Yale University
Senior Project in Statistics and Data Science
S&DS 492

Foul Play: Examining NBA Player Behavior in Response to Foul Trouble

Gabriel Zanuttini-Frank
Advised by Professor Jay Emerson
May 2, 2019

Abstract

The rule of fouling out in the NBA, that a player who commits six fouls over the course of a game is disqualified for the remainder of the game, gives rise to very interesting strategic questions. These include decisions for the coach regarding when to substitute a player given his foul situation, as well as for the player in terms of how aggressively to play to optimize the balance between avoiding committing further fouls and playing effectively. This project performs a statistical analysis of players' defensive tendencies to determine whether foul trouble affects the way players defend, as hypothesized. Distance to the closest offensive (opposing team) player was used as the metric for evaluating defensive aggressiveness. The data used to carry out this analysis was of two types: real-time player movement data and play by play data.

A large portion of this project is devoted to data cleaning and manipulation, as well as the alignment of the two data sets. Particularly in the player-movement data, there appear to be many errors and inconsistencies, likely due to the inaccuracy and volatility of the player tracking cameras in the arenas. The statistical analysis in the end shows several logical trends, but does not yield promising results on the effect of foul trouble on defensive play. We do find a significant effect of distance on the likelihood of a player committing a foul, and further, there does appear to be a positive relationship between a player's number of fouls and his distance from the player he is guarding, controlling for other factors such as score differential, shot clock time remaining, game time remaining, and position. Three different measures of foul trouble are used as predictors for defensive player distance to the closest offensive player: number of fouls, squared number of fouls, and likelihood of fouling out in the game. Because none of these measures show a significant effect on defensive play, we can conclude either that players are not modifying their play in response to foul trouble, that more nuanced models of foul trouble are required to capture the situations in which players do adapt, or that the tracking data does not provide the right kind of information to measure the nature of players' responses.

Contents

1 Introduction

According to Rule 3, Section 1 of the NBA rulebook, "A player is disqualified from the game when he receives his sixth personal foul."[1] The presence of this rule in the game of basketball in many cases imposes a restriction on the amount of time that one individual player can spend in the game. This gives rise to the concept of foul trouble, which can be loosely defined as a situation in which a player is in danger of exceeding the number of permitted fouls and this being disqualified from the game. Due to this impending potential risk of disqualification, or foul trouble, strategic questions arise from the perspective of both coaches and players. At the level of the individual player, there is the question of how their style of play is or should be altered by the risk of being expelled from the game. At the level of the coaching staff, there is the question of whether to, at least temporarily, substitute a player out in order to increase the likelihood of having him available later to play the final and more crucial minutes of the game.

Several other common reasons exist for substituting players off of the floor in the game of basketball: injury, fatigue or rest, and optimal opponent match-up. These are fairly straightforward; an injured or fatigued player will not be capable of playing to his maximum capability at that point in time, a player who perhaps recently returned from injury will require periodic rest to regain strength and stamina, and certain players might be more apt to defend or attack a specific matchup presented by the opposing team. Of these four main reasons for substitution - the three defined above in addition to foul trouble - foul trouble is unique for two reasons: it both has no bearing on the physical state of a player, and its effect is thought to change over the course of a game. For example, in the fourth quarter, it is widely believed, and players will certainly agree, that they play differently whether they have committed two or five fouls. Therefore, holding all other factors describing the state of the game constant, the effect of foul trouble can be isolated and explored.

[1]There is technically an exception to this rule: "If a player in the game receives his sixth personal foul and all substitutes have already been dis-qualified, said player shall remain in the game and shall be charged with a personal and team foul." In this case, a player can in fact remain in the game despite having six personal fouls.

This project focuses on the latter, player strategy-focused effect of foul trouble, and more specifically seeks to analyze the impact on foul trouble specifically on the way in which players defend; whether the presence of foul trouble causes them to play less aggressively, and how much this varies by the degree of foul trouble, the time and score of the game, and across different positions and teams. In order to measure this impact, I must first outline definitions for players' defensive aggressiveness as well as foul trouble.[2]

In order to define player defensive tendencies, or more specifically aggressiveness, I will use their tightness to the player they are guarding. Keeping track of this value is fairly complicated, as NBA defenses, especially certain teams (most notably the Warriors and Rockets) employ strategies that feature countless defensive switches. As a result, over the course of a single defensive possession one player can guard up to five different opponents, making it almost impossible to know who the intended matchup is at any given fraction of a second. To account for this complication, I simply define a player's intended matchup as his closest opponent, and therefore his defensive tightness as the closest distance to an offensive player. Naturally, there are shortcomings that arise from the use of this definition. For example, when the offense is setting a screen, it is likely that the on-ball defender's closest distance to an opponent is the screener rather than the ball handler. Similarly, on drives to the basket, it will often be the case that many players are bunched together and no longer closest to their original matchup. In the case of the screen, these plays happen so quickly that the amount of time in which a player is closest to the screener is very minimal in the grand scheme of the possession. And in the case of a drive to the basket, the amount of time near the rim is very minimal, and the difference in distances between the players is so small that it probably does not matter to whom the closest distance is calculated. Thus, this definition is not perfect, but should be satisfactory in its ability to reveal changes in defensive aggressiveness.[3]

[2]For the purposes of this project, I will ignore the case described in the previous footnote, in which a player can remain in the game with six personal fouls. I can only recall one instance in which this rule was put into effect (Lakers vs. Cavaliers, February 5^{th}, 2014), and I do not believe that players consciously or subconsciously take into account the slim possibility of continuing to play with six fouls, meaning that it should not have any effect on the players' tendencies while in foul trouble.

[3]There does exist a dataset owned by the NBA, titled Summarized Tracking Data, that includes features that would greatly help for the purposes of this project such as shot location, primary defender, and distance to shooter. I unfortunately could not get access to enough games of this data, so I could not

It is possible to track and calculate these distances between players because of the relatively new advent of SportVU cameras to NBA arenas. These cameras were installed prior to the 2013-14 season, and track every player and the ball 25 times per second. They produced the first dataset of its kind in the realm of basketball, allowing for much more granular information on not only the plays in a game but the continuous movement that leads up to an event or change in possession. Therefore, this player tracking data, as it has come to be named, in theory contains the necessary information for measuring players' distances from one another. I will validate this fact in the next section.

The second requisite definition for this project is that of foul trouble. A first, simple, definition for this could simply be the number of fouls that a player has committed thus far in a game. It is certainly the case that a player's number of fouls is positively correlated with the likelihood of fouling out, and in turn, foul trouble. However, this relationship is probably not linear, and it is not difficult to grasp that foul trouble is likely more nuanced than just the number of fouls, as it should take into account at the very least the time remaining in the game. Most would agree that a player with three fouls in the first quarter will play more cautiously than one with three fouls with a minute left in the game, at which point the chances of fouling out are very slim. As a result, I will create two different measures of foul trouble, both defined as the likelihood of fouling out before the end of the game, which incorporate features such as the time remaining and the current score differential.

2 Data Preparation

2.1 Data Sources and Descriptions

As previously mentioned, two kinds of data were collected for this project. The player tracking data and play by play data. The former, a subset of which was in the past released to the public via *stats.nba.com*, is now fully proprietary data in the hands of the NBA, its teams, and its affiliated parties. Therefore, I was not able to obtain multiple

take advantage of it.

seasons' worth of data from an unquestionably reliable source. Rather, I was able to find the SportVU data from every game in the first half of the 2015-16 NBA season (October through January) on the GitHub repository of "sealneaward".[4] In addition to this, I found another repository from "rajshah4"[5] that defines several functions for processing or calculating certain results from the data, a few of which I modified and employed in carrying out my analysis. For example, each individual game was stored as its own zipped JSON file, so in order to import the data into a dataframe in R I modified and used one of the functions defined in rajshah4's script.

The player tracking datasets for each game consist of over two million rows each, and describe the position of each player on the floor as well as the ball 25 times per second. Locations are denoted by x and y coordinates given in feet, which can be directly interpreted as locations on the court. The state of the game is defined by the variables **game_clock**, which shows the number of seconds remaining in the current quarter, **quarter**, and **shot_clock**, which is also given in seconds, rounded to the nearest hundredth. Finally, each player and team has a unique integer identifier, and the **event.id** column supposedly describes the outcome of the current play. Putting these all together yields 11-observation-long blocks (for 10 players and the ball) for each timestamp, extending over the course of the entire game.[6]

Play by play data was scraped using the *jsonlite* package from *stats.nba.com.*[7] This data was more straightforward and easy to manipulate and interpret. It featured 33 variables in total, only 8 of which I ended up keeping. My filtered version of the play by play featured variables denoting the **quarter**, the time remaining in the current quarter, the **score** represented as a string (i.e. "4-2"), a general numeric identifier of the type of play, a more specific numeric identifier for the play type, and a written description of the outcome of the play.[8]

Both of these datasets provided lots of very interesting information, although each

[4]https://github.com/sealneaward/nba-movement-data

[5]https://github.com/rajshah4/NBA_SportVu

[6]An example can be found in Table 1 the Appendix.

[7]Ex: https://stats.nba.com/stats/playbyplayv2?EndPeriod=10&EndRange=55800&GameID=0021500492&RangeType=2&StartPeriod=1&StartRange=0

[8]An example can be found in Table 2 of the Appendix.

one individually would not be sufficient to support the analysis in this project, as the player tracking data did not contain explicit information about the types and outcomes of each play whereas the play by play data did not contain information about the locations of each player. As a result, my next step was to combine the two, which would yield a dataset detailing both the specific kind of play and its outcome as well as the granular player positioning every 25^{th} of a second. With this, it would be quite simple to calculate values such as the number of fouls on each player at every point in time and to carry out certain analyses that are filtered on certain types of plays. For example, it would be easy to examine defensive player aggressiveness only on possessions that ended in three point attempts, as these likely come with more floor spacing and thus simpler identification of primary defenders.

2.2 Data Exploration, Issues, and Inconsistencies

Before merging the datasets and especially before beginning to do any analysis, I dug into and explored the player tracking data. There were several reasons why this was particularly necessary in this case. First, the tracking data is quite novel and I did not have experience in manipulating it. Second, I did not obtain it from the most reputable source. Finally, because the cameras that track the players and ball might not be perfectly reliable, there might be errors in the data. Unsurprisingly, I found numerous issues with the data that did not all seem related back to a single source. Rather, according to my personal inference, some could be attributed to human error from those operating the tracking cameras or the game clock operators, others to mistakes of the imperfect technology, and others perhaps to the inauthenticity of my dataset.

The first oddity that I found while exploring this dataset was in the movements of the shot clock. Player and ball locations are supposed to be measured every 0.04 seconds, and this does appear to mostly be the case. Thus if we look at a histogram of shot clock differences, defined as the difference in values of the shot clock between two adjacent observations in the data, we expect to see mostly differences of 0.04 with a number of larger and positive differences at each change in possession and reset of the shot clock.

However, in no circumstance in basketball is the shot clock ever decreased without the passage of time, so it is curious that there are large negative differences in the shot clock:

The question naturally arises: why are there breaks of more than 1 second, specifically 13 of them in this single game featuring Charlotte and Toronto on January 1^{st}, 2016? One specific example is that at 413.51 of the game clock (or 6:53 remaining) in the third quarter, Charlotte calls a full timeout. Following this stoppage, the tracking data only starts up again at 411.73 of the game clock, almost a full two seconds of game time later. There is no apparent reason for why those two seconds were lost, as the clock was not artificially adjusted during the timeout by the referees. I surmise that, if the cameras are manually controlled, they were not started back up in time for the resumption of play, or that they simply malfunctioned for a few seconds of the game. This is not a unique case, as a similar situation took place in another game as well between Washington and Indiana on January 15^{th}, 2016. Here, Indiana called a full timeout at 5:48 of the first quarter, at which point the cameras stop tracking but then do not resume recording data until the clock reads 5:19. The point at which the cameras start working again does not seem to follow any sort of pattern, as 5:19 is in the middle of a play. This case is more consequential not only because more time went unaccounted for but also because it means that the data is missing the play that occurred between 5:48 and 5:36 that resulted in a missed jump shot. These two examples of inaccuracies in this dataset do give us reason to pause and consider what other types of issues will come to light as well as to question

the validity of this data as a whole. However, from the cases that I investigated, there does not seem to be a pattern for which plays are partially or fully missed by the cameras and thus this dataset, and therefore have no reason to believe that these issues introduces any form of bias in the data.

Following up on the inconsistencies of the shot clock values, I checked whether the shot clock values were consistent for any given game clock value. In other words, do the two clocks always move together and by the same amount, save for when the shot clock is reset? It appears that this is not always the case. Looking at the same game as before between Charlotte and Toronto, there are over 50 instances of the game clock that have multiple distinct values of the shot clock attached to them. Certain cases have explanations that are fairly simple to infer; the shot clock values for the given value of the game clock differ only by a few hundredths of a second and are all within the span of a second. Since the game and shot clocks are manually operated, I suppose that in these situations the shot clock was started slightly before the game clock, leading to multiple shot clock values for the fixed value of the game clock. Other instances, however, are more concerning from a data reliability perspective. At 3:04 of the second quarter, a defensive 3 seconds technical foul is called on Jonas Valanciunas of the Raptors, causing the clock to stop for the Hornets to shoot their awarded free throw. Over the course of this game clock timestamp, the shot clock values range from 18.87 to 13.91. This is worrisome and essentially renders these timestamps untrustworthy, a problem which will later be dealt with. However, all inconsistencies of this sort do appear to occur at clock stoppages. Since this project is concerned with players' distances from each other during live play and given the fact that there are very few of these shot clock variations (in this case they occur in 55 out of 69,819 game clock instances), we can proceed without worrying too much about how this could affect the results of the analysis. While it would be incredibly useful to cross-check the accuracy of each shot clock value with another data source, this would also be incredibly time-consuming, so I will proceed keeping in mind that there remains concern as to the reliability of the shot clock values in this dataset.

Moving on from the issues related to the shot clock values, I seek to better understand

the structure of this dataset. I do so by examining the "blocks" of observations that are present for each player at a given timestamp. These consist of groups of consecutive rows in which the player and time remain constant, while his location changes. I would not expect such blocks to be present in the dataset because the cameras are only supposed to track players during gameplay. Below is a subset of the frequency table of the sizes of these blocks:

Size of block	Frequency
1	214067
2	272699
3	133232
4	96289
5	35777
6	44105
7	9874
8	10603
9	542
...	...
500	6
555	1
588	1
630	2

We first observe from this table that a vast majority of blocks are of size less than or equal to 6. The reason we do not expect the majority of blocks to be of size 1 is that each observation has an event ID attached to it that supposedly describes the present of a distinct event, and a given play can have multiple event ID's attached to it. An event can be anything from a made basket, rebound, or foul to an ejection, substitution, or injury stoppage. So a typical block, of size 3, for example, looks as follows:

player_id	lastname	firstname	position	team_id	x_loc	y_loc	game_clock	shot_clock	quarter	event.id
200768	Lowry	Kyle	G	1610612761	10.77643	38.39262	713.26	13.15	1	2
200768	Lowry	Kyle	G	1610612761	10.77643	38.39262	713.26	13.15	1	3
200768	Lowry	Kyle	G	1610612761	10.77643	38.39262	713.26	13.15	1	4

In addition to the sizes of certain blocks, I would have expected the frequencies to be divisible by 11, since the cameras are supposed to track all players and the ball when they are actively tracking. I first observe the blocks of length greater than 600 to try to infer a cause for these apparent issues. The two blocks of 630 observations occur at the same timestamp, namely a foul at 3:09 of the third quarter. The cameras continue to track only

these two players, Patrick Patterson and Cory Joseph, even as the clock is stopped. It also does not appear that there is a connection between which players continue to be tracked; Joseph was substituted into the game at this point, while Patterson was neither involved in a substitution nor the foul call. Looking at the next two largest blocks of size 588 and 555, they both involve a stoppage in play for a foul and both track Kemba Walker, although in one instance he got fouled and in the other he is not involved in the play. In conclusion, it seems as though the cameras sometimes randomly continue to track players' positions even when the clock is stopped, and there does not appear to be a reason behind which players this happens to. This is useless noise in the data and can be removed.

Now, I attempt to account for the issues that were just exposed and outlined. For a given player ID, event ID, and quarter, I will take the first and last instance of each unique game clock value. Doing this removes the possibility of there being multiple shot clock values for a single game clock and event ID, and eliminates the existence of the large blocks defined above. This filtering also shrinks the size of the dataset by roughly 10%, making it faster to work with later on. To ensure that this process was successful, we can reexamine the table of block sizes as well as the blocks that were previously hundreds of rows long. In fact, the block of Patrick Patterson that was previously 630 rows long is now reduced to 10, and the same is true for Cory Joseph and Kemba Walker. Further, we can see in the table below that there are no longer enormous blocks of unnecessary data, as there are none larger than 16 at this point.

Size of block	Frequency	Size of block	Frequency
1	198557	8	9671
2	254130	9	7
3	124111	10	2810
4	89583	11	1100
5	33558	12	847
6	41440	13	2475
7	9616	16	11

2.3 Data Merging

Perhaps the most difficult undertaking of this project was merging the player tracking and the play by play datasets. According to the documentation available online regarding the player tracking dataset, it should be very simple to merge this data with the exact play by play data that I scraped from *nba.com*, as the event ID variables are supposed to match up nicely with the events described in the play by play data. Upon quick inspection it is clear that the event IDs do span the same range of numbers in each of the datasets, although as mentioned previously the event IDs are given in blocks. If we look at the first block, an example of which can be found at the top of page 10, and cross-check it with the play by play of the same game, using Table 3 in the appendix we find that these three events correspond to a miss, a rebound, and a make, or more specifically a missed driving floating jump shot, a rebound, and a putback dunk. It makes logical sense that these three events be grouped together as constituting a single possession, as a rebound and putback in the span of a second are not equivalent to a full possession leading up to a shot or turnover. The one problem with these blocks, however, is that they overlap with other event IDs. For example, instead of having a new event ID after all of the values of 2, 3, and 4 occurring, an observation with an event ID of 5 pops up in the midst of the block of 2, 3, and 4. Without combing through the play by play data manually and grouping the events that should constitute a single possession, there is no way of controlling this overlap and thus separating out possessions.

My first attempt to circumvent this problem was simply to merge the datasets on the game clock. This idea was quickly quashed because the play by play data only has game clock instances at the ends of possessions rather than at all points in the game, and especially because the player tracking data has the game clock to the hundredth of a second whereas the play by play game clock values are rounded to the nearest integer. In response to this, an option could be to round the game clock to the nearest integer or to create a function that takes the game clock and returns a play occurring within a few seconds. This will capture all of the observations in the player tracking data before the end of the play, but might also include some after the shot was taken or the turnover

was committed. This would affect the result of, say, the average distance to the closest offensive player over the course of the possession because it might include some time when the player's goal is to run back on defense as quickly as possible rather than to guard his opponent as tightly as possible to prevent him from scoring.

Rather than rounding the game clock, the solution that I employed in the end was to denote ends of plays using jumps in the shot clock of more than a tenth of a second. Keeping in mind the issues that I previously discovered regarding the reliability of the shot clock values, we cannot assume that a shot clock jump necessarily indicates the end of a possession. However, we can still go in the other direction; since we know reliably when the ends of possessions occurred from the play by play data, we can only use the shot clock jumps that occur within a few seconds of the end of the play. This is still subject to error if there is no shot clock jump near the time of an end of possession, but it is safer to exclude such plays as opposed to estimating the ends of plays and inevitably counting certain observations that were not part of a play.[9] This merge appeared to work fairly well: around 85% of the unique stoppages recorded in the play by play of the two games being discussed were merged with entries in the tracking data.[10]

2.4 Other Filters Applied

In order to further compress the dataset to make running functions more efficient and to reduce inessential data, I applied a few additional filters on the merged dataset. First, I removed rows that were duplicates of the triple (game clock, quarter, player ID). This removed primarily rows that differed on event ID, which I did not end up using anyway, in addition to certain rows that were kept when eliminating the blocks of player observations. Since I only take the first observation of each set of duplicates, there could be unnatural jumps in player distance traveled between two timestamps as a result of movement during a clock stoppage. However, the functions calculating distance traveled and velocity check

[9]There was even a case in the Charlotte-Toronto game where the tracking data entirely missed a play. The movement data during the play of Jonas Valanciunas' dunk at 11:35 of the first quarter did not reflect this occurrence whatsoever. This is another reason for concern about this dataset, although plays such as this were excluded from the merge of the datasets because they did not have a corresponding shot clock jump.

[10]204/240 for CHA@TOR and 206/245 for WAS@IND

for and eliminate anything faster than 1 foot per .04 seconds, so these jumps should be ignored from the analysis regardless. Next, I found the times at which the ball crossed halfcourt, and removed all observations where the ball is in the backcourt, as the distances from defenders are not informative in these situations. Third, I removed instances of the game clock where there were not exactly 11 observations (five players on each team plus the ball), which were often due to substitutions during stoppages. Finally, based on the location of the ball that indicated which team was in the frontcourt, I filtered out all of the players on offense, leaving a dataset roughly 10% of the original tracking dataset's size, containing only defensive players, their positions only while on defense in their own half of the court, and the outcome of each play.

2.5 Data Validation

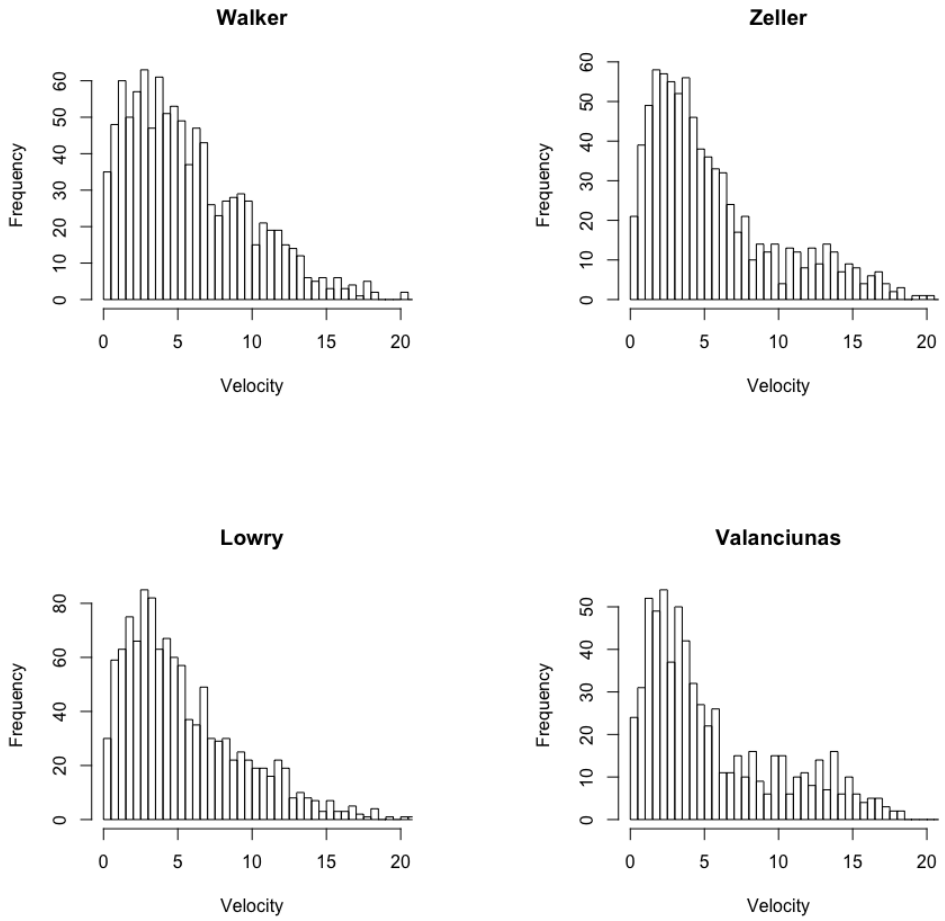
Due to the numerous issues that were discussed in this section, it was necessary to validate, or perform sanity checks, on the data before it could be used to perform any analysis. In a previous project, I compared the average distance covered per game, on both offense and defense, across positions, and found that point guards, shooting guards, and small forwards all cover more distance per game on average than centers. In addition, the teams in the game being explored, the Hornets and Raptors, both feature centers (Cody Zeller and Jonas Valanciunas, respectively) that like to play in the paint more than on the perimeter and therefore exacerbate this gap in average distance covered per game.[11] The table below lists the average distance traveled per second, the percentage of time in the game spent above the free throw line (on the halfcourt side of the free throw line, on either side of the court), and the percentage of time spent in the paint for all ten starters in the game.

We see in this table, which is sorted by time in the paint, the general trends that we would intuitively expect. Most evident is the fact that centers and forwards spends vastly more time in the paint and conversely less time above the free throw line than the guards and wings. Differences in the average distance traveled are much more difficult to

[11]This claim is supported by these statistics: https://stats.nba.com/players/speed-distance/?Season=2015-16&SeasonType=Regular%20Season&TeamID=1610612766

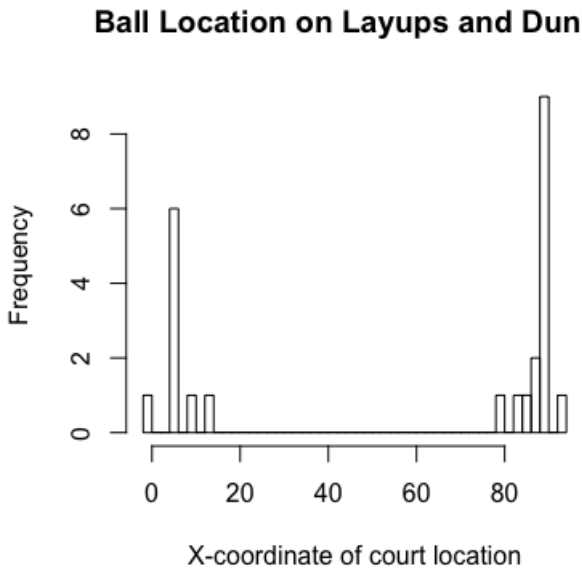
Name	Pos	Avg Dist	Time above FT line	Time in paint
Valanciunas	C	5.10	0.33	0.52
Zeller	C-F	4.92	0.32	0.48
Williams	F	5.19	0.42	0.28
Scola	F	5.29	0.40	0.25
Hairston	F-G	5.20	0.46	0.16
Walker	G	5.47	0.61	0.15
DeRozan	G	5.13	0.46	0.15
Batum	G-F	5.21	0.38	0.15
Lowry	G	5.13	0.59	0.13
Carroll	F	4.86	0.40	0.13

glean for two reasons. First, the values are all quite close, which means that a little bit of noise might be all that separates most of the players in this category. Second, while we do expect the guards to have slightly higher average distances per second (equivalent to average speed), what may be more telling is the distribution of the speeds of the players over the course of the game. Below are the distributions of the speeds of four starters, namely the point guards (on the left) and centers (on the right).



As we would expect, the distributions of the centers' speeds are more skewed right, meaning that the guards travel at higher speeds more often. This is especially evident when comparing the distributions of Walker and Zeller. Because the trends seen in the table above as well as these histograms are in line with our intuition, we can be more comfortable utilizing this dataset as a basis for further analysis.

Lastly, I will perform a quick check of the merged dataset. Here, I examine the ball location at the end of a play that resulted in either a dunk or a layup to ensure that the values are reasonable.



Again, the results are sound. This histogram is bimodal, and the locations of the ball are all close to one of the two baskets. Again, this instills confidence in the reliability of the data.

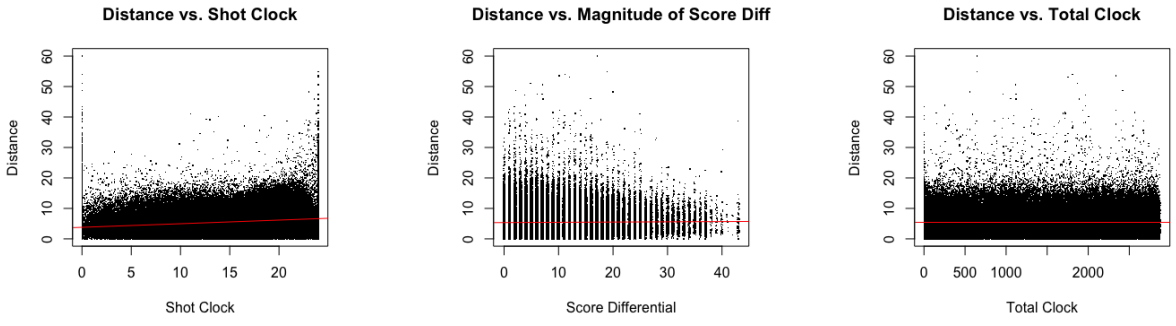
3 Analysis

3.1 Validating closest defender distance as a measure of player agressiveness

The first step in my analysis is to examine the relationships between defenders' distances to their closest opponents and other variables that might be correlated in some way with

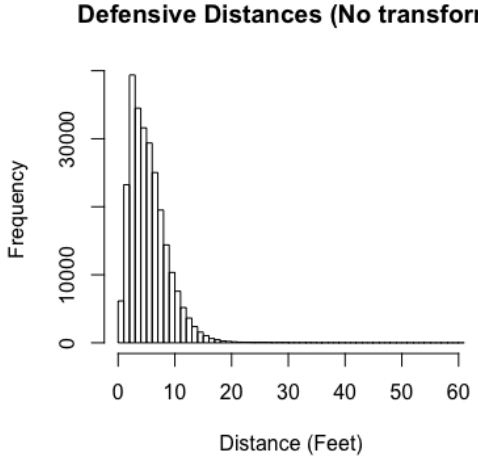
perceived aggressiveness. In doing this, I will use a filtered version of the dataset that contains only instances of the players who are guarding the ball at any given point in time. This is an attempt to reduce the noise in the data, as the players guarding off of the ball have less of a direct incentive to be as close to their man as possible, and their aggressiveness will therefore not be well reflected in the distance measure being used. The on-ball defenders' assignments are much more straightforward, and while they might not always be trying to minimize the distance between themselves and the ball, it is much more likely that they respond to foul trouble by backing off of their man than those not guarding the ball. The player guarding the ball is found by selecting the player out of the five defenders that has the smallest distance to the ball itself.

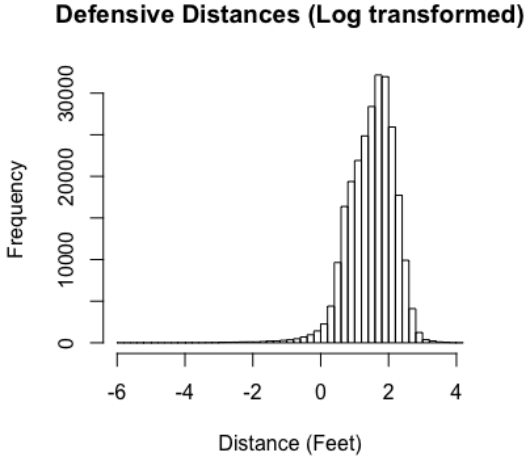
The data used for this analysis is also cut down by a factor of 25, and now contains location values separated by roughly one second each, as compared to the previous ones that were typically separated by .04 seconds. Even though this decreases the total number of observations to build models with, this also reduces the correlation between sequential points in the data. With observations every .04 seconds of game time, there was almost never independence among neighboring observations, as the location of a player at one point in time strongly determines his location .04 seconds later. Such a lack of independence in the data would render the standard errors and t-statistics of any regressions biased and therefore unreliable, which would hinder this analysis from yielding any significant results.



These three plots above demonstrate the relationships between defensive distance, or the distance to the closest offensive player, and shot clock, magnitude of score differential (absolute value of the score differential), and total game clock. Score differential was

converted to an absolute value because I do not expect it to be linear; rather, I hypothesize that players will play tighter defense as the game gets closer, or as the score differential approaches zero. From a quick glance at these plots, we can observe that score differential and total clock do not appear correlated with distance (these correlations are .016 and -.0018), while shot clock and distance are slightly positively correlated (correlation is .188). Before running any linear models on defensive distance, I will check its distribution to see whether a transformation could be applied to make it more normally distributed.





It does appear that log-transforming the distance variable makes it more normally distributed, so we will apply this change prior to conducting further analysis. Now, I continue to explore the relationships of shot clock, score differential, and game clock with defensive distance by running a series of regressions, the results of which are listed below.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.4430	0.0022	660.97	0.0000
scale(total_clock)	-0.0008	0.0023	-0.34	0.7335
scale(abs_scorediff)	0.0076	0.0023	3.25	0.0011
scale(shot_clock)	0.1211	0.0022	55.35	0.0000

In this regression, absolute value of score differential and shot clock are significant while total clock is not. This means that the former two variables do explain some of the variation in defensive distance. Further, the signs of the coefficients of these two variables are as I had hypothesized: as a game gets closer and *abs_scorediff* shrinks, the on-ball defender's distance to ball-handler decreases, and as the shot clock winds down,

the defender also gets closer to his man. Because total clock is not significant in this model, I will emit that from my later analyses. Nonetheless, two out of the three variables examined in this model were significant and demonstrated behavior in the direction that we would intuitively expect. This result gives validation to the use of defensive distance as a measure of aggressiveness, as we associate higher aggressiveness in basketball with closer games and ends of possessions.

3.2 Agressiveness and Fouling Behavior

Next, we would like to show that aggressiveness, which we have defined as proximity to the closest opponent, leads to greater likelihood of fouling. If this is the case, not only would it match our intuition but it would corroborate the idea that players try to play less aggressively when in foul trouble because of the high comparative cost of committing additional fouls. The table below compares, only for on-ball defenders, the mean distance from the opposing player on plays that ended in a foul committed and those that did not, controlling for his number of personal fouls at that point in the game.

Number of Fouls	Did he foul on this play?	Mean Distance
0	FALSE	5.48
0	TRUE	4.36
1	FALSE	5.45
1	TRUE	4.37
2	FALSE	5.43
2	TRUE	4.26
3	FALSE	5.41
3	TRUE	4.36
4	FALSE	5.58
4	TRUE	4.14
5	FALSE	5.70
5	TRUE	3.66

For each distinct number of possible fouls, 0 through 5, the mean distance on plays that ended in a foul committed was lower than that on plays that did not. This is in line with what we would expect, that playing more aggressively, or closer, on defense results in a higher chance of committing a foul. One potential counterargument to this analysis is that it is biased; plays that end in fouls must end with a distance of zero between

the players while the other plays do not, so there is an inherent reason why the fouling plays' mean distances are closer rather than strictly being due to the fact that playing closer makes a player more prone to fouling. There is certainly merit to this argument. However, the amount of time in which the foul is being committed is very small in the grand scheme of the possession, so it should not have too much bearing on the average distance. The differences in the table above seem to be too large to be accounted for by one abnormally small observation. In addition, a logistic regression predicting whether a foul was committed given the distance yields a significant result, with the negative coefficient denoting that distance is inversely related to the likelihood of committing a foul.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.4962	0.0188	-132.95	0.0000
closest_dist	-0.1259	0.0037	-34.36	0.0000

3.3 Measuring the impact of foul trouble on defensive distance

Finally, we can introduce a measure of foul trouble into the model to examine how well it predicts defensive distance. The first definition of foul trouble that I will employ is the most generic and straightforward: a player's number of fouls up to that point in the game.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.4506	0.0020	732.42	0.0000
scale(abs_scorediff)	0.0083	0.0020	4.13	0.0000
scale(shot_clock)	0.1200	0.0020	60.51	0.0000
scale(num_fouls)	-0.0030	0.0020	-1.50	0.1330

Adding a player's number of fouls to the linear model does not improve its performance, and the variable denoting the number of fouls is not significant. This is not entirely surprising, as I did not expect the effect of the number of fouls to be linear. As the game moves closer to the end, I would expect the effect of the number of fouls, or foul trouble, to magnify. For example, if a player has five fouls with eight minutes remaining he will be extremely cautious about his play, whereas if he has three fouls at halftime he likely is not concerned with foul trouble while in the game. To account for this, I will use a

variable for the squared number of fouls in the model instead.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.4506	0.0020	732.42	0.0000
scale(abs_scorediff)	0.0079	0.0020	3.96	0.0001
scale(shot_clock)	0.1201	0.0020	60.58	0.0000
scale(sq_fouls)	-0.0004	0.0020	-0.18	0.8550

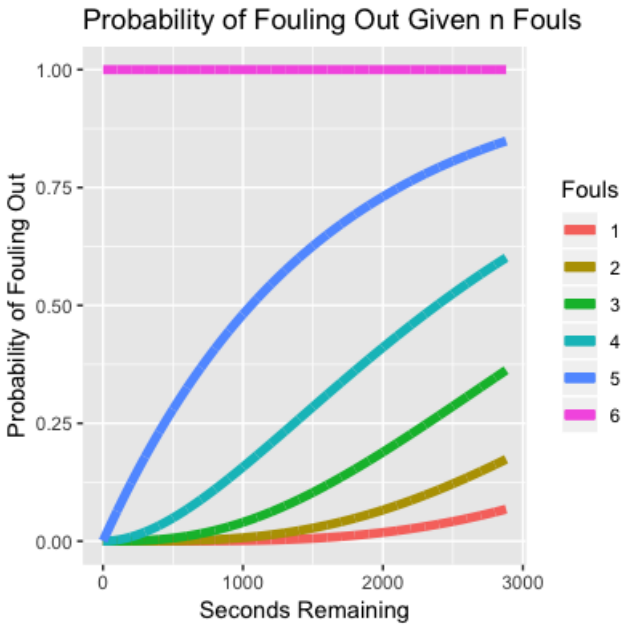
Using squared fouls as a measure of foul trouble does little to improve the model. The coefficients for score differential and shot clock are almost identical to the previous iteration, and the squared fouls coefficient is both minuscule and not close to being statistically significant. Neither of the foul trouble measures used thus far have yielded significant results; however, they were both extremely simplified and did not take into account any other variables that might determine foul trouble status. An extremely glaring omission is the amount of time remaining in the game. A player with three fouls in the last minute of a game is not at all worried about his foul status, as it is virtually impossible to commit three fouls in such a short period of time; on the other hand, one with three fouls in the first quarter has much more reason for concern because there is a lot of time remaining in the game.

3.4 A Markov-chain model of foul trouble

To incorporate time remaining into a measure of foul trouble, I constructed a Markov chain to predict the likelihood of a player fouling out before the end of the game given his current number of fouls and the number of seconds remaining in the entire game. A discrete-time Markov chain is defined by an initial distribution, $\phi(i_0) = P(X(0) = i_0)$, and a transition matrix P, where $p_{ij} = P(X_{t+1} = j | X_t = i)$. In this context, the initial distribution is a one-hot vector indicating the current number of fouls. The transition matrix, in which each state represents the number of fouls a player has, was calculated using 150 games of play by play data from January 2016 and contains the probabilities of the number of fouls that a player will have one second later.[12] These probabilities were estimated using the average number of seconds elapsed between committing the $n-1$ and

[12]The full transition matrix can be found in Table 5 in the appendix

n^{th} fouls for all players in those games. The figure below is an illustration of this model's results, depicting the likelihood that a player fouls out before the end of the game given his number of fouls and the number of seconds remaining in the game.



I then implemented the Markov chain to assign each player at each timestamp a probability of fouling out before the end of the game, and used this as another measure of foul trouble. The regression results employing this variable are as follows:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.4506	0.0020	732.42	0.0000
scale(abs_scorediff)	0.0077	0.0020	3.89	0.0001
scale(shot_clock)	0.1201	0.0020	60.64	0.0000
scale(foulout_prob)	-0.0012	0.0020	-0.61	0.5399

Alas, even this more nuanced definition of foul trouble did not yield a significant result in the model. I also tried adding different variables to the linear model, such as quarter and total clock but neither proved to be significant. Additionally, I subsetted the data into a great deal of ways in an attempt to find a situation in which foul trouble was a significant predictor of distance. Beyond looking at only players guarding the ball, I isolated guards with the thinking that they spend more time on the perimeter and therefore have more physical space with which to adjust their defensive style, while players that play closer to the basket are limited in how much they can back off of their opponent. I also limited

the dataset to plays that ended in jump shots, or shots from outside of the paint. This was due to similar reasoning, namely that plays close to the basket result in players very close together, which might diminish the average distance and obscure variations in this measure. Lastly, I tried looking at only plays in the second half of the games in case the idea of foul trouble being a concern early in the game is overestimated by the Markov model's likelihood of fouling out compared to in the players' minds. Unfortunately, no individual or combination of these filters yielded results showing that a measure of foul trouble significantly explained any variation in defensive distance.

4 Discussion and Conclusions

The fact that none of the analyses conducted in this project did not yield significant results about the effect of foul trouble on defensive aggressiveness leads us to three possible conclusions. The first is that players in fact do not adjust their defensive play in response to their foul trouble status. It is possible that coaches are the ones who worry about players' foul trouble and conduct their substitutions accordingly, while the players continue to play as they would in any situation. Alternatively, there could be an implicit knowledge that the referees are less prone to calling a sixth foul on a player, so the players do not feel very compelled to play differently in order to avoid the possibility of fouling when playing with five fouls.

An alternative conclusion that can be drawn from this project is that the measures of foul trouble that were used did not accurately represent foul trouble in a similar way to how players consider it mentally when deciding how to adjust their play. This could be improved by adding additional elements to the model of foul trouble, such as score differential (closer games will be more tightly called), position (larger players are more likely to have a foul called on them), or quarter (games are refereed differently in the fourth quarter as opposed to the first). This could be done using a Cox proportional-hazards model, which allows the incorporation of arbitrary features in the estimation of survival probability. Since the number of players who do foul out is relatively small, estimating

the survival probabilities would require a larger dataset than the one used in this project.

A third possible conclusion is that my measure of players' defensive aggressiveness does not capture any changes the players make in their defensive play in response to foul trouble. For example, it might be the case that rather than playing further away from their opponents, defenders in foul trouble are less aggressive in other ways such as reaching in less often and keeping their hands farther away or jumping less to contest shots. Both of these tendencies are impossible to capture from the available data. Countless variables factor into the distance between a defender and the player he is matched up with, a few being fatigue, team strategy, the offensive player's shooting ability, the defensive player's quickness. For this reason, the R-squared values of all of the regressions run were extremely low (never greater than .05). I did not expect these values to be high, but this is more evidence that there are lots of factors that defensive distance does not encapsulate. Thus, rather than exploring players' responses to foul trouble through changes in their defensive tendencies or aggressiveness, it might be more telling to examine changes in defensive effectiveness. This would still not incorporate many of the variables just mentioned, but it would capture other changes in defensive tendencies that are not detectable from only looking at the physical distance between players.

One final improvement that could be made to this analysis would be to control for team factors, or even look at player by player differences in changes in defensive aggressiveness in the presence of foul trouble. As I just mentioned, team strategies likely dictate how players defend more so than the players' own decisions. Therefore, much of the variation in defensive distance will be due to team factors, although more data would be necessary for this analysis, as it would make the current dataset extremely sparse.

5 Acknowledgements

I would like to thank Professor Jay Emerson for advising me on this project. Without his constant support, guidance and calming presence throughout the semester, especially when the data seemed unusable, none of this would have been possible. Thank you!

Appendix

Table 1: An example of the player tracking data

player_id	lastname	firstname	team_id	x_loc	y_loc	game_clock	shot_clock	quarter	event.id
200768	Lowry	Kyle	1610612761	10.77643	38.39262	713.26	13.15	1	2
200768	Lowry	Kyle	1610612761	10.77643	38.39262	713.26	13.15	1	3
200768	Lowry	Kyle	1610612761	10.77643	38.39262	713.26	13.15	1	4
202689	Walker	Kemba	1610612766	12.72215	31.93542	713.26	13.15	1	3
202689	Walker	Kemba	1610612766	12.72215	31.93542	713.26	13.15	1	4
202689	Walker	Kemba	1610612766	12.72215	31.93542	713.26	13.15	1	2

Table 2: An example of the play by play data

EventMsgType[a]	EventMsgActionType[b]	Score	quarter	TimeString	HomeDescription	NeutralDescription	VisitorDescription
2	101	NA	1	11:41	MISS Scola 5' Driving Floating Jump Shot	NA	Zeller BLOCK (1 BLK)
4	0	NA	1	11:39	Valanciunas REBOUND (Off:1 Def:0)	NA	NA
1	87	0 - 2	1	11:35	Valanciunas Putback Dunk (2 PTS)	NA	NA
1	48	2 - 2	1	11:25	NA	NA	Zeller 1' Dunk (2 PTS) (Batum 1 AST)

[a]These values are defined in Table 3
[b]Defined in Tables 4 and 5

Table 3: Definitions for EventMsgType

EventMsgType	Definition
1	Make
2	Miss
3	Free Throw (make or miss)
4	Rebound
5	Turnover
6	Foul (any kind)
7	Violation (goaltending, lane violation, kicked ball, delay of game, etc.)
8	Substitution
9	Timeout
10	Jumpball
11	Ejection
12	Start of period (regulation quarter and OT)
13	End of period
18	Instant replay
20	Injury stoppage

Table 4: Definitions for EventMsgActionType

EventMsgActionType	Definition
0	Jump Ball / REBOUND / SUB / Team Rebound
1	Jump Shot / Bad Pass Turnover / Timeout / Delay of game violation / P.FOUL
2	Lost ball Turnover / S.FOUL
3	L.B.FOUL - loose ball foul
4	Traveling Turnover / Off Foul
5	Layup
7	Dunk
8	3 Second Violation Turnover
9	C.P.FOUL
10	Free Throw 1 of 1
11	Free Throw 1 of 2
12	Free Throw 2 of 2
14	FLAGRANT.FOUL.TYPE1
16	Free Throw Technical
17	T.Foul (Def. 3 Sec [player_name])
18	Free Throw Flagrant 1 of 2
19	Free Throw Flagrant 2 of 2
20	Free Throw Flagrant 1 of 1
25	Free Throw Clear Path 1 of 2
26	Free Throw Clear Path 2 of 2
27	Personal Block Foul
28	Personal Take Foul
29	Shooting Block Foul
39	Step Out of Bounds Turnover
40	Out of Bounds Lost Ball Turnover
41	Running Layup
42	Driving Layup
43	Alley Oop Layup
44	Reverse Layup
45	Out of Bounds - Bad Pass Turnover Turnover
46	Running Jump Shot
47	Turnaround Jump Shot
48	Dunk
49	Driving Dunk
50	Running Dunk
52	Alley Oop Dunk
55	Hook Shot
57	Driving Hook Shot
58	Turnaround Hook Shot
63	Fadeaway Jumper
66	Jump Bank Shot
67	Hook Bank Shot
71	Finger Roll Layup
72	Putback Layup
73	Driving Reverse Layup
75	Driving Finger Roll Layup
76	Running Finger Roll Layup
78	Floating Jump Shot

Table 5: Definitions for EventMsgActionType (continued)

EventMsgActionType	Definition
79	Pullup Jump Shot (can be 3pt)
80	Step Back Jump Shot (can be 3pt)
81	Pullup Bank Shot
82	Driving Bank Shot
83	Fadeaway Bank Shot
85	Turnaround Bank Shot
86	Turnaround Fadeaway Shot
87	Putback Dunk
93	Driving Bank Hook Shot
97	Tip Layup Shot
98	Cutting Layup Shot
99	Cutting Finger Roll Layup Shot
100	Running Alley Oop Layup Shot
101	Driving Floating Jump Shot
102	Driving Floating Bank Jump Shot
103	Running Pull-Up Jump Shot
108	Cutting dunk shot

Table 6: Markov chain transition matrix for fouling

	0	1	2	3	4	5	6
0	0.99909	0.00091	0.00000	0.00000	0.00000	0.00000	0.00000
1	0.00000	0.99926	0.00074	0.00000	0.00000	0.00000	0.00000
2	0.00000	0.00000	0.99923	0.00077	0.00000	0.00000	0.00000
3	0.00000	0.00000	0.00000	0.99916	0.00084	0.00000	0.00000
4	0.00000	0.00000	0.00000	0.00000	0.99924	0.00076	0.00000
5	0.00000	0.00000	0.00000	0.00000	0.00000	0.99934	0.00066
6	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	1.00000