Using Cumulative Win Probabilities to Predict NCAA Basketball Performance

Mark Bashuk
Business Intelligence, RaceTrac Petroleum
Atlanta, GA USA 30339
Email: bashukma@gmail.com

Abstract

Traditional ranking methods such as Average Scoring Margin (ASM) or the Ratings Percentage Index (RPI) are limited in their accuracy and usefulness because they focus on just the final score of the game, not how the game arrived at the final score. In this paper I propose a new method that looks at cumulative win probabilities over the duration of a game to measure both team and an individual player's performance. Using five years of gameplay data to generate a Win Probability Index for NCAA basketball, I am able to create an open system that allows anyone to measure the impact, in terms of win probability added, of each play. This method is more accurate than either the ASM or RPI while also providing a more detailed level of player and play specific detail. My initial design includes input adjustments for conference play, home/away/neutral site games, and a team's strength of schedule. Outputs of this study include player rankings, team rankings, and strength of schedule rankings. Detailed explanations of my methodology and the simulations used to build the model, a comparison to existing methods, and an exploration of futures uses of this data are included.

1 INTRODUCTION

The fact that the last play of a game can completely change how we feel about and evaluate a game has always frustrated me. My belief is that even though a last-minute buzzer-beater can be exhilarating (or punishing if you are on the wrong side of it), the outcome of a single play should not influence our overall judgment of a team's performance. Many rating systems combine margin of victory and pace data to create their team metrics and attempt to handle this issue by pointing out that a 1 point loss and a 1 point win are not much different and will average themselves out over a season. Even though this is an improvement over using just Wins and Losses (like the RPI), I believe it is still limited in its usefulness because the final score rarely tells the whole story. If a team is up by 20 at half-time, and gives up a few garbage-time buckets to win by 9, it should not be judged the same as a team that was tied at the half, had a 4pt lead with a minute left, and made all their free throws at the end to win by 9. My system, the SevenOvertimes[1] Method (SOM), uses cumulative win probabilities to score each game, and then combines a team's average cumulative win probability with its strength of schedule to rank teams and predict future game outcomes.

2 METHODOLOGIES

The central premise of SOM is that we can use cumulative win probabilities to accurately measure team performance. The process to rank the teams involves three steps: acquire the play-by-play data, process the game metadata, and then calculate the rankings. Most of the play-by-play and metadata data is available on ESPN.com, NCAA.com, or CBSSports.com, and the process to calculate the rankings is done within SQL stored procedures.

2.1 Acquiring the Play-by-Play Data

Developing the process to acquire and import the play-by-play data was, from a technical perspective, the most difficult aspect of this process. I attempted to use many different sources, and each presented its own problems. I

[1] The name SevenOvertimes is a loose reference to the six-overtime game played by UConn and Syracuse in the 2009 Big East Tournament. I registered the domain SevenOvertimes.com shortly after that game and sat on it for two years until I started posting my team rankings on the site. The name does not have any significance beyond that.

ultimately decided to use ESPN.com because it was the most reliable and provided the best access to game metadata and daily schedules. Some unexpected challenges arose, including developing processes to handle overtime games, "team plays" such as timeouts and shot-clock violations, and incremental updates – only updating one day's worth of games at a time.

2.2 Processing the Metadata

In addition to processing the play-by-play data of each game, I needed a separate process to handle all of the game metadata. This included the teams, the score, the location, and the game date. For reporting purposes, the data is stored in a simple star schema with a single `fact_plays` table. The `fact_plays` table can be joined to a `Dimension_Games`, `Dimension_Teams`, and `Dimension_Plays` table. An entity-relationship diagram of the data warehouse can be seen in Figure 1.

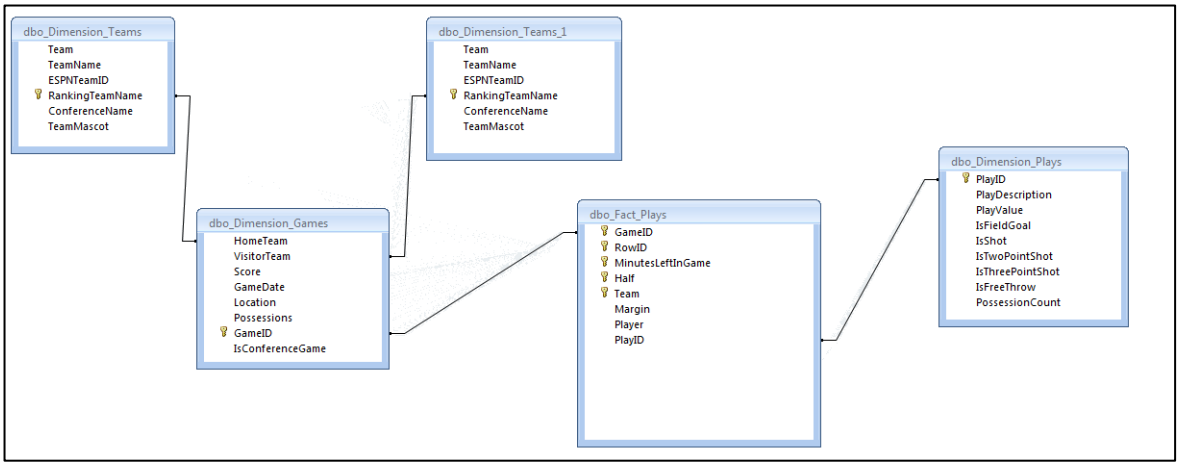


Figure 1: Entity-Relationship Diagram for the core tables used in the data warehouse for SevenOvertimes. There are other tables not pictured, including `Dimension_Conferences`, `Dimension_TimeBuckets`, and `WinProbs`.

2.3 Calculate the Rankings

The process used to calculate the rankings involves two steps, which run after each day's games are processed. The first step populates a `WinProbabilitiesByPlay` table that is a combination of the play-by-play data (stored in `fact_plays`) and the Win Probability Matrix[2] (stored in `WinProbs`). For the games without play-by-play data, I estimate the average win probability by using 50% + the final scoring margin[3].

The second step of the process combines the Cumulative Win Probabilities (CWP) of each game, with each team's strength of schedule to produce a single team rating metric. The simulations (discussed in detail in Section 3) were used to tweak the in-game probabilities and strength of schedule calculation to optimize the overall team metrics. Links to all of the SQL code used to acquire the play-by-play data, process the metadata and calculate the rankings can be found in Appendix 1.

3 OVERIVEW OF SIMULATIONS

3.1 Strength of Schedule and Team Ranking Simulation

Three simulations were run to find the optimal values for the three variables used in the ranking and predictions calculations. The first simulation tested 100 combinations of each teams CWP and its Strength of Schedule

[2] A Win Probability Matrix is a table that gives the expected win probability for a game based on the time left and the lead. For example, if the home team is up by 3 with 4 minutes left in the game my Win Probability Matrix tells us that the home team has a 75.7% chance of winning based on the population of games used to build the matrix. A portion of the Win Probability Matrix and a link to the entire table can be seen in Appendix 2.
[3] This is a placeholder until I am able to develop a more rigorous method of estimating the win probability, or find a more reliable source of data.

(SOS). At the end of the year, the optimal rating system was a combination of 31% of a team's CWP and 69% of its SOS. The results for this analysis, as seen in Figure 1, were very consistent over the three years (2009-2011) used in the analysis. Other systems, such as Ken Pomeroy's, have initial values for each team that slowly degrade as the season progresses [1]. In an effort to mirror that effect, my initial value for SOS is 31% and it goes up by .275% per day until it reaches 69% on the last day of the season.

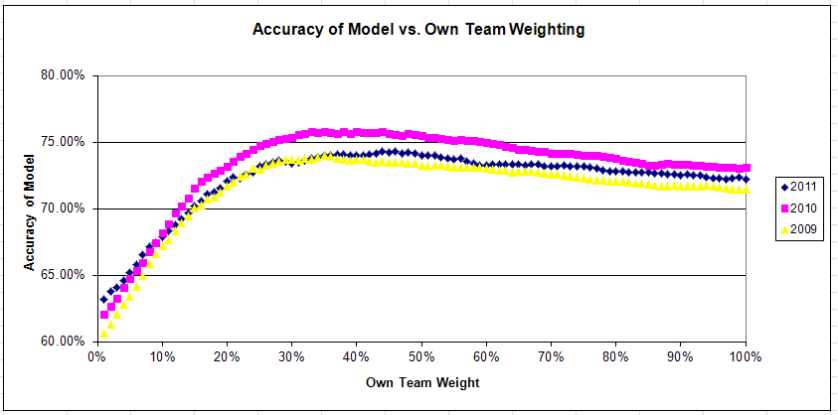


Figure 1: Plot of Model Accuracy vs. Own Team Weighting in the formula for 2009-2011. The above plot shows a similar trend for all three years, and that the ideal weighting is approximately 31%.

3.2 Home Court Advantage Simulation

The second simulation tested different home court advantages by giving the home team a "bump" of a certain percentage and removing the same percentage from the visiting team. The results for this experiment were not nearly as consistent as the SOS simulations seen above. Even though the results in Figure 2 do show a peak near 2.3%, I decided it was more efficient to simply use a home court advantage of 4.74 points. An analysis of 13,937 games from 2007-2012 shows a mean home court advantage of 4.74 points per game which is consistent with results shown by [2] and [3]. The game data used to calculate the Home Court Advantage can be found in Appendix 3.

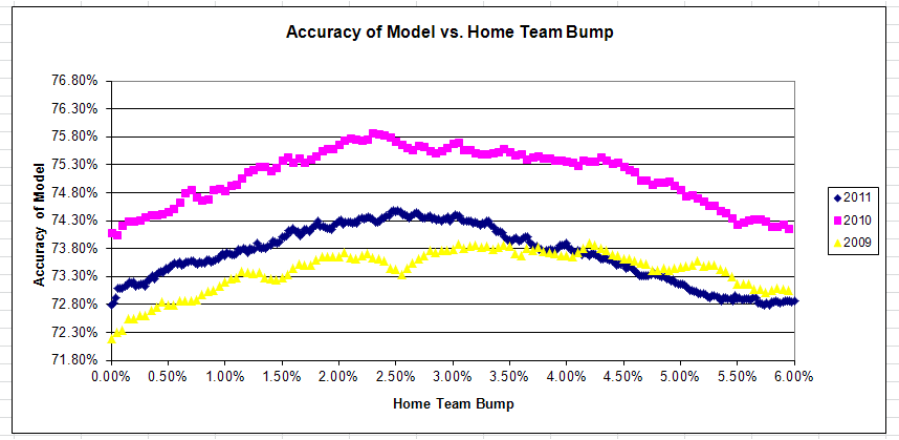


Figure 2: Plot of Accuracy of Model vs. Home Team Bump.

3.3 In-Game Segments Weighting Simulation

The third simulation was designed to test my hypothesis that a team's performance at the end of game is more telling than their performance at the beginning of the game. I originally planned to split the game into ten 4-minute

segments and test four different values for each segment. However, this required 4^10 simulations, and at nearly 30 seconds per run, I did not have the time or computer resources capable of running the simulation. This led me to use eight segments of three values (.33, .66, 1.00) each. This simulation required 6,561 (3^8) trials, and took over 50 hours to complete. The chart seen in Figure 3 shows the accuracy of the trials, grouped by segment weight. The striped columns represent the accuracy at each time segment, with a weighting of .33. The checkered columns represent a weighting of .66, and the darkest section represents 1.00. This chart shows that the accuracy of the models increases as the earlier time segments were given less weighting, and that the end of the game is given more weighting. The most accurate simulation (#2107) had segments weights of (.33, .33, .33, .33, 1.00, 1.00, 1.00, 1.00) and predicted 74.87% of 2011's games correctly. Using a uniform weighting throughout the game proved to be accurate 74.0% of the time, which means the improvements seen in the more accurate simulations match my initial hypothesis that the end of the game is more important than the beginning.

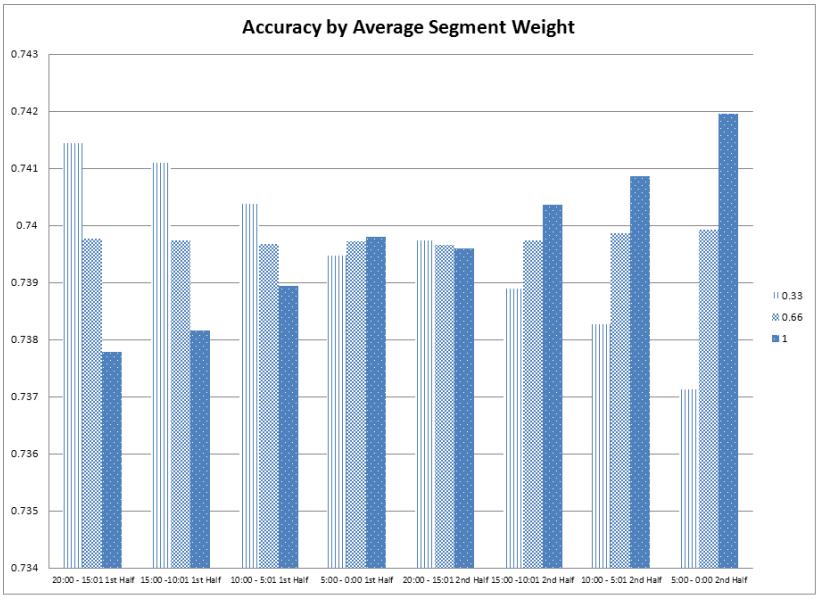


Figure 3: As the early part of the game (first 15 minutes) becomes less important and the end of the game (last 15 minutes) becomes more important, the accuracy of the model increases significantly.

4 COMPARISONS TO OTHER SYSTEMS

The most straightforward way to determine the accuracy and relevance of a rating system is to test it on a daily basis and see how accurate it is at predicting the outcome of games. Much of the work in comparing existing rating systems[4] has already been compiled by Beck [4]. However, there are slight differences between the populations of games used by each system, which can lead to significant differences in the interpretation of these results. SOM, in addition to Ken Pomeroy and Jeff Sagarin, produce predictions for every game featuring two Division 1 teams. The Las Vegas odds might not apply to games featuring some of the smaller Division 1 teams, while they might apply to games featuring major programs against Non-Division 1 opponents (e.g. any games against Chaminade in the Maui Invitational) [5]. Picking the home team to win can only be used for games not played on a neutral court, so it also has a different population of games than the other systems. The accuracy of home court advantage to pick games over the past ten years has been 68.8% [6]. As of the writing of this paper, the aggregate accuracy of sites tracked by Beck, for all games with a spread in the 2011-2012 season, was 73.85% with a standard deviation of about 1.36%.

[4] Systems tracked by Beck include powerratings.com, Compughter Ratings, Jon Dokter, Sonny Moore, StatFox, Sagarin Elo, Sagarin Predictive, the opening line in Las Vegas, and the closing line.

Table 1: Comparisons of Ratings Systems

Rating System	Accuracy
Ken Pomeroy (2010-2011 Season)[5]	77.7%
Vegas Opening Line (2011-2012 Season)	75.2%
Sagarin Predictive (2011-2012 Season)	72.9%
SevenOvertimes (2011 – 2012 Season)	72.6%
NCAA Home Court Advantage (1999-2009)	68.8%
Final RPI Ranking (2010-2011 Season)[6]	64.0%

The formula used to predict the margin of a given game, α, is based on the rankings of the home team (x_h), the visiting team (x_v) and the home court advantage (γ), and can be seen in Equation 1.

Equation 1: $$\alpha = 100 * (x_h - x_v) + \gamma$$

In addition to predicting the outcomes of games, I am including a "confidence" metric that represents my confidence that the home team will win straight up. The formula to compute the confidence, β, of a game with a spread of φ can be seen in Equation 2.

Equation 2: $$\beta = 48.85\% + 3.35\varphi$$

The formula from Equation 2 was derived from an analysis of all games played from 2007 to 2011, and a link to the data and a plot of the data can be found in Appendix 3. The relationship between the spread and the likelihood that home team wins for NCAA basketball comes close to the relationship[7] shown by Stern for NFL football [7].

5 LIMITATIONS OF SYSTEM

Despite efforts to make this system as accurate as possible, there are several key factors limiting its accuracy and usefulness. The single largest problem is the availability of play-by-play data. Sites such as NCAA.com, ESPN.com, and CBSSports.com have play-by-play data available, but many games are either missing entirely or feature incomplete or corrupted data. Examples of corrupted data include scores listed inside the player fields, player names missing, or play descriptions missing entirely. SOM assumes that all the teams are playing at full strength for all games -- it does not currently account for the rigors of travel, injuries, or suspensions. This is one of the advantages of using the opening and closing lines of a game to predict the outcomes – the human-input element of the line will take this into account, while an automated system may overrate a team missing their star player.

Team Rankings are based on CWP, but player ratings are based on Win Probability Added[8] (WPA). Two problems with looking WPA at the player level are that it simply credits a player for scoring (with little regard to efficiency), and that it is not able to account for individual defense. If a defender spends most of his energy shutting down the other team's star player, the WPA metric is not aware of the defender's influence on the game. Because of this, WPA is very flawed in measuring an individual player's performance, but is still accurate at measuring the collective performance of a team. A table of the top 15 players in WPA can be seen in Appendix 4.

6 FUTURE ADJUSTMENTS TO SYSTEM

With more computing resources I could re-run the time-segment simulation with smaller buckets and more options, or I could calculate an individualized home court advantage instead of using the same 4.74 point advantage for all teams. Other simulations I have also identified would be to test how to improve the weight of the SOS factor as the season progresses (as opposed to the uniform increase currently used) and a test to create a "recency effect" that would

[5] Accuracy of Ken Pomeroy sourced from: http://www.teamrankings.com/blog/ncaa-basketball/under-the-teamrankings-hood-part-3-pros-and-cons
[6] Accuracy of RPI for 2010-2011 season detail: http://www.sevenovertimes.com/sql/rpiaccuracy.csv
[7] The relationship shown by Stern is $\beta(NFL) = 50.00\% + 3.00\varphi$
[8] WPA is calculated by finding the difference in win probability between any two consecutive plays in the play-by-play data.

vary the weight of games played as the season progresses. I would like to include coaching data, more adjustments for conference games, as well as adjustments for rivalry games and TV broadcast info. I would also like to get more play-by-play archived to improve the Win Probability Matrix, or even convert the Win Probability Matrix to a continuous function using a logarithmic regression.

In addition to tweaking this system to improve its NCAA basketball accuracy, it could easily be adapted to predict NBA games. Even though it would require significantly more work, the idea of using cumulative win probabilities to measure team performance could be applied to any sport or league – Major League Baseball, NFL football, or soccer are all viable candidates for this system.

7 CONCLUSIONS

I have demonstrated, in detail, a new system for measuring the performance of both teams and players in NCAA basketball. Using a series of simulations I have been able to adjust the system to improve its accuracy, and I have identified future tweaks that will continue to improve its accuracy. The outcome of the simulations has allowed us to quantify the home court advantage in NCAA basketball (4.74 points per game), how much the strength of schedule and team performance should be factors in measuring team quality (69% vs. 31% at the end of the season), and how to properly weight the time segments of a game to see their influence (First 25 minutes are worth 36%, Last 15 minutes are worth 64%). I believe that SOM is very useful for ranking the performance of teams in any sport, and with continued development it can improve greatly.

8 ACKNOWLEDGEMENTS

I would like to thank everyone who helped me develop the model and allowed me to talk about college basketball analytics with them way more than they probably wanted. This includes Andrew Bashuk, Brendan Murphy, Patrick O'Shea, Ben Slocum, Jake Slocum, Tim Dumbacher, Clay McNeil, Robel Bekele, Tyler Warner, Matt Langsen, Chris Dyches, Jeff Cobb, Mark Maclachlan, Steve Lockwood and Chad Monden. I would like to especially thank Kent Polzin at RaceTrac Petroleum for lending the computing time used to run the time-segment simulations.

9 REFERENCES

[1] Pomeroy, Ken. "The Kenpom.com Blog." *2012 Pomeroy College Basketball Ratings*. The Forecast Factory, 16 Nov. 2010. Web. 10 Jan. 2012. <http://kenpom.com/blog/index.php/weblog/what_happens_to_pre-season_ratings_when_its_not_pre-season_anymore/>.

[2] Harville, David A., and Michael H. Smith. "The Home-Court Advantage: How Large Is It, and Does It Vary from Team to Team?" *The American Statistician* 48.1 (1994): 22-28. Print.

[3] "College Basketball Betting - Hoops Wagers." *Free College Football Picks - NFL Picks*. Bettorsworld. Web. 10 Jan. 2012. <http://www.bettorsworld.com/pinnacle-pulse/23.htm>.

[4] Beck, Todd. "ThePredictionTracker Basketball Predicitions."*ThePredictionTracker.com*. 10 Jan. 2012. Web. 10 Jan. 2012. <http://www.thepredictiontracker.com/bbresults.php>.

[5] "Sports Betting and Gambling Odds Online." *Sports Betting Odds, Picks and Statistics*. Covers Media Group. Web. 10 Jan. 2012. <http://www.covers.com/pageLoader/pageLoader.aspx?page=/data/ncb/teams/pastresults/2011-2012/team2361.html>.

[6] Moskowitz, Tobias J., and L. Jon. Wertheim. *Scorecasting: The Hidden Influences behind How Sports Are Played and Games Are Won*. New York: Crown Archetype, 2011. Print.

[7] Stern, Hal. "On the Probability of Winning a Football Game." *The American Statistician* 45.3 (1991): 179-83. Print.

10 APPENDICES

Appendix 1

Import Play-by-Play Data

- **-http://www.sevenovertimes.com/sql/dailyschedules.sql**
- **-http://www.sevenovertimes.com/sql/spteamloop.sql**
- **-http://www.sevenovertimes.com/sql/spallinone.sql**

Process Game Metadata

- **-http://www.sevenovertimes.com/sql/spisconferencegame.sql**
- **-http://www.sevenovertimes.com/sql/spupdategamedate.sql**
- **-http://www.sevenovertimes.com/sql/spscoreloop.sql**
- **-http://www.sevenovertimes.com/sql/spcleanupdimgames.sql**

Calculate Rankings

- **-http://www.sevenovertimes.com/sql/spcalculatewinprobabilities.sql**
- **-http://www.sevenovertimes.com/sql/spcalculaterankings.sql**

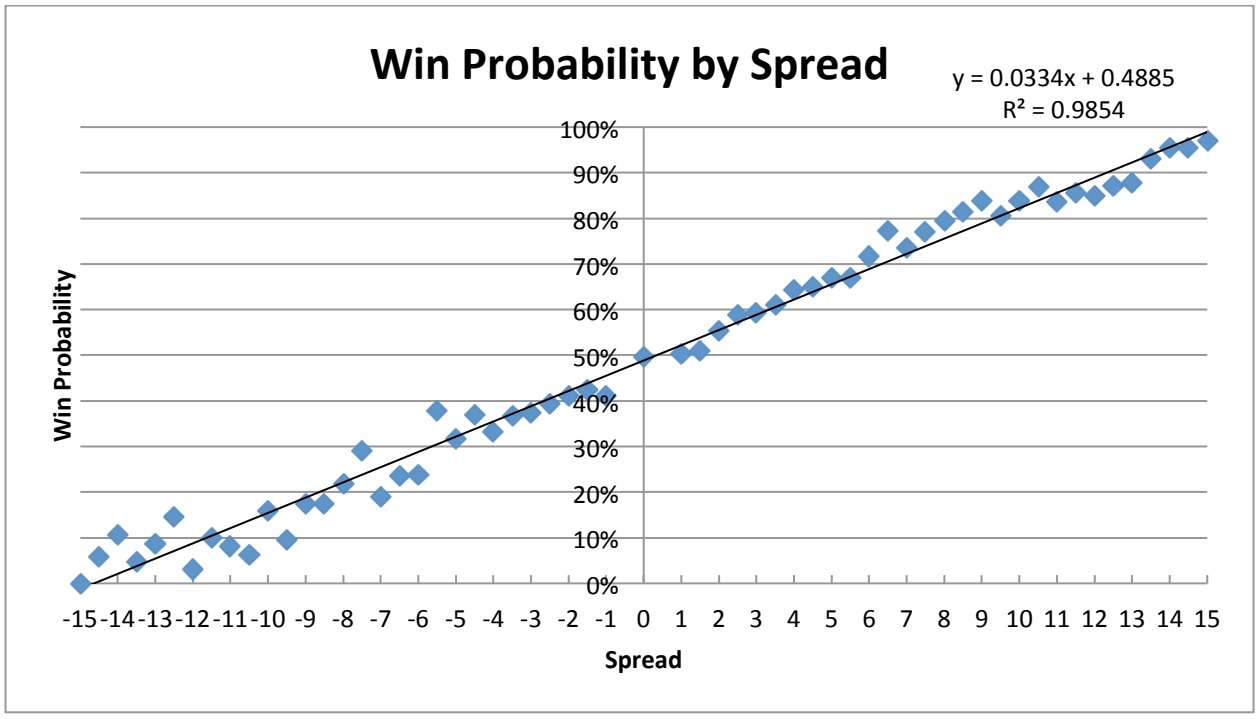
Appendix 2

Lead / Min Left	0	1	2	3	4	5	6	7	8	9	10
-15	0%	0%	0%	2%	13%	11%	10%	8%	1%	7%	3%
-14	0%	0%	0%	0%	3%	0%	0%	11%	7%	3%	15%
-13	0%	0%	0%	5%	2%	0%	0%	3%	3%	11%	4%
-12	0%	0%	3%	0%	0%	2%	3%	2%	3%	1%	2%
-11	0%	0%	3%	1%	1%	1%	1%	2%	4%	7%	4%
-10	0%	0%	0%	0%	0%	1%	2%	1%	4%	4%	7%
-9	0%	0%	2%	5%	4%	9%	8%	6%	13%	5%	14%
-8	0%	0%	2%	8%	6%	5%	7%	4%	15%	10%	7%
-7	0%	2%	9%	6%	11%	9%	12%	12%	12%	10%	15%
-6	0%	3%	11%	3%	4%	6%	4%	7%	15%	21%	26%
-5	1%	5%	11%	10%	12%	18%	14%	8%	26%	19%	18%
-4	1%	16%	14%	7%	16%	24%	14%	28%	24%	27%	32%
-3	3%	12%	19%	15%	30%	22%	36%	48%	34%	39%	27%
-2	11%	27%	34%	32%	25%	31%	36%	39%	34%	29%	36%
-1	25%	46%	41%	36%	32%	39%	58%	40%	51%	36%	40%
0	47%	55%	49%	51%	45%	45%	50%	49%	52%	47%	53%
1	66%	62%	56%	62%	57%	71%	57%	55%	48%	55%	70%
2	84%	76%	68%	58%	77%	69%	61%	69%	69%	70%	62%
3	94%	81%	76%	78%	76%	73%	78%	78%	72%	62%	67%
4	96%	92%	82%	82%	75%	80%	63%	67%	77%	82%	74%
5	99%	95%	92%	79%	77%	90%	89%	89%	87%	88%	76%
6	99%	95%	95%	94%	96%	88%	88%	90%	90%	84%	83%
7	100%	96%	94%	97%	97%	96%	95%	93%	96%	94%	82%
8	100%	98%	99%	95%	91%	99%	98%	97%	97%	86%	95%
9	100%	100%	99%	97%	97%	97%	95%	98%	89%	96%	94%
10	100%	100%	100%	95%	96%	99%	93%	92%	90%	90%	97%
11	100%	100%	100%	100%	99%	99%	97%	92%	99%	99%	94%
12	100%	100%	100%	100%	98%	98%	99%	96%	99%	96%	93%
13	100%	100%	100%	100%	100%	100%	98%	97%	99%	100%	100%
14	100%	100%	100%	100%	100%	98%	99%	100%	100%	100%	100%
15	100%	100%	100%	100%	100%	100%	100%	100%	99%	100%	100%

Note: The table above is a portion of the Win Probability Matrix used by SOM. Chart shows the estimated Win Probability for the home team in the last 10 minutes of a game, when the lead is between -15 and 15 points.

Source Data: http://www.sevenovertimes.com/sql/winprobmatrix.csv

Appendix 3



Source Data: http://www.sevenovertimes.com/sql/spread_analysis.csv

Row Labels	Avg. Home Score	Avg. Away Score	HCA	# of Games
2008	68.66	63.16	5.50	2986
2009	66.77	62.76	4.01	2948
2010	70.33	65.68	4.64	3280
2011	68.32	63.89	4.44	3428
2012	69.20	63.51	5.69	1295
Grand Total	**68.62**	**63.88**	**4.74**	**13,937**

Source Data: http://www.sevenovertimes.com/sql/hca.csv

Appendix 4

Top 15 Players in Win Probability Added for 2011-2012 Season

Rank	Player	Team	Win Probability Added	Avg WPA	# of Plays
1	Mike Moore	Hofstra	8.821	0.022	405
2	Doug McDermott	Creighton	7.815	0.022	352
3	John Jenkins	Vanderbilt	7.503	0.026	288
4	Terrell Stoglin	Maryland	7.494	0.024	311
5	Zack Rosen	Pennsylvania	7.478	0.021	352
6	Jordan Clarkson	Tulsa	7.470	0.023	325
7	Reggie Chamberlain	UMKC	7.439	0.024	308
8	Antwan Carter	Longwood	7.339	0.015	486
9	Tyler Bernardini	Pennsylvania	7.280	0.022	338
10	Velton Jones	Robert Morris	7.222	0.020	361
11	Kevin Olekaibe	Fresno State	6.964	0.024	296
12	Ashton Gibbs	Pittsburgh	6.947	0.020	349
13	Ramone Moore	Temple	6.887	0.025	271
14	Ryan Willen	Lafayette	6.741	0.022	309
15	Jeremiah Bowman	Longwood	6.729	0.018	365

Top 15 Clutch Players in Win Probability Added for 2011-2012 Season

Rank	Player	Team	Win Probability Added	Avg WPA	# of Plays
1	Evan Roquemore	Santa Clara	2.019	0.106	19
2	Kevin Olekaibe	Fresno State	2.325	0.101	23
3	Maurice Jones	Southern California	1.672	0.088	19
4	Andre Jones	Winthrop	1.317	0.082	16
5	Jared Cunningham	Oregon State	1.321	0.078	17
6	Humpty Hitchens	James Madison	2.137	0.076	28
7	Brandon Young	Depaul	1.286	0.076	17
8	Michael Williams	South Florida	1.266	0.074	17
9	Mike Moore	Hofstra	1.630	0.074	22
10	John Jenkins	Vanderbilt	1.477	0.074	20
11	Khalid Mutakabbir	Presbyterian	1.600	0.070	23
12	Kyle Cain	Arizona State	1.315	0.069	19
13	Austin Morgan	Yale	1.291	0.068	19
14	Deonte Burton	Nevada	1.289	0.068	19
15	Drew Ferry	Cornell	1.216	0.068	18

Note: Clutch is defined as average WPA for games within five points in the last five minutes of a game. Players must have 10 plays to qualify. Players are sorted by average WPA.

Note: Data is current for all games played thru January 9th, 2012.