Learning Stochastic Models for Basketball Substitutions from Play-by-Play Data

Harish S. Bhat, Li-Hsuan Huang, and Sebastian Rodriguez

Applied Mathematics Unit, University of California, Merced,
Merced, United States
{hbhat,lhuang33,srodriguez48}@ucmerced.edu

Abstract. Using play-by-play data from all 2014-15 regular season NBA games, we build a generative model that accounts for substitutions of one lineup by another together with the plus/minus rate of each lineup. The substitution model consists of a continuous-time Markov chain with transition rates inferred from data. We compare different linear and non-linear regression techniques for constructing the lineup plus/minus rate model. We use our model to simulate the NBA playoffs; the test error rate computed in this way is 20%, meaning that we correctly predict the winners of 12 of the 15 playoff series. Finally, we outline several ways in which the model can be improved.

Keywords: Substitutions, plus/minus, minutes, continuous Markov chain

1 Introduction

If one watches a basketball game played in the NBA, one cannot help but notice that players are substituted in and out with regularity. A large difference between basketball and other sports such as football/soccer and baseball is that a player who is substituted out of the game can return to play at a later time. Substitutions can substantially alter the strategy employed by the 5-man unit on the court. Many teams field, at different times of the game, different lineups in order to change the emphasis placed on aspects such as (but not limited to) rebounding, pace and fast break opportunities, or long-range shooting.

In short, an NBA team is actually a collection of different 5-man units. On average, in the 2014-15 regular season, teams used 15.1 different 5-man units per game. In this work, we use play-by-play data to build stochastic models for the dynamics of these 5-man units. Combining this model of substitutions with scoring models for each 5-man unit, we obtain generative models that can be used to simulate games. The ultimate goal of these models is to answer questions such as: in a 7-game series between two teams, what is the probability each team will win? Motivated by this goal, in the present work, we seek baseline continuous-time stochastic models that can be used as a starting point for further modeling efforts.

Our work builds on different strands of the literature. Discrete-time Markov chain models of basketball have been considered in [11], for instance. One particularly successful model uses a discrete-time Markov chain to rank NCAA basketball teams [5]. Classification methods from machine learning have been applied to basketball "box score"-type data to make daily predictions of the winners of college basketball games [10]. Continuous-time stochastic models have been considered by [9], though in these models the lineup of players on the court is ignored. Finally, very recent work models the spatial location of all players on the court during the game [1], with possessions modeled as a semi-Markov process. Perhaps the work closest to ours is [6], which develops a probabilistic graphical model to simulate matches, including changes to team lineups. One of the main conclusions of [6] is that the outcome of individual games and series are sensitive to changes in the team lineup. Our work uses this as a starting point for modeling.

2 Data Collection

Although sources such as NBA.com and ESPN provide statistics of teams and individual players, we found it difficult to obtain statistics on the performance of 5-man units or lineups. To obtain this information, we mined 2014-15 regular season NBA play-by-play data from `knbr.stats.com`, supplemented by data on substitutions taking place between quarters from `Basketball-Reference.com`. For each play-by-play HTML page, we used `Beautiful Soup`, a Python package, to scrape the information needed from the text descriptions of particular plays.

To give an example of how the data appears after processing, we present Table 1. Each row of this data set corresponds to a group of 10 players who played on the court for a positive amount of time before at least one substitution was made by either team. Columns 1-3 record the date of the game and the identities of the home and visiting teams. Columns 4-8 record the identities of the five players on the court for the home team, while columns 9-13 record the same information for the visiting team. Column 14 contains the number of seconds this group of 10 players (5 from each team) played just before one substitution was made by either team. Columns 15-17 record the number of play-by-play events that have occurred for the home, visiting, and both teams since the last substitution. Columns 18-19 record the home and visiting scores at the time just before the substitution was made.

Column 20, the last column, records the change in point differential. Let the current home and visiting scores (recorded in columns 18-19) be H_i and V_i, respectively. Then the change in point differential Δ_i is

$$\Delta_i = (H_i - V_i) - (H_{i-1} - V_{i-1}), \tag{1}$$

with the understanding that the initial scores are $H_0 = V_0 = 0$. This quantity is the "plus/minus" of the two 5-man units on the court. If we start the first row of Table 1, we see that at the time the first substitution is made, $\Delta_1 = -2$, corresponding to home and visiting scores of 15 and 17, respectively. At the time

1 (Date)	2 (Home Team)	3 (Visiting Team)	4 (Home Player 1)	5 (Home Player 2)	6 (Home Player 3)	7 (Home Player 4)	8 (Home Player 5)	9 (Visiting Player 1)	10 (Visiting Player 2)	11 (Visiting Player 3)	12 (Visiting Player 4)	13 (Visiting Player 5)	14 (Seconds Played)	15 (Home Events)	16 (Visiting Events)	17 (Total Events)	18 (Home Score)	19 (Visiting Score)	20 (Δ_i—see Eq. (1))
20150127	Mia	Mil	478	479	480	487	481	57	426	425	431	427	350	13	21	34	15	17	-2
20150127	Mia	Mil	479	480	487	481	484	57	426	425	431	427	149	8	27	14	20	22	0
20150127	Mia	Mil	480	487	484	485	478	57	426	425	431	427	124	7	32	12	22	24	0
20150127	Mia	Mil	487	484	485	478	185	57	425	427	430	429	97	14	6	13	29	30	1
20150127	Mia	Mil	478	484	485	185	483	425	429	430	428	432	73	4	4	8	29	30	0

Table 1. Sample rows of data frame produced by scraping play-by-play data.

the next substitution is made, the score is 20 to 22 in favor of the visiting team. Because the differential is still -2, the *change* in differential is zero, i.e., $\Delta_2 = 0$.

Viewed from the point of view of the home 5-man unit, this means that even though the unit scored 5 points on offense, the unit yielded 5 points on defense. We see that the Δ_i value encapsulates both the offensive and defensive performance of a particular 5-man unit. It is better to score only 3 points on offense and yield 0 points on defense than it is to score 20 points on offense while yielding 25.

3 Substitution Models

Our model consists of two parts: (i) a model for substituting one 5-man unit by another, and (ii) a model for how each 5-man unit contributes to the overall score of the game. In this section, we begin by describing a continuous-time Markov chain model for substitutions. We construct one Markov chain for each of the 30 teams in the NBA; let M_i denote the transition rate matrix of the Markov chain for team i. The Markov chain for team i is completely specified by M_i. Each state of M_i is a different 5-man unit that appears in the training data for team i. Let N_i be the number of states for M_i; using the entire 2014-15 regular season as training data, we obtain the following counts: For each i, we infer the $N_i \times N_i$ transition rate matrix M_i using the MLE (maximum likelihood estimate) [2, 7]:

$$\widehat{M}_i^{j,k} = \frac{\#(j \to k)}{\alpha(j)}. \tag{2}$$

Here $\widehat{M}_i^{j,k}$ is the estimate for the (j,k)-th entry of M_i, $\#(j \to k)$ denotes the number of observations of a transition from state j to state k, and $\alpha(j)$ denotes the total time spent in state j. All of these values can be computed using the play-by-play data.

1 (Atl)	2 (Bkn)	3 (Bos)	4 (Cha)	5 (Chi)	6 (Cle)	7 (Dal)	8 (Den)	9 (Det)	10 (GS)	11 (Hou)	12 (Ind)	13 (LAC)	14 (LAL)	15 (Mem)
401	669	309	516	372	426	387	446	523	466	518	841	504	504	481

16 (Mia)	17 (Mil)	18 (Min)	19 (NO)	20 (NY)	21 (OKC)	22 (Orl)	23 (Phi)	24 (Pho)	25 (Por)	26 (SA)	27 (Sac)	28 (Tor)	29 (Uta)	30 (Was)
415	404	472	702	362	420	358	586	563	807	574	470	541	265	417

Table 2. For each of the 30 NBA teams, we record the total number of 5-man units used by the team during the 2014-15 regular season. In our Markov chain model, this is the number of states N_i for each team $i \in \{1, 2, \ldots, 30\}$.

To validate this model's performance on the training set of all regular season 2014-15 NBA games, we simulate 8200 games for each team. We count the number of substitutions made by each team and divide by 100 to obtain a Monte Carlo estimate for the total number of substitutions made by each team in one full season of play. The simulation follows standard algorithms for sampling from a continuous-time Markov chain. Assume the system is currently in state j. We then simulate exponentially distributed random variables with rates given by row j of the transition rate matrix. The minimum of these samples gives us both the time spent in state j as well as the identity of the new state k to which we transition. We initialize the simulation using the most common 5-man unit for each team, and we terminate the simulation once it reaches 2880 seconds, corresponding to a regulation-length NBA game.

Using results for 29 of the 30 teams, the correlation between the simulated and true number of substitutions is 0.8634. For one team, the Boston Celtics, simulations predict 4627.57 substitutions in one season, while the true number is 1792. This is one indication that there are surely far better distributions than the exponential to model the time spent in one state before transitioning. We discuss ongoing work in this direction in Section 5.

In Fig. 1, we plot the true and simulated times played by each of the 5-man units across all 30 teams (left panel, Pearson correlation of 0.834), and the true and simulated times played by each of the 492 NBA players (right panel, Pearson correlation of 0.915). All times are in minutes. The data from the last panel has been plotted on log-scaled axes; the reported correlation is for the raw data.

Overall, the in-sample fit between true and simulated unit and player times indicate that our model is a reasonable starting point to account for substitutions and 5-man unit playing team. Clearly, further research is necessary to improve the fit and develop a more predictive model of 5-man unit time. An obvious area for improvement is to model the number of fouls committed by each player on a team. Because a player must leave the game immediately after committing a sixth foul, a player is more likely to be substituted out of the game as he

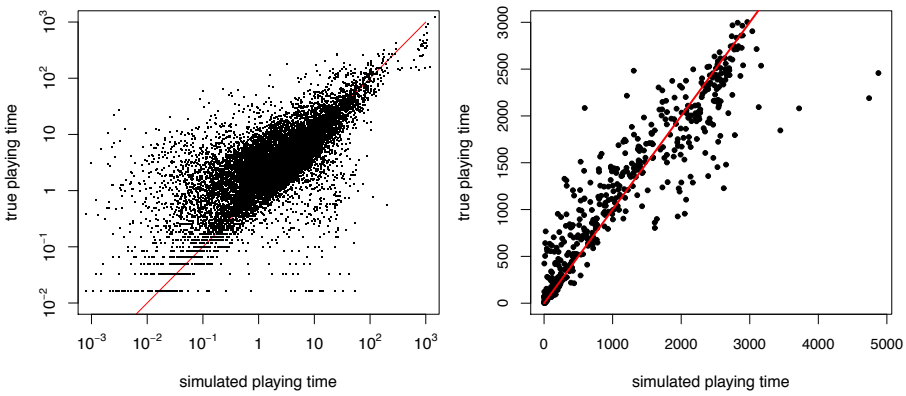

Fig. 1. We plot true and simulated times played by each 5-man unit (left) and each player (right). For both plots, we have plotted the line $y = x$ in red; deviations from this line constitute model error. Simulations are carried out using a continuous-time Markov chain model for substitutions inferred from play-by-play data. Note that the left plot has log-scaled axes.

accumulates more fouls. Another idea is to allow the Markov transition rates to depend on how many minutes remain in the game and the game score; towards the end of blowout games, where one team leads another by a large margin, we see teams rest their regular players in favor of bench players.

In what follows, we will show that the model developed here, despite its deficiencies and though it ignores which team actually won each regular-season game, is capable of prediction.

4 Scoring Models and Results

The second part of our model considers the change in point differential (or plus/minus) rate for each 5-man unit. We refer to this as our scoring model, even though the concept of point differential incorporates both offensive and defensive performance, as described in Section 2.

4.1 Results for the 2014-15 NBA Regular Season

When simulating the continuous-time Markov chain substitution model, if the system spends τ units of time in state i, we multiply τ by the scoring rate associated with this state. This yields a change in point differential for a particular segment of game time. Summing these point differential changes across a 48-minute game, we obtain an aggregate point differential. Again, we initialize the system in the state corresponding to the lineup most often used by the team. To

simulate a game between two teams, we simulate each team's aggregate point differential separately; the team with the larger value is then declared the winner.

In the most basic scoring model, we assign to each 5-man unit an average scoring rate. That is, across the entire training set, we sum the change in point differentials for a particular 5-man unit and divide by the total time this 5-man unit spent on the court. Using this scoring rate, we simulate each of the 1230 regular season games 100 times and average the results for each game. We produce from this simulation three confusion matrices corresponding to true and predicted winners (H = home, V = visiting):

$$\begin{array}{c} \\ H \\ V \end{array}\!\!\begin{array}{c} \begin{array}{cc} H & V \end{array} \\ \begin{bmatrix} 506 & 202 \\ 200 & 322 \end{bmatrix} \end{array} \qquad \begin{array}{c} \\ H \\ V \end{array}\!\!\begin{array}{c} \begin{array}{cc} H & V \end{array} \\ \begin{bmatrix} 329 & 94 \\ 152 & 280 \end{bmatrix} \end{array} \qquad \begin{array}{c} \\ H \\ V \end{array}\!\!\begin{array}{c} \begin{array}{cc} H & V \end{array} \\ \begin{bmatrix} 220 & 54 \\ 90 & 186 \end{bmatrix} \end{array}$$

Rows correspond to predictions while columns correspond to the truth. From left to right, we show results on all games (overall accuracy of 0.67), games in which the predicted margin was ≥ 5 points (overall accuracy of 0.71), and games in which the predicted margin was ≥ 10 points (overall accuracy of 0.73).

4.2 Results for the 2014-15 NBA Playoffs

Because we used regular-season data to train the model, we must consider the above results to be training set results. To develop test set results, we consider the recently concluded NBA playoffs. For each best-of-7 playoff series, we predict the winner, the expected margin of victory, and the probability of victory. Note that the margin here is in terms of the game score, i.e., if one team sweeps another, the margin is 4, whereas if the series goes to a seventh game, the margin will necessarily be 1. We present our predictions on the left and the truth on the right: Overall, our model correctly predicts 11 out of the 15 playoff series winners. Two of the erroneous predictions were made on series that were decided in a seventh and final game.

Ridge Regression. The next scoring model we present is built using ridge regression [4]. Each NBA team plays 82 games in a regular season. For team i, consider the $82 \times N_i$ matrix that indicates the number of seconds each 5-man unit played in each game. Let this matrix be X, and let $\boldsymbol{y}$ be the 82×1 vector giving the margin of victory or defeat for each game. The rough idea is to find $\boldsymbol{\beta}$ such that $X\boldsymbol{\beta} = \boldsymbol{y}$. In this case, $\boldsymbol{\beta}$ will contain a plus/minus rate for each 5-man unit.

There are two caveats. First, because $N_i > 82$ for all i, the linear system is underdetermined. We choose ridge regression over LASSO for this problem because we would like to determine a nonzero plus/minus rate for as many 5-man units as possible. If this rate happens to be close to zero, then that is acceptable, but we see no reason to promote sparsity as in LASSO. The second caveat is that while the usual ridge regression penalty is $\|\boldsymbol{\beta}\|_2^2$, in our case, following this procedure yields worse results than the average scoring rate model described

Series	Winner	Margin	Probability	Winner	Margin
NO at GS	GS	1.43	0.75	GS	4
Dal at Hou	Hou	0.08	0.50	Hou	3
SA at LAC	SA	0.24	0.51	LAC	1
Mem at Por	Por	0.39	0.58	Mem	3
Mem at GS	GS	1.07	0.64	GS	2
LAC at Hou	LAC	0.72	0.64	Hou	1
Hou at GS	GS	1.63	0.77	GS	3
Bkn at Atl	Atl	2.05	0.82	Atl	2
Bos at Cle	Cle	2.38	0.86	Cle	4
Mil at Chi	Chi	0.92	0.66	Chi	2
Was at Tor	Tor	1.04	0.68	Was	4
Was at Atl	Atl	1.75	0.81	Atl	2
Chi at Cle	Cle	0.91	0.67	Cle	2
Cle at Atl	Cle	0.32	0.55	Cle	4
Cle at GS	GS	0.32	0.58	GS	2

Table 3. Predictions (left, with non-integer values of margin) and ground truth (right) for 15 NBA playoff series. The above results are test set results using the continuous-time Markov chain substitution model and the simple average scoring rate model. The model correctly predicts 11/15 of the winners.

above. Therefore, we change the penalty to $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2^2$, where $\boldsymbol{\beta}_0$ is the vector of average scoring rates used in the earlier scoring model. We can implement this easily by considering $\boldsymbol{\beta} = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1$. Then the ridge objective function is:

$$J_\lambda(\boldsymbol{\beta_1}) = \| \underbrace{(\boldsymbol{y} - X\boldsymbol{\beta}_0)}_{\boldsymbol{y}'} - X\boldsymbol{\beta}_1\|_2^2 + \frac{\lambda}{2}\|\boldsymbol{\beta}_1\|_2^2.$$

Passing $\boldsymbol{y}'$ and X to a ridge regression solver then yields, for a fixed value of λ, a minimizer $\boldsymbol{\beta}_1$. We use 10-fold cross-validation on the training set to determine an optimal value of λ; we then rerun the ridge regression on the entire training set using this optimal λ. This yields $\boldsymbol{\beta}_1$, which we add to $\boldsymbol{\beta}_0$ to obtain the scoring rate model. Of course, this procedure is repeated for each team.

Using ridge regression, we improve our training set performance, as displayed in the following confusion matrix: $\begin{bmatrix} 509 & 194 \\ 197 & 330 \end{bmatrix}$. The overall accuracy is now 0.682. We also see a slight improvement in test set performance as display in the left-most table in Table 4, as we are now correctly predicting 12/15 or 80% of the playoff winners. Among the models developed in this paper, the ridge regression model is the best. Again, two of the incorrect predictions are for series that were decided in seven games.

Support Vector Regression. The next scoring model we consider is support vector regression (SVR) with a radial basis function kernel. For team i, we extract from the training data all rows and columns corresponding to 5-man units from team

Winner	Margin	Prob.	Winner	Margin	Prob.	Winner	Margin	Prob.
GS	1.74	0.78	GS	2.50	0.90	GS	3.40	1.00
Hou	0.44	0.57	Hou	3.20	0.90	Hou	2.60	0.90
SA	0.42	0.54	LAC	0.80	0.90	LAC	1.40	0.90
Por	0.29	0.56	Por	1.70	0.90	Por	1.00	0.70
GS	0.32	0.53	Mem	0.30	0.90	GS	0.80	0.50
Hou	0.01	0.53	Hou	2.10	0.90	Hou	1.50	0.70
GS	0.88	0.63	Hou	0.30	0.90	Hou	0.20	0.60
Atl	2.15	0.82	Bkn	2.50	0.90	Bkn	2.40	1.00
Cle	2.07	0.88	Cle	0.80	0.90	Cle	2.50	0.80
Chi	1.11	0.71	Mil	0.80	0.90	Mil	1.50	0.70
Tor	0.88	0.64	Was	1.80	0.90	Was	0.30	0.50
Atl	1.36	0.72	Was	3.20	0.90	Was	0.30	0.60
Cle	1.04	0.70	Chi	2.90	0.90	Cle	2.00	0.80
Cle	0.31	0.54	Atl	1.90	0.90	Cle	0.10	0.50
GS	0.16	0.51	GS	2.70	0.90	GS	0.10	0.50

Table 4. Test set results for ridge regression (left, 80% accuracy), support vector regression (center, 46% accuracy), and k-nearest neighbor regression (right, 66% accuracy). Note that the ridge regression scoring rate model results in a correct prediction for 12 out of the 15 playoff series; this is the best model considered in this paper. For the order of the playoff series and true winners, please see Table 3.

i. This yields, for each team, a training matrix with approximately 1500-2500 rows and exactly N_i columns. We fit one SVR model to each training matrix. Then, when simulating a game, we use this SVR model to predict the change in point differential generated by a particular 5-man unit over a particular stretch of time.

Test set results for the SVR model are given in the central table in Table 4. Because this model is more computationally intensive than the prior models, we simulated each NBA playoff series 10 times rather than 100 times. Overall, we see that only 7/15 or 46% of series winners have been predicted correctly.

Nearest Neighbor Regression. The final scoring model we consider is a k-nearest neighbor regression model with $k = 3$. We train this model on the same set of matrices used to train the SVR model. Playoff predictions are given in the rightmost table in Table 4. In situations where both teams won 5 of the 10 simulated series, we chose the team whose expected margin was positive. Overall, we see that 10/15 or 66% of series winners have been predicted correctly.

4.3 Additional Model Evaluation and Usage

To assess whether our test set prediction accuracy is meaningful, we have built three "box score" models. These models select—as a playoff series winner—the team that has (i) scored the most points in the regular season, (ii) achieved the best regular season winning percentage, and (iii) achieved the highest playoff

seeding. Respectively, these models correctly predict 8/15, 11/15, and 12/15 of the playoff series' winners. Of course, our model is more complex than these box score models; naturally, we should expect our model to be capable of answering more complex questions than a box score model is capable of answering.

Our model is particularly well suited to answer "What if?" questions involving player/lineup usage. For example, the model can be used to assess the impact of a player being injured. Atlanta's Kyle Korver, one of the best three-point shooters in the NBA, was injured and did not play after the first two games of the playoff series against the Cleveland Cavaliers. From the next to the last row of Table 4, we see that the continuous-time Markov chain with ridge regression scoring rate predicts that Cleveland should win the series against Atlanta with a probability of 0.54 and a margin of less than one game (specifically, 0.31 games). These results assume that the usage of players mirrors that of the regular season, i.e., that Korver is healthy and able to play. As a test, we have removed from the Atlanta Hawks' transition matrix any 5-man lineup that involves Korver. Rerunning the simulation, we now find that Cleveland should win the series with a probability of 0.79 and a margin of almost 2 games (specifically, 1.72 games). This is closer to the real result, a 4-game series sweep by Cleveland.

While we have simulated the effect of a player not being to play at all, we note that we can also simulate more subtle scenarios such as (i) a player only being able to play a limited number of minutes per game, or (ii) a coach making a conscious decision to use particular lineups more often against a given opponent.

We view our model as a modular component to be incorporated into (rather than to replace) models that involve traditional predictors such as those used in the box score models above. Our best model uses ridge regression to infer the scoring rate for each 5-man unit, but completely ignores informative data such as who actually won each regular season game. In future work, we seek to use this information to generate improved predictions for the outcomes of games.

5 Conclusion

Given the simplicity of the model employed, our results are encouraging. There are several clear directions in which the model can be generalized and improved. First, at the moment, we are using a basic frequentist procedure to infer the transition rates of the continuous-time Markov chain. In ongoing work, we seek to compare this procedure against more sophisticated techniques such as variational Bayes and particle-based Monte Carlo inference [8, 3]. Second, the continuous-time Markov chain assumes that the holding time in each state has an exponential distribution. We seek to generalize this to a distribution that more accurately models the data; this will yield a semi-Markov process as in [1]. While we have tested nonlinear regression models such as SVR, we have not conducted extensive cross-validation studies to find more optimal values of parameters for these models. For these nonlinear models, it may be beneficial to consider several years worth of training data. Finally, we expect that our scoring model can be improved by incorporating the effect of the opposing 5-man unit on the court.

References

1. Cervone, D., D'Amour, A., Bornn, L., Goldsberry, K.: A multiresolution stochastic process model for predicting basketball possession outcomes (2015), arXiv:1408.0777
2. Guttorp, P.: Stochastic Modeling of Scientific Data. Chapman & Hall/CRC (1995)
3. Hajiaghayi, M., Kirkpatrick, B., Wang, L., Bouchard-Côté, A.: Efficient continuous-time Markov chain estimation. In: Proceedings of ICML 2014. pp. 638–646 (2014)
4. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer, 2 edn. (2009)
5. Kvam, P., Sokol, J.S.: A logistic regression/Markov chain model for NCAA basketball. Naval Research Logistics 53, 788–803 (2006)
6. M-H. Oh, S. Keshri, G.I.: Graphical model for basketball match simulation. In: MIT Sloan Sports Analytics Conference (2015)
7. Metzner, P., Dittmer, E., Jahnke, T., Schütte, C.: Generator estimation of Markov jump processes. Journal of Computational Physics 227, 353–375 (2007)
8. Opper, M., Sanguinetti, G.: Variational inference for Markov jump processes. In: Advances in NIPS 20. pp. 1105–1112 (2007)
9. Peuter, C.D.: Modeling Basketball Games as Alternating Renewal-Reward Processes and Predicting Match Outcomes. Master's thesis, Duke University (2013)
10. Shi, Z., Moorthy, S., Zimmermann, A.: Predicting NCAAB match outcomes using ML techniques—some results and lessons learned. In: Proceedings of the MLSA Workshop at ECML/PKDD 2013 (2013)
11. Shirley, K.: A Markov model for basketball. In: New England Symposium for Statistics in Sports. Boston, MA (2007)