A Spatial Analysis of the NCAA Basketball Tournament

By

Brian S. Ward
CH2M HILL

Brian R. Davenhall
CH2M HILL

Bryce R. Wells
Athlon Sports, Inc.

2006 ESRI International User Conference Proceedings

Abstract

The NCAA Men's Basketball Tournament is one of the most popular events in both athletics and entertainment. Each year, millions of people watch the games, but not only that: they also bet on the games; travel hundreds or thousands of miles to see their teams play; and most importantly - they compete in NCAA Tournament bracket contests. This study takes a look a historical NCAA Tournament results since 1985 (the year the tournament expanded to 64 teams), to determine whether or not there is a spatial correlation between the distance of competing teams to their game sites, based on either their seeding and/or the Vegas lines, relative to their success in Tournament games. The end result will be not only the results of the study, but also a fascinating visualization of the last 20 years of the NCAA tournament.

Introduction: March Madness

Each year around the Ides of March, sports fans across the United States are obsessively captured by the frenzy of the National Collegiate Athletic Administration (NCAA) Men's Basketball Tournament. What was once a niche market shared primarily by the teams, cheerleaders, alumni, and fans of such schools as UCLA, the University of Kentucky, and the University of North Carolina, has over time become a cultural phenomenon. On a world scale, the closest comparison in sports is the World Cup – if on a smaller, more frequent scale than the World Cup. Whether it is the emotionless, anticipated financial gain from beating the Las Vegas odds-makers, the camaraderie of joining a company bracket contest, the eager anticipation of exciting finishes and maddening CBS coverage, or a genuine interest in a particular team's success, millions upon millions of fans spend their waking hours in front of their televisions – or, more recently, tuned in through the immediacy of the Internet inside their cubicles, parting only when nature or Starbucks calls.

The process begins on Sunday evening, after the last major conference tournaments have ended. The NCAA Men's Basketball Tournament Selection Committee, comprised of leadership representing many of the NCAA's athletic conferences, provides their selections to CBS, who then announces the tournament bracket. Databases across the country are immediately populated with the teams, so that all with interest may either print out their bracket or enter a fantasy bracket challenge. During the announcement, television coverage invariably takes us around the nation to observe those "bubble teams" that may or may not be invited. Rapturous excitement or crushing depression is reflected in the young faces of the players, while coaches nervously fumble with their clipboards, knowing that the committee's decision may mean the difference between receiving a pink slip and gaining a six-figure raise.

The tournament's participants have been divided into four regional groupings comprised of sixteen teams each. An individual region's sixteen teams are seeded (i.e., ranked) 1 through 16, with the most accomplished team gaining the 1 seed. The 16 seed is considered fodder for the 1 seed in the first round, as no 1 seed has ever lost on the first Thursday or Friday of the tournament. The first round's other games consecutively match up the next best and next worst teams, progressing from the 2 versus 15 match up to the 8 versus 9 match up. Each round rids the bracket of half of its remaining participants, until the remaining teams from each region convene at the often anti-climatic Final Four. As an example, the 2005 initial tournament bracket can be viewed in Figure 1.

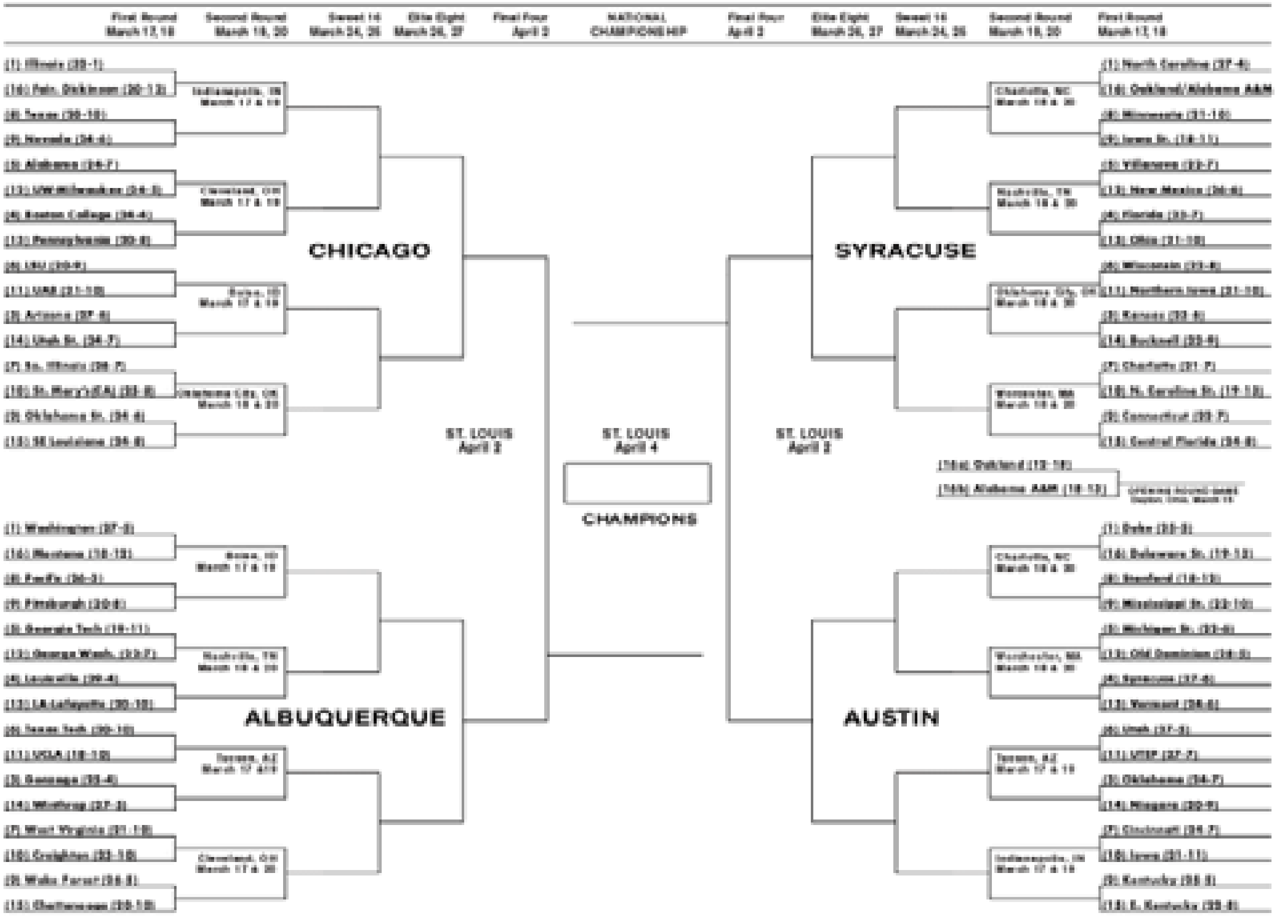

Figure 1 – 2005 NCAA Men's Basketball Tournament Bracket

As the event begins in earnest on Thursday and Friday, the immediate concern within the average fan is less about who will be crowned the champion and more about which 12 seed will upset a 5 seed this year. This highlights the observation that the "madness" surrounding the tournament is far bigger than the games, or the participating schools, or the personalities involved. At its core, the tournament has become about picking the upsets – which are truly what make the first round so magical for its observers. Anticipation and excitement gradually decreases as teams advance past the first and second rounds into the Sweet Sixteen, then the Elite Eight, and finally – the ultimate prize for those fans concerned with team bragging rights – the Final Four. While a Final Four weekend cannot begin to compare with the hyperactive madness of the first round, it contains the primary goal for those participating in the event – the crowning of the NCAA Men's Basketball National Champion.

On the flip side, an improper amount of money changes hands during the event that is the NCAA Tournament. In Las Vegas, bookies and odds-makers make a fine living by establishing betting lines, and raking in the dollars of those willing to participate via their disposable – hopefully – income, in this gambling enterprise. It is important to establish now that this paper in no way attempts to affect the "sport" of gambling on college basketball. Rather, it is inspired solely by curiosity and the overwhelming obsession that its writers are consumed by during this three week event. As such, the results are for informational and entertainment purposes only – in other words, the paper's authors claim and wish neither credit nor blame for one's windfall or losses.

Spatial Factors in Athletics

With so much effort given to anticipating the outcomes of tournament games, it seemed interesting to consider whether or not there were simple geographic factors that might influence these outcomes. Hu and Zidek suggest that home team advantage is significant in predicting the outcome of NBA playoff games and propose separating home and away games. Although NBA playoff games are played on "true home courts" it does seem reasonable that GIS can be used to develop an additional covariate (i.e., home court advantage) for games on neutral courts. It‘s conceivable that this covariate could incorporate A GIS-related component such as Euclidean distance, route distance, elapsed time, or some combination of these data and/or other information.

Straub suggests that jet lag can be severe for teams traveling across several time zones and cites Oren et al. as establishing that the more time zones crossed the greater the jet lag. Based on these conclusions it appears plausible that the "time zone effect" could be a measurable phenomenon. If this is in fact true, then it can be conceived that as the distance to game destination increases so too do the number of changes in time zone resulting in a negative impact (i.e., a greater likelihood of losing) to teams traveling large distances to game destinations.

The above two examples serve to illustrate that spatial factors can influence the outcome of athletic events and set precedence for including spatial data for predictive modeling.

Proposed Analysis

As has been established, there is strong evidence of a correlation between home-court advantage and the outcomes of sporting events. Over the 21 years included in the study, teams participating in the NCAA Men's Basketball Tournament have represented 46 states, with game locations in 38 states. While there is a cursory attempt by the Selection Committee to place teams within logical geographical areas, more often than not teams are sent a significant distance away from their campuses to participate in the games. This paper's primary aim is to study the NCAA Tournament since its expansion in 1985 to 64 teams, and to determine whether or not there exists a predictive correlation between the distance traveled by a team and its likelihood for success.

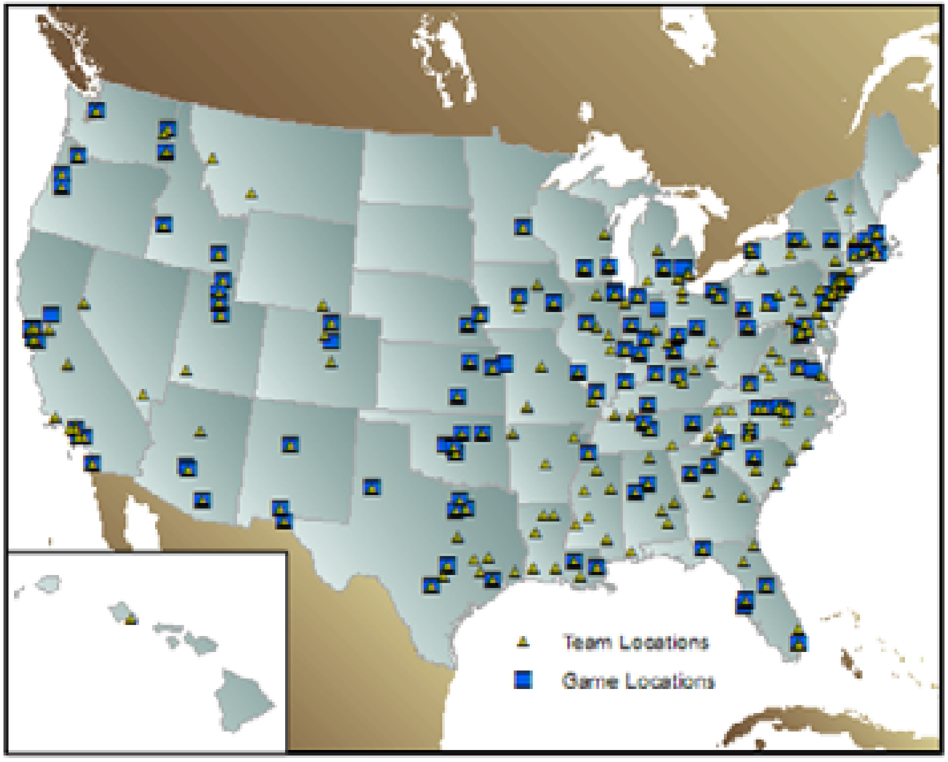

Figure 2 - Relevant Spatial Locations, 1985-2005

Evidence seems to suggest a strong correlation between distance and success, when selective studies are applied and additional – possibly more relevant – information is dismissed. For instance, the University of Hawaii is clearly the most spatially remote team that has ever participated in the Tournament. Three times in the past 22 years the Hawaii Rainbows – now Rainbow Warriors – have flown across the Pacific Ocean to the mainland in order to participate in the NCAA Tournament. The combined one-way distance traveled by these Hawaii teams is 18,138 kilometers (11,246 miles). Conversely, their opponents – twice Syracuse and Xavier – have traveled just 5,004 kilometers (3,102 miles) to play in these games. The results of these games seem to support the idea that distance is related to success, as Hawaii lost all three games by double figures.

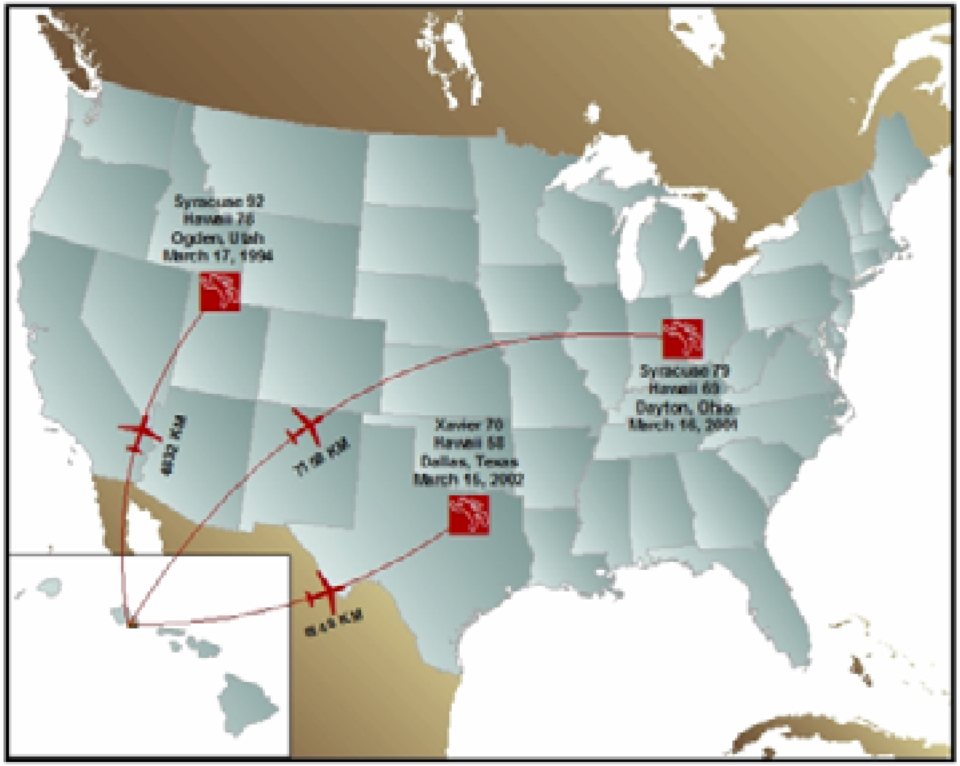

Figure 3 – Anecdotal Evidence, University of Hawaii

Another seemingly relevant example is the 1987 run by the University of Indiana. Though by no means an underdog – the Hoosiers were one of the top teams in the nation, and were granted a 1 seed by the committee – Bobby Knight's team faced an almost unseemly home court advantage in the first four rounds. For the first two rounds, near formalities for a team the caliber of this Indiana group, they merely had to travel up State Road 37 to Indianapolis. Just 75 kilometers (47 miles) away, the bus was on the road for barely an hour. Conversely, their opponents Fairfield and Auburn had to travel a combined 1,894 kilometers (1,174 miles). Indiana had little trouble, scoring almost 200 points on the way to a pair of blowout victories.

The Sweet Sixteen and Elite Eight provided significantly more drama, without a great deal more travel. Again, Indiana traveled just 173 kilometers (107 miles) to the venue in Cincinnati, Ohio, while its opponents – Duke and Louisiana State – combined to journey 1,741 kilometers (1,079 miles). Indiana Sophomore Rick Calloway ensured that Duke would not defeat the Hoosiers, scoring 21 points en route to an 88-82 victory that was more comfortable than the final margin would indicate. The LSU game two days later would not prove as easy, as Dale Brown's group staked a 12 point second half lead, and seemed to have Indiana backed into a corner. The key play down the stretch came from unlikely hero, Joe Hillman, who rarely made it off the bench for the Hoosiers. His "old-fashioned" three-point play, on a pass from Bloomington legend Steve Alford, sparked Indiana to a one point victory, 77-76. Under the often scornful eye of Coach Brown, the Tigers surely had a long trip home to Baton Rouge.

Bobby Knight's group would also travel to bayou country, as the 1987 Final Four would take place at the Superdome in New Orleans. Compared to the first two trips, Indiana had a great deal farther to travel, as New Orleans is over 1000 kilometers (620 miles) south of Bloomington. However, the other three Final Four participants – semifinal opponent Nevada-Las Vegas, Syracuse, and Providence – traveled an average of 2,148 kilometers (1,332 miles) to take part in the annual media frenzy.

In the National Semifinal game between Indiana and UNLV, the aptly named, road-weary Running Rebels of UNLV had no trouble playing at a frantic pace, as the teams combined to score 100 points in the first half of play. Indiana held the lead by six points, and a Kojak-esque Jerry Tarkanian would spend most of the remainder of the game with his signature white towel stuffed in his mouth. UNLV ultimately could not overcome its opponent, and Indiana pulled out a 97-93 win on the strength of Alford's 33 points. In the other semifinal game, Syracuse defeated fellow Big East member Providence, 77-63.

The resulting National Championship game was one of the most memorable games in NCAA Tournament history. Syracuse alumnus Jim Boeheim and his Orangemen matched the Hoosiers stride for stride, and it was apparent early that the game would come down to the final minutes. While Steve Alford was the team's MVP and a consensus All-American, it would be Keith Smart that dealt the championship blow for Indiana. Trailing 73-72 with time running out, teammate Daryl Thomas spotted Smart along the left baseline. Smart caught the ball and, in one smooth motion, hurled it towards the basket as time expired. While the net settled into stillness, the surrounding arena exploded as Indiana secured the National Championship. The state of Indiana would adopt the French Quarter for the evening, while the Hoosiers and their fans reveled in their particular brand of euphoria – as if they had discovered yet another home city on their way to basketball history.

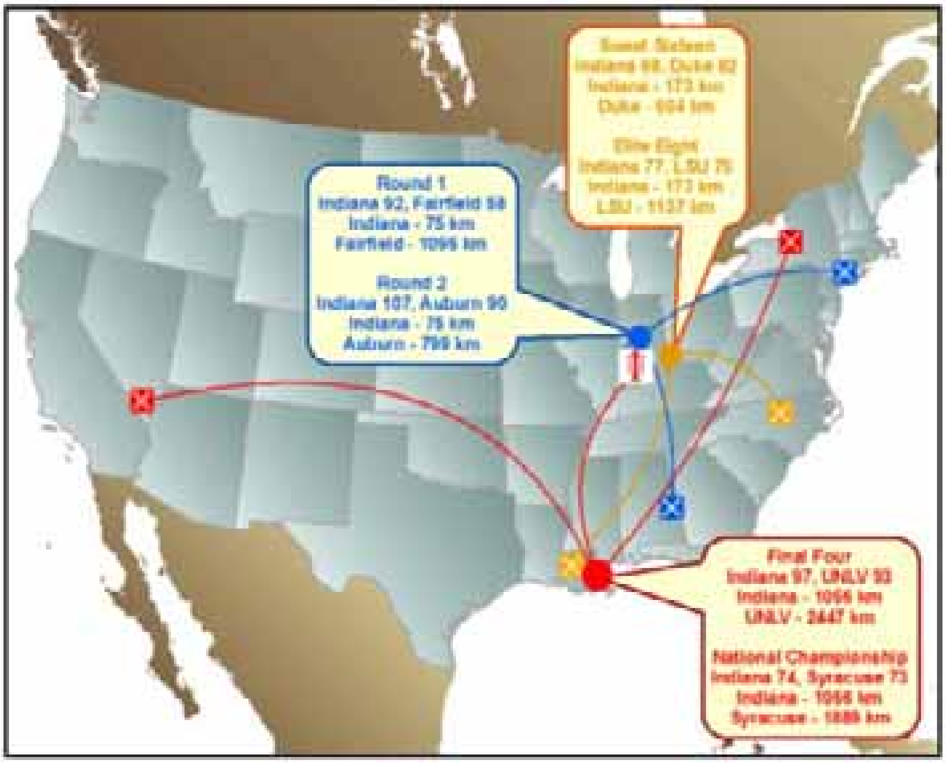

Figure 4 – Anecdotal Evidence, Indiana University

Before this becomes a journalistic tour of nostalgia for Indiana faithful who might be reading, it seems prudent to digress towards the meat of the effort. Anecdotal evidence can be persuasive in many cases, and is used to support any variety of arguments. But in reality, it cannot be relied upon to establish a correlation such as the one being sought. Rather, a statistical methodology is required, and the proposed approach will now be outlined.

The goal of the remainder of this paper is to determine whether the Euclidean distance between home city and destination city is a valid predictor of the outcome of NCAA Tournament games, either independently or in conjunction with other possibly relevant covariates. Will the variable of distance traveled contribute to the predictive capability of the model, and is it also possible that the effect of time zone changes and the idea of home court advantage can be quantified in order to participate in the model?

Methods

Since NCAA Men's Basketball Tournament participants and results are part of historical public record, the decision was made to use an already compiled database can be found at HoopsTournament.net. This spreadsheet-style Access database contains a great deal of relevant information for each game played in the NCAA Tournament between its inception and 2005. Having such data in a central location proved extremely beneficial in preparation for the analysis. The database was normalized, producing additional tables related to the primary "Games" table – such as "Locations" and "Teams". Each of these tables included the city and state of the game location or school.

Using the ESRI Data & Maps compilation, the United States Cities shapefile was used to create an Address Locator based on the city and state fields. The access tables, "Locations" and "Teams", were then opened in ArcMap, and the data was geocoded using the new Address Locator. Most of the entries in each of the tables were matched with 100% confidence, though a few remained that had to be matched interactively. Where there was a comparable city located in the Address Locator – such as Rutherford, New Jersey, serving as a proxy to East Rutherford, New Jersey – the city in close proximity was used. It is the belief of the authors that these situations, generally producing less than 10 kilometers (6 miles) of difference, would be of acceptable spatial accuracy for analysis. In situations where there was no acceptable proxy, Wikipedia.com was used to locate the actual Latitude and Longitude (WGS84) of the city. The Latitude and Longitude were input as records in a new table, and the new table was then imported into a new Personal Geodatabase (PGDB) feature class using the "Add X, Y" tool in ArcMap. Specific instances where Latitude/Longitude or a proxy city were used are outlined in Appendix A.

It was then important to establish the distance between each game location and its participating teams. The "Locations" and "Teams" feature classes were projected from Geographic (WGS 84) coordinates to the USA Contiguous Equidistant Conic (metric) projection. A metric projection was chosen rather than an English system of measurements-based projection somewhat arbitrarily; although if a reason were to be given, it could most accurately be attributed to the authors' preference for using the more logical system of measurements.

Hawth's Analysis Tools were chosen to determine the Euclidean distance between the points. Hawth's tools are accessed as an extension to ArcMap. With the teams and locations feature classes loaded into an ArcMap document, the Analytical tool – Distance Between Points (Between Layers) – produced a comma delimited text file including the distance for each coincidence of all teams to all locations. This produced an almost unmanageable 35,000 unique records. The file was then brought into Access as a new table and a query was written to relate it to the original Games table through a multiple-field join of "Locations" and "Teams".

For this analysis, a parsimonious *a priori* model selection and inference strategy was chosen. This incorporated the authors' basic understanding of the NCAA Tournament, in helping to drive the analysis using logical indicators. This strategy consisted of constructing a set of candidate models that were postulated, prior to the statistical analysis, to correctly model game outcomes. We constructed 10 different models based on available literature from previous studies and our own personal observations, experiences, and assumptions as to what might best predict the outcome of a game. This type of predictive strategy relies heavily on the knowledge of the investigator and reports only on evaluated models. We were not interested in conducting an exploratory *a posteriori* "data dredging" type of statistical analysis that simply employs an iterative process to examine different combinations of covariates to develop a predictive model. It was thought that such an analysis would potentially introduce false indicators, perhaps even leading to an over-fitted model.

We included the following covariates in our analysis: tournament seed (seed), RPI Rank (rpi), and Euclidean distance (distance) to the game destination. We created three derived covariates by calculating the differences between competing teams for tournament seed (seed_diff), RPI Rank (rpi_diff), and distance (distance_diff). For each model we performed a logistic regression analysis, in the R program for statistical analysis, using a generalized linear model (GLM) procedure with the logit link function. We evaluated the Akaike Information Criteria (*AIC*) statistic and ranked competing models using this statistic to determine the best-fitting model. Models that reported the lowest *AIC* values received the highest rankings. We used a sample size n=2,656. The models we evaluated and their results are presented in Table 1.

Results

The model that included distance_diff, seed_diff, and rpi_diff reported an *AIC*=2903.565 and received the highest ranking. The model that included only the difference in seed between competing teams (model 6) reported an *AIC*=2911.157. This model ranked fourth and reported an AIC value very close to the highest ranking model. The models that included only distance or distance_diff reported *AIC* values equal to 3675.582 and 3660.256, respectively and received the lowest rankings of all models. The full model, including all six covariates, received the third highest ranking.

Model	Covariates	*AIC*	*Rank*
1	distance	3675.582	10
2	seed	3171.378	5
3	rpi	3518.293	8
4	distance_diff	3660.256	9
5	seed_diff	2911.157	4
6	rpi_diff	3251.220	7
7	seed_diff, rpi_diff	2906.064	2
8	distance_diff, rpi_diff	3242.727	6
9	distance_diff, seed_diff, rpi_diff	2903.565	1
10	FULL MODEL	2909.565	3

Table 1 – Akaike Information Criteria (AIC) model results

Discussion

The results of our analysis will likely not become required study for bettors – or even the casual bracketologist. It is rapidly clear by looking at the results that RPI and seed are much more effective means of predicting game outcomes than distance. The derived covariates used by looking at each of the differences in RPI, seed, and distance were in each case more effective than those values alone – though, the greatest improvement and most impressive results again were provided by the differences in RPI and seed. Furthermore, the results suggest that the difference in seed between competing teams is the best predictor covariate, of those evaluated here. The combined difference in RPI and seed model (model 7), which received the second highest ranking, reported only a marginally better AIC value. Though only marginally, adding the difference of distance to model 7 (model 9) did improve the AIC value. The outcome was not completely unexpected since tournament seed and RPI are closely related. In fact, the Selection Committee uses RPI as a major part of their criteria for developing team seeds, and subsequently team pairings for tournament games.

There is significant evidence through other empirical studies that home court advantage can be indicative, generally, of game outcome – when viewed independently of any other factors. The authors believe that this trend could hold true for NCAA Tournament games as well, that independent of other factors a team that's playing very close to home would have an advantage over a team that has traveled a great distance. However, based on the number of samples available, it is not possible to prove this theory through anything resembling sound statistical evidence.

As with common score and outcome predicting models that weigh the home court advantage, it is quite likely that a reasonable predictive model could be created for the NCAA Tournament that included a distance factor. This factor should build along the lines of the difference in distance covariate used in this study, and should be weighted more heavily when one team has a very short distance to travel for its game – close enough to effectively provide a home court advantage. However, the model used in this study as it appears would not provide such an enhancement.

Future studies should evaluate both frequentist and Bayesian methods for statistical analysis and inference. Modeling procedures could also include exploratory methods, although this is not what the authors of this paper are proposing in this paper, such as stepwise-regression or best subsets. Additional statistics could be evaluated, as well as testing for goodness-of-fit for high ranking models.

Additionally, there are several other potential covariates of interest that might be researched in the near future. Among the spatial (albeit some only loosely spatial) examples are: travel across time zones, particularly those match ups where one team has crossed two or more time zones while its opponent is in its home time zone; and, cumulative travel, for teams appearing in the Sweet Sixteen, Elite Eight, and Final Four.

Finally, regardless of the authors' further investigations into spatial indicators of NCAA Tournament success, they will undoubtedly be tuned in next March – as they are every year – trying to figure out which 12 seed will upset a 5 seed this year.

Acknowledgements

Mr. Ward would first and foremost like to thank Barry Goheen, for alerting him firsthand to the magic of “March Madness”, on March 21, 1988; Jon Shannon “McCopacetic” McElvaney for his diligent support and championship throughout the process; and Dad, for meeting me in Phoenix a couple of years ago to watch our Commodores become fodder on Connecticut’s long-distance National Championship run.

Mr. Davenhall would like to acknowledge the countless teams that tirelessly compete each year in hopes of reaching the tournament and claiming the coveted Championship; the supporters and fans of this event that make it so exciting; and the family and friends that share in this “March Madness” each year.

Mr. Wells would like to thank Billy McCaffery, the only good thing to come out of Duke University, for leading the 1993 Commodores to the Sweet 16 and enlightening him to the thrills and the spectacle that is the NCAA Tournament. And to his wife Tricia who was the only person crazy enough to travel with him to Oxford, MS, to see Matt Freije break the all-time scoring record for Vanderbilt University. And finally to his son Greyson, for understanding that his dad doesn't have, but wishes he did, the requisite basketball genetic makeup to pass down to him.

Appendix A: Spatial adjustments of location and team data

Teams table, proxy cities used (team; actual city, state; proxy city, state):

- Alcorn State; Lorman, MS; Hermanville, MS
- Boston College; Chestnut Hill, MA; Boston, MA
- Bucknell; Lewisburg, PA; Sunbury, PA
- Cal State-Northridge; Northridge, CA; San Fernando, CA
- Coastal Carolina; Conway, SC; Myrtle Beach, SC
- Fordham; Bronx, NY; New York, NY
- Rhode Island; Kingston, RI; Newport, RI
- St. John’s; Jamaica, NY; New York, NY
- Wagner; Staten Island, NY; New York, NY
- Niagara; Niagara, NY; Niagara Falls, NY
- Oakland; Rochester, MI; Detroit, MI

Team table, latitude/longitude used (team; city, state) from Wikipedia.com:

- Alcorn State; Lorman (Hermanville), MS
- Campbell; Buies Creek, NC
- Colgate; Hamilton, NY
- Dartmouth; Hanover, NH
- Eastern Washington; Cheney, WA
- Fairfield; Fairfield, CT
- Lebanon Valley; Annville, PA
- Manhattan; Riverdale, NY
- Mississippi; Oxford, MS
- Mississippi Valley; Itta Bena, MS
- Monmouth; West Long Branch, NJ
- Morehead State; Morehead, KY
- Mt. St. Mary’s; Emmittsburg, MD
- Niagara; Niagara (Niagara Falls), NY
- Oakland; Rochester (Detroit), MI
- Pepperdine; Malibu, CA
- Prairie View; Prairie View, TX
- Rider; Lawrenceville, NJ
- St. Francis, PA; Loretto, PA
- St. Mary’s; Moraga, CA
- Virginia Military; Lexington, VA
- Western Carolina; Cullowhee, NC

Locations table, proxy cities used (actual city, state; proxy city, state):

- East Rutherford, NJ; Rutherford, NJ
- Jamaica, NY; New York, NY
- Kingston, RI; Newport, RI
- Landover, MD; Washington, DC
- Notre Dame, IN; South Bend, IN
- Rosemont, IL; Chicago, IL
- Williamsburg, PA; Altoona, PA

Bibliography

Ben-Naim, E. Vazquez, F., and Redner, S. "Parity and Predictability of Competitions: Nonlinear Dynamics of Sports." Dynamics Days Asia Pacific 4 Proceedings 2006.

Benson, J. HoopsTournament.net. 21 March 2006. <http://www.hoopstournament.net>

Beyer, H. Hawth's Analysis Tools for ArcGIS. 30 March 2006. <http://www.spatialecology.com/htools/overview.php>

Carlin, B.P. "Improved NCAA Basketball Tournament Modeling Via Point Spread and Team Strength Information." *The American Statistician* 1996: 50, 39-43.

Graham, A. "Resurrected Hoosiers get new life". *The Bloomington (Indiana) Herald-Telephone* 23 March 1987.

Graham, A. "End to come Monday, for IU, but it could be a happy one". *The Bloomington (Indiana) Herald-Telephone* 29 March 1987.

Hammel, B. "Hoosiers get away fast". *The Bloomington (Indiana) Herald-Telephone* 13 March 1987*.*

Hammel, B. "Hoosiers turn up the volume". *The Bloomington (Indiana) Herald-Telephone* 15 March 1987*.*

Hammel, B. "Calloway comes down to earth, brings Duke with him". *The Bloomington (Indiana) Herald-Telephone* 21 March 1987.

Hammel, B. "Hoosiers stop Rebels, gain NCAA final again". *The Bloomington (Indiana) Herald-Telephone* 29 March 1987.

Hammel, B. "IU wins NCAA crown". *The Bloomington (Indiana) Herald-Telephone* 31 March 1987.

Hu, F. and Zidek, J.V. "Forecasting NBA Basketball Playoff Outcomes Using the Weighted Likelihood." Technical Report, University of British Columbia 2003.

Kvam, P. and Sokol, J.S. "A Logistic Regression/Markov Chain Model for NCAA Basketball." *Naval Research Logistics* 2005.

Lerra, M.A. "The Role of Familiarity in the Home Advantage." Thesis, Wesleyan University 2003.

Oden, D.A., Reich, W., Rosenthal, N.E., & Wehr, T.A. *How to Beat Jet Lag: A Practical Guide for Air Travelers*. New York: Henry Holt 1993.

Reid, M.B. "Least Squares Model for Predicting College Football Scores." Thesis, University of Utah 2003.

Schwertman, N.C., Schenk, L., and Holbrook, B.C. "More Probability

Models for the NCAA Regional Basketball Tournaments." *The American Statistician* 1996: v50, p34-38.

Straub, W.F. "The Effects of Diaphragmatic Breathing and Sleep Training on Sleep, Jet Lag, and Swimming Performance." *The Sport Journal* 2003: v6, n1.

Wikipedia. 30 March 2006. <http://www.wikipedia.com>

Yang, T.Y. and Swartz, T. "A Two-Stage Bayesian Model for Predicting Winners in Major League Baseball." *Journal of Data Science* 2004: v2, p61-73.

Contact Information

Mr. Brian S. Ward
CH2M HILL
19 S. Tejon Street, Suite 500
Colorado Springs, CO 80903
US
719-477-4917
brian.ward@ch2m.com

Mr. Brian R. Davenhall
CH2M Hill
2525 Airpark Drive
Redding, CA 96001
US
530-229-3253
brian.davenhall@ch2m.com

Mr. Bryce R. Wells
Athlon Sports, Inc.
2011 18th Avenue South
Nashville, TN 37212
US
615-293-4011
bryce_wells@yahoo.com