Multivariate Dynamic Linear Models and NBA Player Tracking

Megan Robertson

April 28, 2017

1 Introduction

This project investigates the utility of time series modeling applied to player tracking data from National Basketball Association (NBA) games. The NBA began collecting player tracking information over the past few seasons, and this data provides the opportunity to explore basketball games at a deeper level beyond the usual box score statistics. For my Master's thesis, I used this data to create classification models predicting whether a shot would be made based on the movement of the players prior to taking a shot. These models were unsuccessful in predicting shot outcome and did not incorporate the spatial or temporal characteristics of the data. As a result I wanted to use the data for another project to investigate its spatial and temporal aspects.

Various projects using the tracking data have been conducted and can be found online. Scholars have investigated shooting ability, evaluated the three point shot, and more. [2] Every NBA team has access to the tracking data, but most of the research being conducted by teams is not published due to the competitive nature of the league. If teams find the tracking data useful to create game strategy or value players, they will keep it to themselves so their competitors cannot use it. Therefore it is not possible to know the extent of the prior work conducted with this data, but I have yet to find anything indicating the data has been used to fit dynamic linear models.

The goal of this project is to determine if an NBA player's movements are statistically exchangeable. Multivariate dynamic linear models with were fit using discount factor based methods to the tracking data for Avery Bradley, a guard on the Boston Celtics. Multiple models were fit to two different uninterrupted sections of game play. Feed forward intervention was used to account for Bradley's change in directions throughout the game that resulted in large forecast errors during the sequential updating of the model. The models differ by the values of the discount factors, and the marginal likelihoods of these values are compared in order to determine if the movement of the player is statistically exchangeable. In addition, the number of necessary interventions in each model accounts for the dynamics of the game and is informative about the movement of the player. Similar marginal likelihoods and the same number of interventions would mean that the two different game segments are statistically similar. If the marginal likelihoods and number of interventions are different, then the movement of the player is not statistically exchangeable.

2 The Tracking Data

Every NBA arena has had tracking cameras since the 2013-2014 season. The technology, known as SportVU data, is owned by the STATS company. The cameras shoot twenty-five images per second and the system translates the locations of the ten players and the ball to x,y coordinates. [3] The x-coordinate corresponds to the position relative to the sidelines and ranges from 0-94 feet, the length of a professional court. The y-coordinate measures position relative to the baselines and is between 0-50 feet, the width of a professional court.

The tracking data received by the teams contains the locations of the players and the ball over the course of a game. It also contains information about events such as shots being taken, fouls and more. However, this version is proprietary and not available to the general public. Therefore, this project uses a non-proprietary version known as the moments data that could previously be scraped from the NBA website. A single moment consists of the locations for the ball and ten players from one image taken during the game. This data contains the locations of the players and the ball as well as the time on the game clock and shot clock. The moments data does not contain information about the times of events such as passes, shots, fouls, or rebounds. The movements of a player can be recreated from the moments data by plotting their coordinates over time. Figure 1 displays the coordinates for Avery Bradley during three segments of a game.

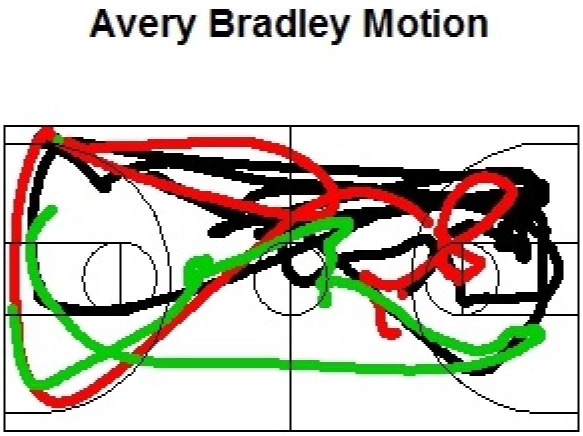


Figure 1: Three segments of Bradley's game motion. Code to draw the court image was kindly provided by Beau Coker and Will Eastman.

There are 0.04 seconds in between each observation since the cameras capture 25 images per second. The data was smoothed by reducing the observations to the ones occurring every 0.24 seconds. A logistic transformation was applied to the x and y locations to account for the fact that players' movements are restricted by the boundaries of the court. There will be smaller fluctuations in location when a player is near a boundary. The tracking data used for this project is from the November 21, 2015 contest between the Brooklyn Nets and the Boston Celtics. [1]

3 The Model

3.1 Mutlivariate DLM

A multivariate dynamic linear model was fit to game segments for Avery Bradley. After applying the logistic transformation, the data was differenced so that the velocity of the player was being modeled. The model is defined as follows:

$$\boldsymbol{\Delta y_t} = \boldsymbol{F_t'\theta_t} + \boldsymbol{v_t},\ v_t \sim N(0, v_t\Sigma)$$

$$\boldsymbol{\Delta y_t} == \begin{pmatrix} dx_t - dx_{t-1} \\ dy_t - dy_{t-1} \end{pmatrix} \text{where dx, dy represent the logistic transformed data}$$

$$\boldsymbol{\Theta_t} = \boldsymbol{G_t\Theta_{t-1}} + \boldsymbol{\Omega_t}, \Omega_t \sim N(0, W_t, \Sigma)$$

$$F = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, G = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

$$\boldsymbol{\Theta_t} = \begin{pmatrix} \alpha_{xt} & \alpha_{yt} \\ \beta_{xt} & \beta_{yt} \end{pmatrix}$$

Thus, the parameters of the state vector are as follows:

$$\alpha_{x,t} = \alpha_{x,t-1} + \beta_{x,t-1} + \omega_{x,t}$$

$$\alpha_{y,t} = \alpha_{x,t-1} + \beta_{y,t-1} + \omega_{y,t}$$

$$\beta_{x,t} = \beta_{x,t-1} + \omega_{x,t}$$

$$\beta_{y,t} = \beta_{y,t-1} + \omega_{y,t}$$

The α parameters correspond to the current velocity (in either the x or y direction as indicated by the subscripts), and the β parameters correspond to the current accelerations. Forward filtering was used to fit the models and calculate the marginal likelihood for β and δ. More details on models and the algorithms can be found in later sections and in Chapter 10 of Prado and West. [4]

3.2 Feed Forward Intervention

Feed forward intervention was introduced into the model-fitting process to account for the quick changes in directions that define a basketball game. The previous location is not the only aspect that affects where a player will travel next. Events such as a steal, rebound and more influence a player's movement. For example, if a player steals the ball on defense they will rapidly reverse their movement in the x-direction. Originally, play-by-play data was scraped to account for these changes in direction, but the data was not detailed enough to account for all the changes and thus was not appropriate to include in a model.

At each iteration of the forward filtering step, the error term for the one-step ahead forecast is calculated. As shown in Prado and West, this error term follows a T-distribution. For each iteration where the error term was in the extreme tails of the T-distribution, both discount factors were reduced so that the next predictions would depend less on the previous observations. For each iteration of sequential updating where intervention occurs, the parameters are not updated and there is no contribution to the calculation of the marginal likelihood for the discount

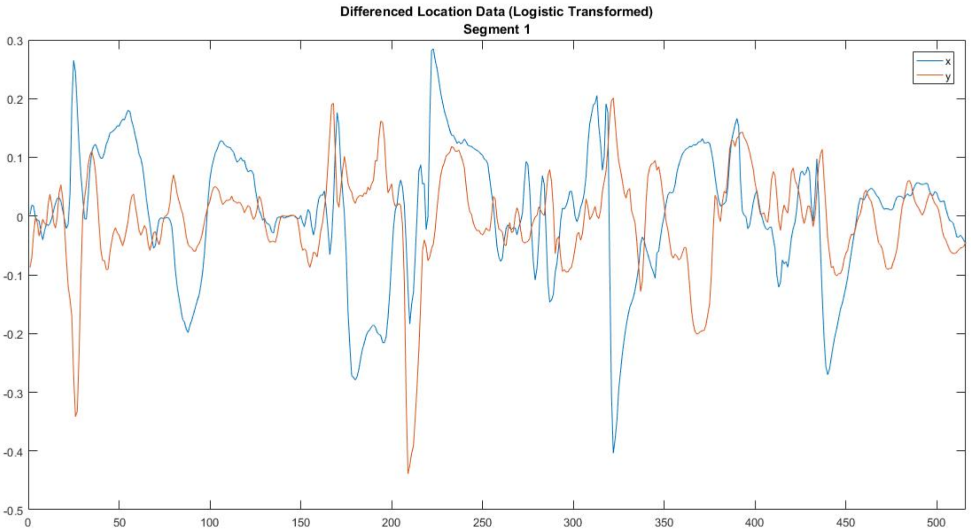


Figure 2: Differenced x,y locations for one game segment

factors β and δ. This allows the model to adjust to an extreme change in direction by not depending on a previous location that is drastically different from the future locations of the player.

Feed forward intervention was also used to remove any potential outliers Outliers are not surprising since the tracking cameras are not always accurate and thus should be removed. More information on feed forward intervention can be found in Section 11.2 of West and Harrison, 1997. [5]

3.2.1 Feed Forward Tuning

The application of feed forward intervention required the tuning of the reduced discount factors as well as the cut-off values in the tails of the T-distribution to determine when intervention should occur. Models were fit with different values for the cut-off in the tails of the T-distribution of the error term as well as for the reduced values of β and δ used in intervention. Smaller values of β and δ flatten the posterior distribution and make the T-distribution more variable.

Models with different values for these parameters were fit, and they were examined together with the data. A good choice for the T-distribution cut-off value would result in interventions being triggered where there are quick changes in either the x or the y direction. Appropriate choices for the reduced discount factors would reduce the dependence on the previous information enough such that there is not a need for another intervention immediately after. These were the characteristics sought after in the tuning stage. In addition, it was important to select values that did not cause the marginal likelihoods of β and δ, described in the next section, to converge too quickly.

Figure 3 displays the differenced logistic-transformed locations of Avery Bradley during a segment of the game. The yellow stars indicate instances of intervention. An intervention was

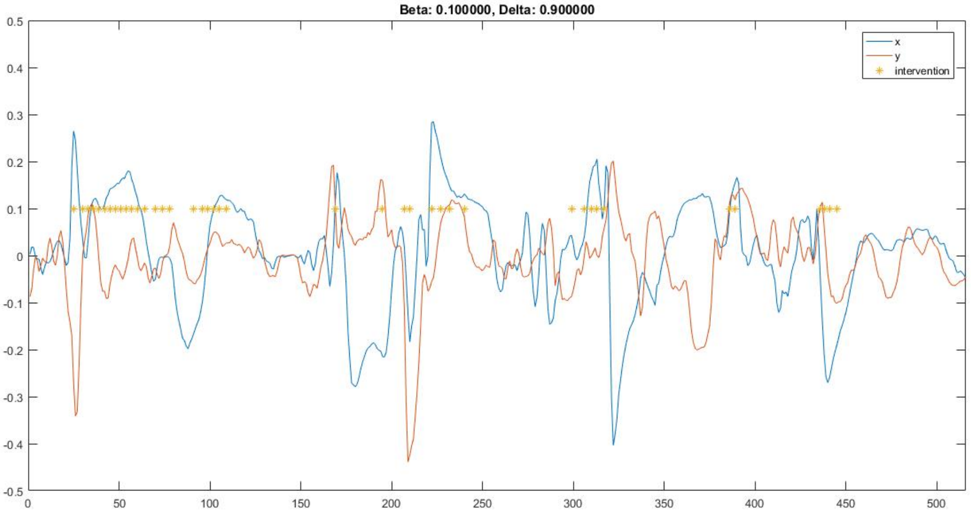


Figure 3: Data and instances of reduced discount factors.

triggered if the forecast error for either Bradley's x or y location was in the lower or upper 1% of the T-distribution. The values selected for the reduced discount factors were $\beta = 0.$ and $\delta =$

3.3 Marginal Likelihood for β and δ

Once the values for the T-distribution tail cut-off and the reduced discount factors were determined, multiple multivariate DLMs were fit. Forward filtering, as described in Prado and West Section 10.4 [4], was used to calculate the marginal likelihoods for β and δ. The values tested for β and δ were the numbers at increments of 0.01 between 0.8 and 1. Thus 441 models were fit and used in the computation of the marginal likelihoods for the discount factors.

Because the forecast error at each iteration in forward filtering follows a T-distribution, the marginal likelihood for each model (each pair of (β, δ)) can be computed using the probability distribution function of the T- distribution. Power discounting was used since marginal likelihoods tend to concentrate quickly. A power discount factor of $\alpha = 0.90$ was introduced in the calculation of the marginal likelihood so that useful comparisons could be made between the marginal likelihoods for different segments of the game. The convergence of the marginal likelhoods was also taken into account when selecting values for the reduced discount factors used during intervention. Matlab code was adapted from lectures for the model fitting process and Mike West helped add the necessary changes for the intervention and power discounting steps. Once the correct adaptations to the lecture code had been made, there were no major coding issues.

4 Results

Models were fit for two segments of the contest between the Celtics and the Nets. These sections were selected because they were the longest stretches for which Avery Bradley was in the game and there were no stoppages in play. The first segment begins at 4:37 remaining in the first quarter and ends with 2:32 on the clock. Figure 2 displays the data from this segment. The second game segment begins with 11:40 remaining in the fourth quarter and concludes at 9:47 left in the game. This data is displayed in Figure 4.

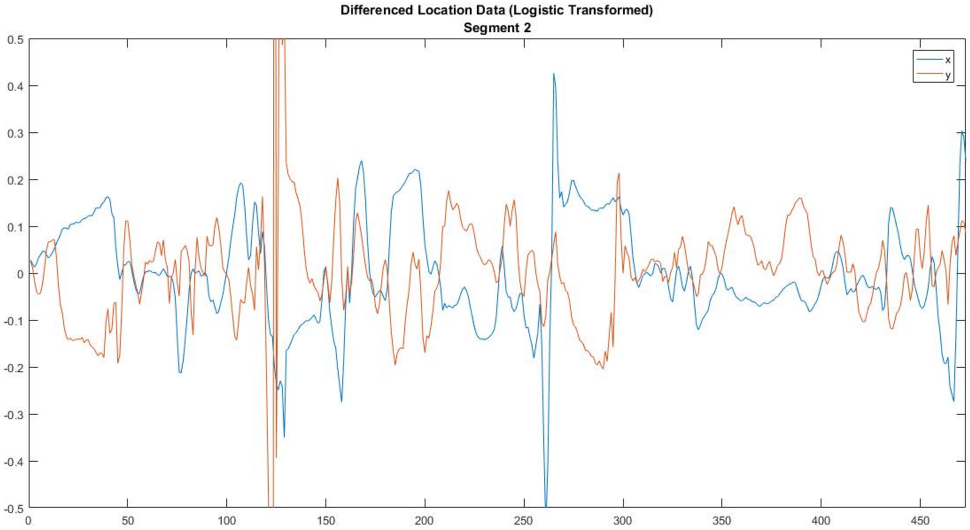


Figure 4: Data from the second game segment.

The multivariate dynamic linear model with discounting and feed forward intervention as described in the previous section was fit to both of the segments. The power discount factor used in the calculation of the marginal likelihood was selected to be $\alpha = 0.9$ to prevent the marginal likelihoods for β and δ from converging too quickly. The reduced discount factors used during intervention were $\beta = 0.1$ and $\delta = 0.9$.

The two game segments had similar marginal likelihoods for δ in that both were right-skewed. However, the δ marginal likelihood for game segment 2 has a more gradual decline in density as the δ values increase. The marginal likelihood for segment 1 drops off more quickly as the δ values increase.

The marginal likelihoods for the β values for the different game segments also differ. For the first segment, smaller values of β are more likely based on the data. On the other hand, larger values of δ are more likely in the second game segment.

The number of interventions for each model fit were also recorded. The mean number of interventions per unit time for models fit to the first game segment was 0.48 and the mean for the

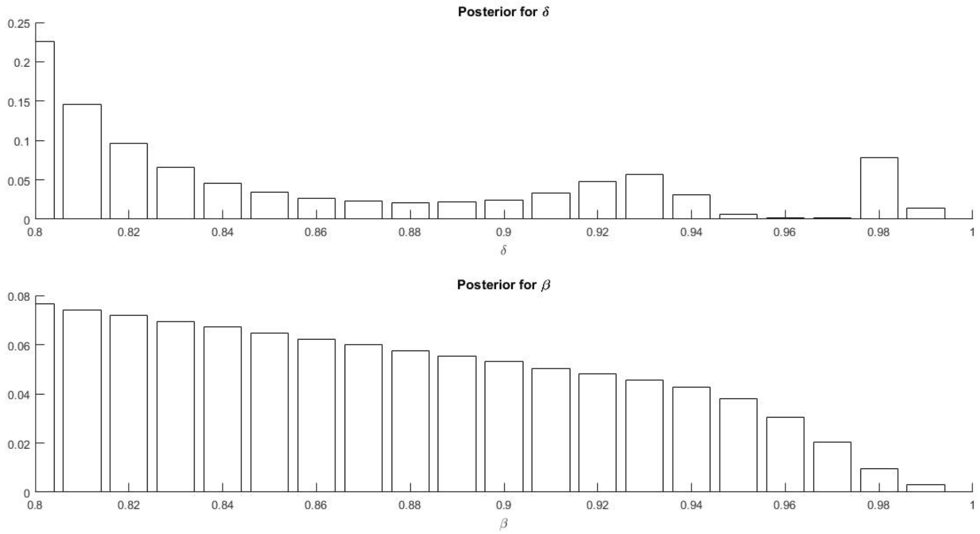


Figure 5: Marginal likelihoods for β and δ, game segment 1.

second game segment was 3.82.

A player will change directions after the occurrence of various events in the game. If a player is on offense and a teammate misses a shot and the other team gets the rebound, he is going to change direction along the x-axis in order to protect his team's basket. Thus the number of interventions is informative about the dynamics of a game. The figures in Appendix 1 display the play by play accounts for both game segments from the ESPN website. [5]

During the first segment, there are multiple events that lead to a change in possession of the basketball. The Nets grab a defensive rebound, the Celtics steal the ball and hit a shot, the Celtics get a defensive rebound and make a shot, the Nets miss and the Celtics grab the rebound, the Nets rebound a missed Celtics shot, and the Celtics steal the ball again and score. In the second segment, the Celtics get a defensive rebound, the Nets get a defensive rebound and the Celtics make a shot. The first game segment has more events that would cause a change in the x-direction, so one would expect more interventions in fitting models to the first segment. However, the play by play data does not account for all of the game events that would cause a player to change directions. Thus it is possible that the second game segment had some steals or passes that caused a change in direction that resulted in the need for intervention that do not show up in ESPN's play by play account.

5 Conclusion

The goal of the project was to determine if the movement of an NBA player during games is statistically exchangeable. The models fit for Avery Bradley's movement during two stretches of a game suggest that the movement is not statistically exchangeable. The marginal likelihoods

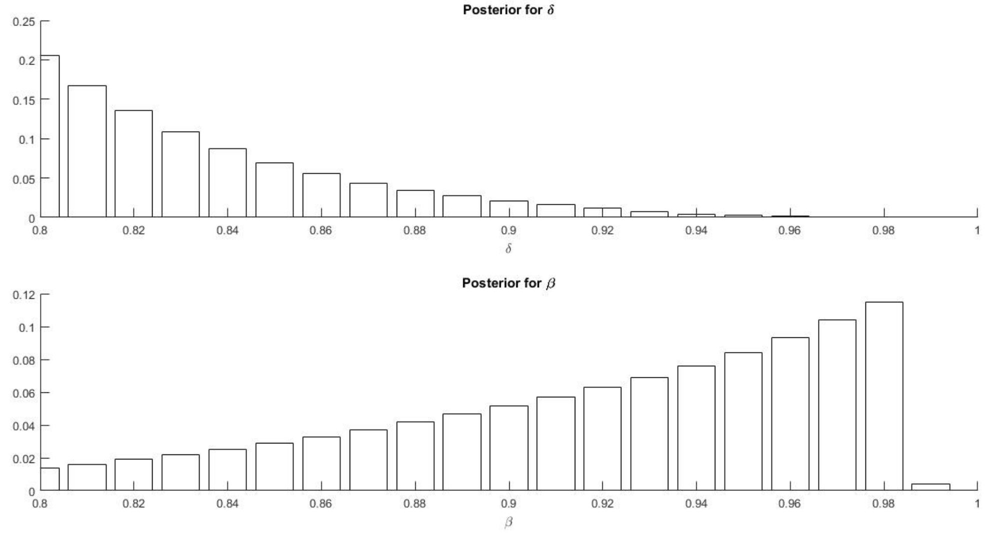


Figure 6: Marginal likelihoods for β and δ, game segment 2.

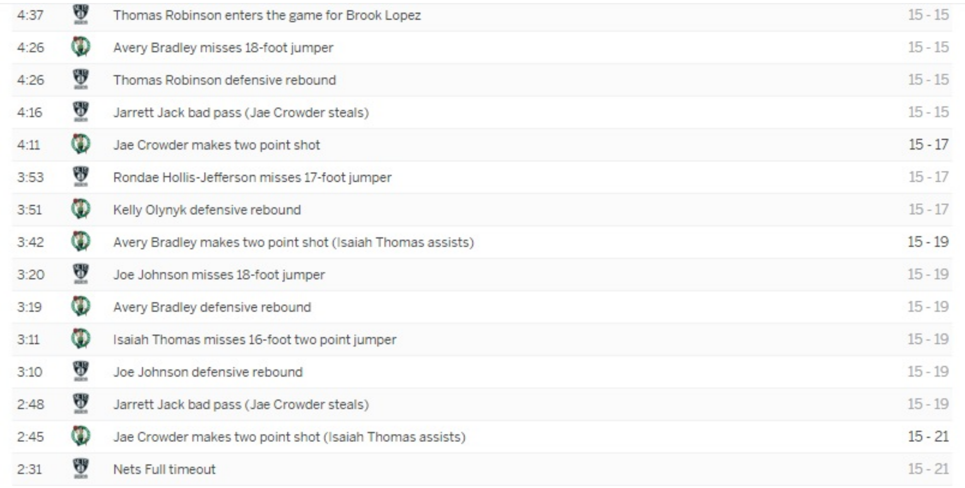
for the β and δ discount factors of the multivariate dynamic linear models are not the same for different game segments. In addition, the average number of interventions per unit time for each segment was different.

These results are not surprising since the movement of the player, particularly their change of direction, is going to be heavily influenced by the events of the game as well as the movements of other players on the court. Feed forward intervention is used to account for these changes by making the model less dependent on previous observations when the forecast error is large, but more information could improve the interventions.

The models could be be improved by incorporating co-variates such as whether or not a shot is being taken, if a player rebounds a missed shot, or if a steal has occurred. Including this information would provide a more informative method to account for the quick changes in direction that can be seen in both of the game segments used for modeling. Knowing the reason for why a player is changing direction is more informative since different events might result in more or less gradual changes. For example, a change in direction after a steal is probably going to happen much faster than a change in direction after a made basket.

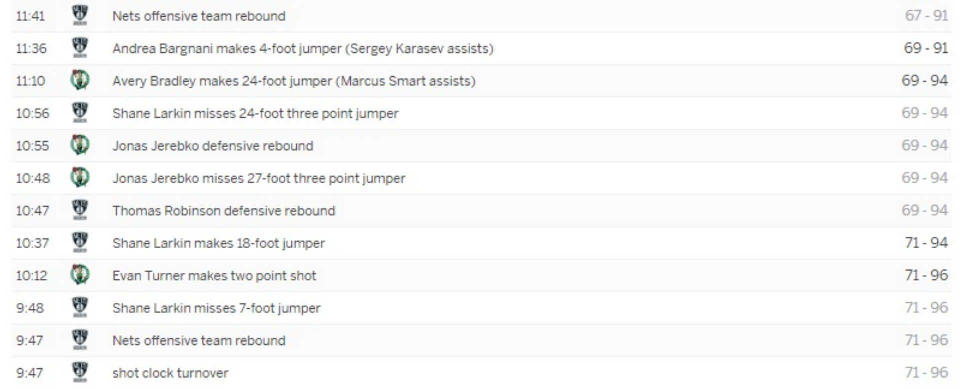
Even though the models could be improved by incorporating co-variates, the results are still promising. The project answered the question of whether the movement of an NBA player is statistically exchangeable. The results are not surprising, but they demonstrate the importance of how the dynamics of a basketball game affect how a player chooses to move throughout the course of the contest.

6 Appendix 1. Play by Play Data



4:37	Thomas Robinson enters the game for Brook Lopez	15 - 15
4:26	Avery Bradley misses 18-foot jumper	15 - 15
4:26	Thomas Robinson defensive rebound	15 - 15
4:16	Jarrett Jack bad pass (Jae Crowder steals)	15 - 15
4:11	Jae Crowder makes two point shot	15 - 17
3:53	Rondae Hollis-Jefferson misses 17-foot jumper	15 - 17
3:51	Kelly Olynyk defensive rebound	15 - 17
3:42	Avery Bradley makes two point shot (Isaiah Thomas assists)	15 - 19
3:20	Joe Johnson misses 18-foot jumper	15 - 19
3:19	Avery Bradley defensive rebound	15 - 19
3:11	Isaiah Thomas misses 16-foot two point jumper	15 - 19
3:10	Joe Johnson defensive rebound	15 - 19
2:48	Jarrett Jack bad pass (Jae Crowder steals)	15 - 19
2:45	Jae Crowder makes two point shot (Isaiah Thomas assists)	15 - 21
2:31	Nets Full timeout	15 - 21

Figure 7: Play by play account, game segment 1.



11:41	Nets offensive team rebound	67 - 91
11:36	Andrea Bargnani makes 4-foot jumper (Sergey Karasev assists)	69 - 91
11:10	Avery Bradley makes 24-foot jumper (Marcus Smart assists)	69 - 94
10:56	Shane Larkin misses 24-foot three point jumper	69 - 94
10:55	Jonas Jerebko defensive rebound	69 - 94
10:48	Jonas Jerebko misses 27-foot three point jumper	69 - 94
10:47	Thomas Robinson defensive rebound	69 - 94
10:37	Shane Larkin makes 18-foot jumper	71 - 94
10:12	Evan Turner makes two point shot	71 - 96
9:48	Shane Larkin misses 7-foot jumper	71 - 96
9:47	Nets offensive team rebound	71 - 96
9:47	shot clock turnover	71 - 96

Figure 8: Play by play account, game segment 2.

7 References

[1] neilmj, Basketball Data. *GitHub*, `https://github.com/neilmj/BasketballData/tree/master/2016.NBA.Raw.SportVU.Game.Logs`, 2016.

[2] Silz, Bradley. An Investigation of Three-Point Shooting through An Analysis of NBA Player Tracking Data `https://arxiv.org/ftp/arxiv/papers/1703/1703.07030.pdf`, December 2016.

[3] Partnow, Seth. Nylon Calculus 101: Intro to SportVU `http://nyloncalculus.com/2015/08/13/nylon-calculus-101-intro-to-sportvu/`, 2015.

[4] Prado, Raquel and Mike West. Time Series: Modeling, Computation and Inference. CRC Press: New York, 2010.

[5] West, Mike and Jeff Harrison. Bayesian Forecasting and Dynamic Models. Springer: New York, 1997.

[5] ESPN. Play by Play Data. `http://www.espn.com/nba/playbyplay?gameId=400828066`, 2015.