the season. Using the same records, and across every matchup, we were able

to predict a stronger and weaker team for each match by their pre-computed RPI score on the date prior to their matchup. For the purposes of this investigation, we are designating the team with the higher RPI score at each game as being its predicted Winner. The results of a seasons worth of RPI comparison are shown in Figure 2.

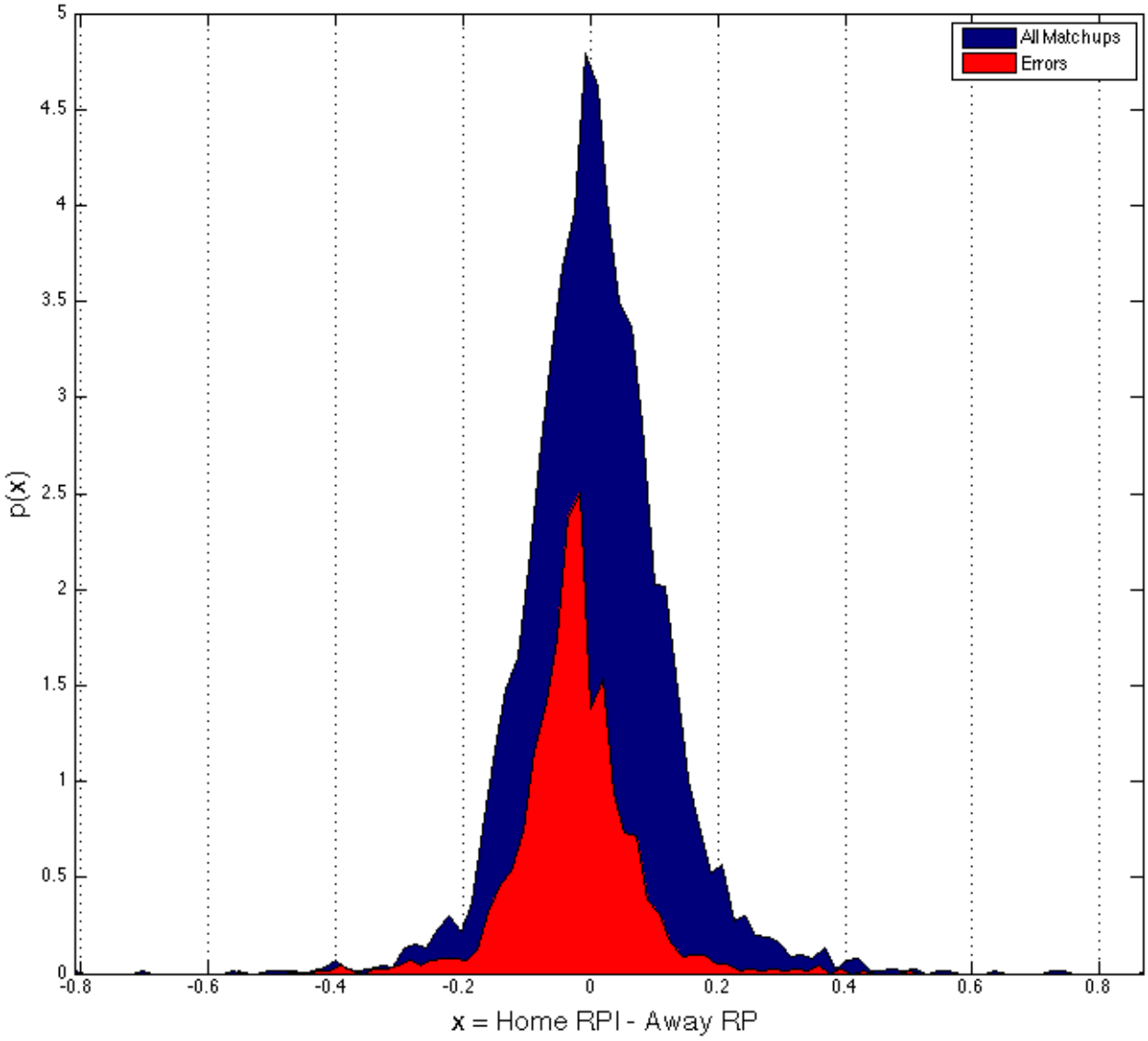

Figure 2: Probability Density of RPI Differences

The figure is a probability density of all the matchups as a function of the difference in RPI score of the home and away teams before to the game. The x-axis is in order of increasing home team favor towards the right of the figure. The error rate can be derived from this figure by the ratio of the area under the red curve (total number of errors) to the area under the blue curve (total number of matchups).

Our RPI-derived error rate across the 2008-2009 season was 33.43%, evaluated across $5,070$ games. The RPI in-tournament error for this season was 27%. The errors for the season were distributed as shown in Figure 3.

The errors are confined to a generally narrower band of RPI score differences than the overall distribution. Their expected value was -0.0164, which indicates a skew of RPI errors when the Away team was projected to win, however we have yet to determine it's statistical significance. This is a representation of the common

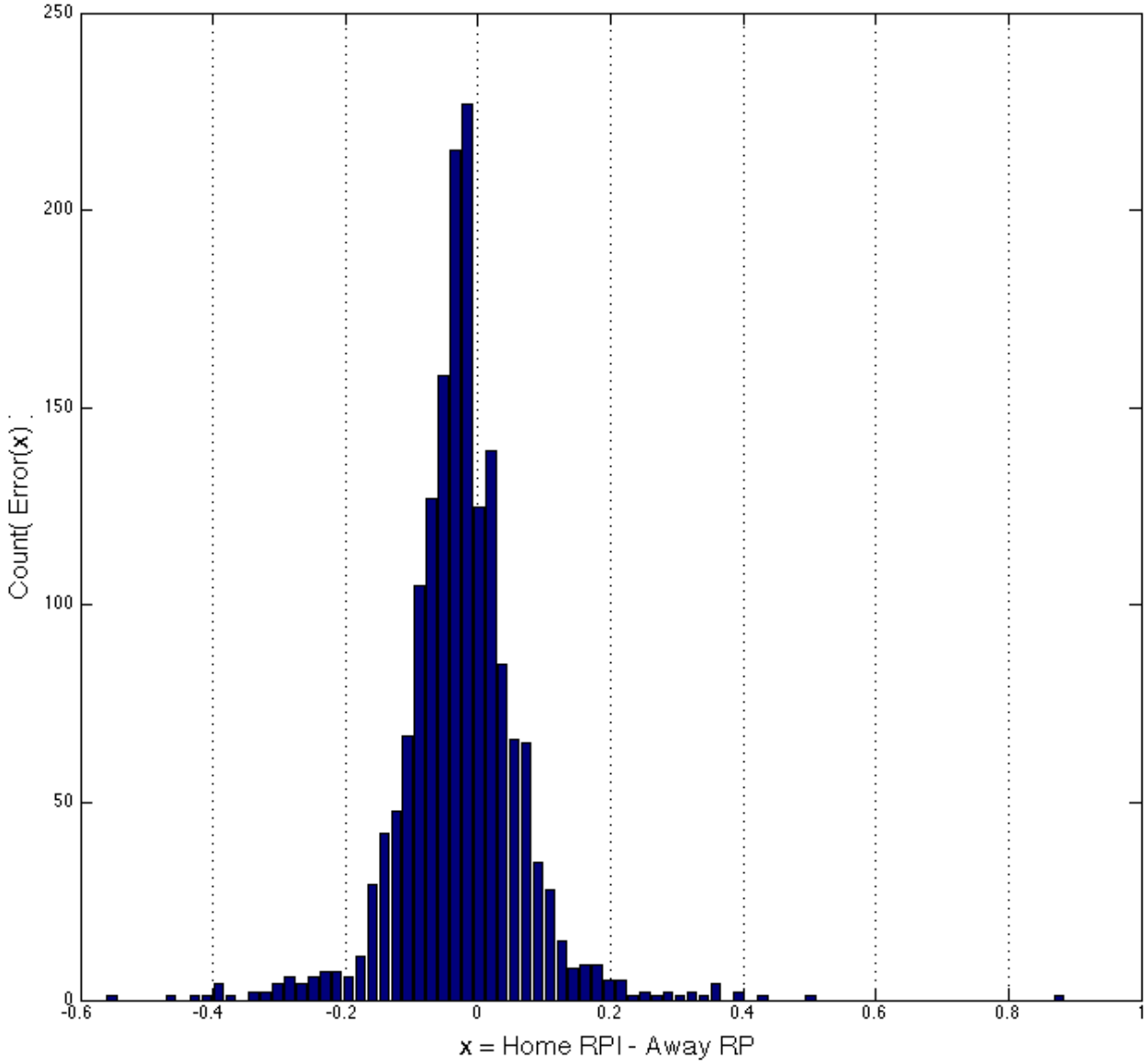

Figure 3: Distribution of Errors as a Function of RPI Difference

effect in sports called Home-Court Advantage.

6 Future Work

Our first step will be to apply the principles of machine learning to the RPI algorithm. The NCAA uses weights of $e, f = 1.4$ and $d = 0.6$ for a teams away and home wins, respectively. These numbers are too round to be statistically tuned, and therefore, we think there is a good chance for their improvement. After re-computing the RPI tables with all the information we need to readily weight the parameters of the rating ourselves, we can optimize RPI using iterative gradient descent. We are very curious to see both the magnitude of the improvement as well as the amount of change in the parameters. Since RPI is a central feature in our implementation of OLRE, once we have the new values will compare the performance of OLRE with and without the tuned RPI.

Our proposed implementation of AdaBoost[5] is lagging from what we stated in our proposal. While we have been defining features from our dataset, we have not defined or trained nearly enough of them to satisfy the large number of classifiers required for sufficient boosting. We are going to automate scripts to build these features and train them over the next few days.

The reach goal for our project, after a discussion with Professor Torresani, is to implement k-means clustering across the players using their game statistics and biological data as they perform against each team. From this we would get clusters of similar players to calculate an expectation on the number of points that player will score and sum across all players on the team to get an expected team score. Then a simple comparison of expected score will give us the game outcome. While this seems promising and is in the spirit of what we set out to do, it may be a stretch that we complete it by the end of term.

References

[1] M. Brown, P. Kvam, G. Nemhauser, and J. S. Sokol, "Insights from the lrmc method for ncaa tournament prediction," *MIT Sloan Sports Analytics Conference*, vol. 53, Mar. 2012.

[2] K. Chong-Adler. (2012, Mar.) Espns tournament challenge sets bracket record with 6.45 million entries. [Online]. Available: http://frontrow.espn.go.com/2012/03/espns-tournament-challenge-sets-bracket-record-with-6-45-million-entries/

[3] CincyFan007. 1. [Online]. Available: http://games.espn.go.com/tournament-challenge-bracket/en/entry?entryID=970323

[4] ESPN. Ncaa-mens college basketball. [Online]. Available: http://espn.go.com/mens-college-basketball/

[5] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, Aug. 1997.

[6] P. Kvam and J. S. Sokol, "A logistic regression/markov chain model for ncaa basketball," *Naval Research Logistics*, vol. 53, 2006.

[7] ——, "An improved lrmc method for ncaa basketball prediction," *Journal of Quantitative Analysis in Sports*, vol. 6, no. 3, May 2010.

[8] J. B. Lang, "Homogeneous linear predictor models for contingency tables," *J Amer Statist Assoc–Theory and Methods*, pp. 121–134, 2005.

[9] B. T. West, "A simple and flexible rating method for predicting success in the ncaa basketball tournament: Updated results from 2007," *Journal of Quantitative Analysis in Sports*, vol. 4, no. 2, Apr. 2008.