The Value of Energy:

Quantifying the Relationship Between Energy Expenditure and Player Performance in the National Basketball Association

A thesis presented

by

Harrison Chase

to

the Department of Statistics and the Department of Computer Science

in partial fulfillment of the honors requirements

for the degree of

Bachelor of Arts

Harvard College

Cambridge, Massachusetts

March 31, 2017

Abstract

This paper uses player tracking data from a subset of games from the 2015-16 NBA season to estimate the relationship between energy expenditure and player performance. Using locational tracking data, several measures of energy expenditure are calculated, along with one baseline measure. To account for each individual player's skill level, a ridge regression is performed and the residuals are used as the dependent variable in a second regression. Binning is used to estimate the relationship between measures of energy expenditure and player performance, and the coefficients from binning are subsequently used to estimate parametric transformations of energy expenditure that are included in the final model. Finally, generalized additive models (GAMs) are used to estimate the relationship between energy expenditure and performance for individual players.

I find three notable results. First, energy expenditure is a significant predictor of a player's deviation from their average skill level. Second, player tracking data can be used to generate measures of energy expenditure that allow for significantly more accurate predictions of player performance than measures not derived from player tracking data. Third, there is strong evidence that players require a period of time when they first enter the game to warm up, during which they perform at a lower level than normal.

Acknowledgements

First and foremost, I would like to thank my thesis advisors, Professor Mark Glickman and Professor Michael Mitzenmacher. I am forever indebted to them for their invaluable guidance and many suggestions throughout the school year. This thesis would not have been possible to write without them, as they truly played a part in both convincing me to write a thesis and helping select suitable topic once I had decided to do so.

I would also like to deeply thank several other people: Alex D'Amour for being gracious enough to talk to me about work he had done in the field and providing the original inspiration for this work; Brittni Donaldson and Charles Rohlf from STATS LLC for not only granting me access to this incredible data, but also providing many suggestions and answering any questions I might have; my parents for their support and encouraging me to write a thesis, as well as for proofreading it instead of enjoying their vacation; and my brothers for pretending to be interested in my thesis when I talked about it to them. Finally, this work is (hopefully) easy to read only because of the additional editing and suggestions from Stephen Yen and Daniel Smith.

Contents

1 Introduction

It is commonly accepted in sports that players perform at a lower level when they are tired. For players in the National Basketball Association (NBA), there are two primary factors affecting their energy levels during a game. First is the amount of rest they have had before the game. It has been shown that when teams play a game two days in a row they perform significantly worse in the second game [Winston, 2012]. The second factor that must be considered when discussing in-game energy levels is how much energy a player has expended in the game prior to that moment. It has been shown in other sports that a player's performance decreases in the second half of a game, the implied cause of which is their fatigue accrued from the first half [Rampinini et al., 2009, Gabbett, 2013]. Despite the prevalent assumption of a relationship, there has been no research done quantifying the exact effect of energy expenditure on player performance.

This paper fills this gap in the literature by using player tracking data from the 2015-16 NBA season to quantify the relationship between energy expenditure and player performance. I use *energy expenditure* to refer to the amount of calories expended by players when they are on the court. I use *player performance* to refer not to a player's nominal performance but rather the deviation of their performance from their average skill level. This paper finds three notable results. First, energy expenditure is a significant predictor of a player's deviation from their average skill level. Second, player tracking data can be used to generate measures of energy expenditure that allow for significantly more accurate predictions of player performance than measures not derived from player tracking data. Third, there is strong evidence that players require a period of time when they first enter the game to warm up, during which they perform at a lower level than normal.

The relationship between energy expenditure and player performance is an important issue for NBA teams. The primary in-game duty of a coach in the NBA is determining which five players are on the court, commonly referred to managing the *rotation* of a team. It is

common sense that a primary factor that a coach takes into consideration when managing the rotation of his team is the energy levels of his players. A star player may be substituted out if he has been on the court for a while and the coach suspects his performance may be so negatively affected by his fatigue that it would be better to play a weaker, but fully rested, player. As there has been no attempt to quantify this relationship, a coach makes decisions like these based on his gut intuition as to how a player's performance is affected by energy expenditure.

The sparsity of the literature on energy expenditure and player performance can likely be attributed to the difficulty of assessing energy expenditure. The most direct way to measure energy expenditure is through rates of oxygen consumption [Passmore and Durnin, 1955]. The traditional method of measuring oxygen consumption is to collect and examine the exhaled air of the subject, but this approach is obviously not suitable to use on NBA players during a game [Patton, 1997]. Thus, this paper draws on existing literature which relates oxygen consumption to velocity and acceleration data in order to estimate energy expenditure.

Only recently has data been collected for the NBA from which it is possible to calculate velocity and acceleration. In 2013, six SportVu cameras which track the position of all ten players (as well as the ball and referees) 25 times a second were installed in every NBA arena. Figure 1 shows the six camera angles and how they are used to triangulate player and ball location.

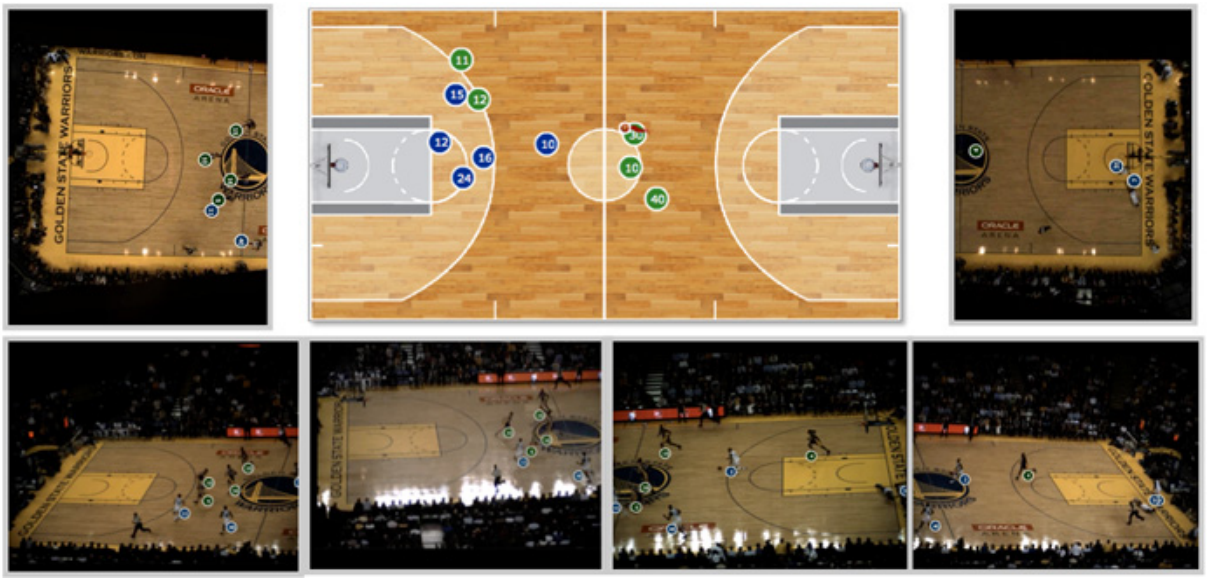

Figure 1: Picture of the SportVu cameras in action, with the view from the six different cameras on the edges and then a digital rendering of the location of the players (and ball) in the center (from stats.nba.com).

The player tracking data that these cameras provide has led to an impressive number of new insights into the sport of basketball. The MIT Sloan Sports Analytics Conference, one of the premier conferences for quantitative sports analysis, has featured work utilizing player tracking data from the SportVu cameras at an increasing rate over the past five years. In 2012, the first conference paper to use this data was presented and won the grand prize [Maheswaran et al., 2012]. Since then, other papers presented have utilized player tracking data to quantify the following aspects of basketball (listed in chronological order): player acceleration [Maymin, 2013], the relationship between offensive rebounding and transition defense [Wiens et al., 2013], interior defense [Goldsberry and Weiss, 2013], player decision-making [Bornn et al., 2014], the "hot-hand" effect [Bocskocsky et al., 2014], rebounding [Maheswaran et al., 2014], recognizing on-ball screens [McQueen et al., 2014], overall defense [Franks et al., 2015], and ball screen defense [McIntyre et al., 2016].

The papers above from this one conference over five years cover a broad spectrum of topics, but as mentioned earlier there has been no research quantifying the relationship between energy expenditure and player performance. This paper is the first to do so. The

structure of the paper is as follows. Section 2 summarizes the related literature regarding the approximation of energy expenditure from velocity and acceleration data, the methods commonly used to estimate player skill, and the results obtained from trying to relate these two variables in other sports. Section 3 discusses the data I use and any data wrangling that is performed, and provides motivation for four measures of energy expenditure that I create along with five transformations of these measures that I use in the final model. Section 4 walks though the creation of the final model and discusses the implications of my findings. Finally, Section 5 concludes with a summary of the paper and a discussion of future work.

2 Previous Work

In this section I highlight three separate areas of existing literature. First, I summarize the literature regarding the estimation of energy expenditure from locational tracking data. Second, I present the methods used to most accurately measure player skill in the NBA. Third, I discuss the paucity of previous literature relating energy expenditure and player performance, not just in basketball but in all sports, highlighting the novelty of this paper.

2.1 Estimating Energy Expenditure

Although previous work using player tracking data in basketball to calculate energy expenditure is nonexistent, several other sports have their own tracking systems, and analyses performed with that data are instructive. In particular, soccer is a sport for which there are comparatively many papers calculating energy expenditure from tracking data, as player movement has been tracked in soccer for a much longer period of time than it has been in basketball [Polglaze et al., 2016]. Even though soccer differs greatly from basketball, papers regarding energy expenditure in soccer are still of value as they provide insight into how to best use locational tracking data to estimate energy expenditure.

When working with tracking data, energy expenditure must be estimated from measures that can be calculated solely from positional data, like velocity, acceleration, and distance. One measure of energy expenditure that has recently risen in popularity is *metabolic power*, defined as the product of energy cost (i.e. how much energy was expended) and velocity [di Prampero, 1986]. Existing research has shown that energy cost does not vary with velocity [Margaria et al., 1963]. However, this relationship was established using subjects who ran at unfluctuating speeds, and therefore is only appropriate to use if constant velocity is assumed [Polglaze et al., 2016]. This assumption is problematic for calculating energy expenditure in professional basketball - prior work done on calculating acceleration from player tracking data data shows a wide range of acceleration vectors [Maymin, 2013]. Indeed, constant

velocity is a bad assumption for most sports, and thus those examining energy expenditure using soccer tracking data have had to overcome this obstacle as well.

One approach to fix the problematic assumption of constant velocity has been to recognize that accelerating on an even surface is equivalent to running at constant velocity up an incline [di Prampero et al., 2005]. When sprinting, the total acceleration acting on the runner's body (written as g') is equivalent to the vectorial sum of forward acceleration (a) and gravity (g).

$$g' = \sqrt{a^2 + g^2} \qquad (1)$$

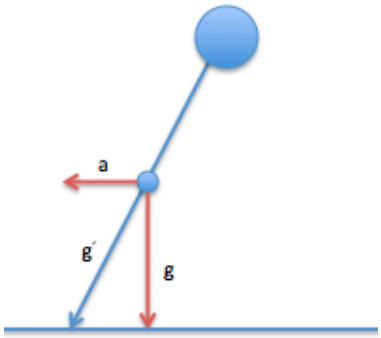

Figure 2: On the left is formula describing the total acceleration working on the runner's body, on the right is a graphical representation of this formula (figure recreated from di Prampero et al. [2005]).

The total acceleration g' acts along both the x axis (due to the sprinter's forward acceleration) and along the y axis (due to gravity). This force can be rotated such that it acts only in the vertical direction, leaving a horizontal force of zero and therefore implying constant velocity. The total acceleration g' is currently acting at an angle of $\arctan(g/a)$, and thus the angle that the vector g' must be rotated by in order to to achieve a horizontal force of zero (written as α) is simply the complement of said angle.

$$\alpha = 90 - \arctan(g/a) \quad (2)$$

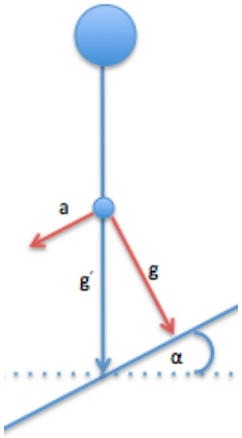

Figure 3: On the left is the formula for the equivalent angle of incline, on the right is a graphical representation from which this formula can be derived (figure recreated from di Prampero et al. [2005]).

With α calculated, the slope of the incline (written as i) can be calculated as

$$i = \tan(\alpha)\,. \quad (3)$$

Running at a constant velocity has been linked to the energy cost of running as a function of the incline upon which one is running. Minetti et al. [2002] measure oxygen consumption of subjects as they run on a treadmill at varying degrees of incline. Their oxygen consumption is then used to calculate the energy cost of running at each specific incline, and the relationship between energy cost and incline is found to be best fit by a fifth degree polynomial. Specifically, they find the energy cost of running (written as c) can be written as a function of the incline (i) in the following manner:

$$c = 155.4i^5 - 30.4i^4 - 43.3i^3 + 46.3i^2 + 19.5i + 3.6\,. \quad (4)$$

di Prampero et al. [2005] suggest that c, multiplied by a constant to reflect that a purely vertical force of g' implies a different mass, can be used to estimate the cost of running at non-constant speeds on an even surface. The multiplicative constant (written as m') can be found by calculating the ratio between m_1 and m_2 such that $m_1 g = m_2 g'$, which after

rearrangement can be found as:

$$m' = \frac{g'}{g} = \sqrt{a^2/g^2 + 1}\,. \tag{5}$$

With the energy cost calculated as cm', metabolic power can be calculated as the product of the energy cost of running and the velocity at which one is running. Thus, metabolic power is calculated as

$$cm'v\,. \tag{6}$$

Osgnach et al. [2010] use the exact procedure outlined above when examining energy expenditure in Serie A soccer matches.

While metabolic power is regarded as the the most complete way to calculate energy expenditure from tracking data, other variables often are used instead, due to ease of calculation. For example, studies have used distance traveled and acceleration as simple proxies for energy expenditure [Varley et al., 2013]. One proxy of energy expenditure derived solely from acceleration values is Vector Dynamic Body Acceleration (VeDBA), which is simply the vector sum of acceleration vectors [Qasem et al., 2012]. Finally, another measure of energy expenditure to be considered is minutes played. The derivation of this measure does not require tracking data and thus serves as baseline for which I can attempt to see how much of a better fit, if any at all, the variables listed above that are derived from player tracking data provide.

2.2 Estimating Player Performance

When quantifying the relationship between player performance and energy expenditure, the identities of the players must be controlled for. I am interested primarily in how a player's deviations from their average skill level can be explained by how much energy they have expended up to that point. If the identities of the players on the court were to not be taken

into account, my analysis would not calculate the relationship between a player's deviation from their average skill level and energy expenditure, but rather the relationship between skill and energy expenditure. Thus, in order to estimate the relationship between energy expenditure and player performance, I must first have a method of estimating the skill of a player.

Estimating player skill in the NBA is a topic of great interest to many, and as such there is comprehensive existing literature on the subject. The methods considered the most complete ways of estimating player skill are derived off of Adjusted Plus-Minus (APM), a methodology first introduced in 2004 [Engelmann, 2017]. In the original analysis, each data point represents a series of possessions where the players on the court are the same. The change in the home team's lead (positive if the home team scores, negative if the away team scores) is the dependent variable. For the independent variables, each player is encoded in their own variable, set equal to 1 if the player is on the court and on the home team, 0 if the player is not on the court, and -1 if the player is on the court and on the away team. The coefficient estimates correspond to player impact.

A natural extension of this model is to compute separate coefficients for a player's offensive and defensive value. This is done by creating separate variables for offense and defense [Ilardi and Barzilai, 2007]. In this approach, players are no longer represented as a single variable but as a set of variables, of which one is equal to 1 when the player is on the court and on offense and zero otherwise, and the other of which is equal to 1 when the player is on the court and on defense and zero otherwise. In this approach, the dependent variable is also changed to be equal to the number of points scored on that possession, and home court advantage is included in an additional variable which is equal to 1 when the home team is on offense and 0 otherwise.

APM proves to be quite predictive, but it has several flaws and was further improved upon in 2010 [Sill, 2010]. Recognizing that APM is prone to overfitting and suffers from collinearity for players who do not play significant minutes, a regularization parameter is

introduced, switching from ordinary least squares regression to *ridge regression*, or "Tikhonov Regularization." When comparing the results via cross validation, ridge regression proves to be far more accurate than least squares regression. This type of approach is not only academically accepted as the most accurate method of estimating player skill, but has also pervaded popular culture. *Regularized Adjusted Plus Minus* (RAPM) is a statistic derived from the above mentioned type of ridge regression and is the base for the most advanced measures of player skill available on ESPN and other major sports media outlets.

2.3 Relating Energy Expenditure to Player Performance

Numerous papers discuss the calculation of variables intended to estimate energy expenditure across many sports, but few attempt to relate energy expenditure to player performance. When Polglaze et al. [2016] summarize recently published papers that involve calculations of energy expenditure from tracking data, many of the analyses listed are mainly comparative, examining the differences in energy expenditure by position or by location on the field. Of the analyses that do attempt to link energy expenditure to other variables, most of them use energy expenditure as the dependent variable, and examine the effect of other variables (score margin, match status, substitution frequency, etc.) on energy expenditure. While analyses in this vein can (and should) be done with player tracking data for basketball, I find it more interesting to examine how energy expenditure influences other variables, such as player performance.

Of the studies mentioned by Polglaze et al. [2016] that attempt to relate energy expenditure to player performance, they primarily relate summary statistics to final season standings [Rampinini et al., 2009, Gabbett, 2013]. By summary statistics, I refer to total distance run per match and other aggregations of energy expenditure measures on a match level. Working with summary statistics ignores the information that one could gain by analyzing player performance within a match alongside their energy expenditure up to that moment. Additionally, the relationship between summary statistics on a match level and how teams end

up doing does not provide insight as to how a player's performance changes with energy expenditure, but rather only implies that players with a certain level of energy expenditure achieve certain results. A small number of studies do attempt to correlate energy expenditure with player performance or tactical decisions within a game. There is evidence that soccer players perform worse technically in the second half compared to the first [Rampinini et al., 2009]. An experiment also finds that basketball players, when made to play two games in a certain time period, perform worse tactically in the second game [Sampaio et al., 2014]. While these studies present quantitative findings that back up common intuition (the intuition being that players play worse when they are tired), they do little to establish a functional relationship between energy expenditure and player performance. Rather, these papers look mainly at performance across large time frames (halves) and do not attempt to relate energy expenditure by players in the first half to their performance decline in the second.

There are several explanations for why there is such a scarcity of literature relating energy expenditure to player performance. First, as mentioned before, tracking systems have been around for a relatively short amount of time, and thus there has not been much time for researchers to analyze the data. Second, the sport for which tracking data has existed the longest (soccer) does not lend itself particularly well to such analysis for various reasons. For one, a goal - the event that affects winning the game most directly, and should therefore reflect player performance - occurs fairly infrequently in soccer. Although other events (passes, tackles, shots) could be used as a proxy for player performance, these are not always perfectly correlated with player performance. For example, tackles are not a perfect representation of defensive performance because when a defender is forced to make a tackle they must have allowed their man to get the ball - a good defender might stop their man from getting the ball in the first place, either with good positioning or by intercepting the pass [Anderson and Sally, 2014]. A second reason why it is difficult to establish a relationship between energy expenditure and player performance in soccer is that there are no clear-cut

periods to compare performance between. The only stop in play is at halftime, and, unlike in American sports like baseball and football, there are not clear stoppages in play between events.

Basketball, however, is a sport that lends itself much better to a quantitative analysis of the effects of energy expenditure. Points are scored very frequently in basketball, so it is possible to use those as a signal of player performance. Basketball can be broken down into possessions, where one team a team's possession ends after they either turn it over or score. Although there is not a pause in the game between events (unlike in football, where teams have to snap the ball, and in baseball, where a pitcher has to throw the ball), studying player performance on a possession basis seems both feasible and informative when working with basketball player tracking data.

3 Data and Methodology

This section outlines the practical and theoretical work needed to model the relationship between energy expenditure and player performance. First I give a detailed description of the data that I use. I then provide formal definitions of four different measures of energy expenditure, and show how they can be calculated from player tracking data. Given these measures of energy expenditure, I go through the process of feature engineering and use these measures to construct five variables that I attempt to relate to player performance in the final model. Finally, I discuss the theoretical approach to modeling player performance as a function of energy expenditure that I take, along with any simplifying assumptions I make.

3.1 Data

I am provided with two types of data from Stats.inc. One is possession data, available from play-by-play. In the original format, I am given a file that contains ten rows per possession. Figure 4 shows the data as provided from Stats.inc, before any data wrangling. For the first ten lines, all of which pertain to the same possession, the only variables that are changing are the $PLAYER\_ID$ and $TEAM\_ID$ variables.

GAME_CODE	OFF_TEAM_ID	DEF_TEAM_ID	POSSESSION_ID	PLAYER_ID	TEAM_ID	PERIOD	SCORE_BEFORE	SCORE_AFTER	SCORE_DEF_TEAM	SHOT_TIME	POSSESSION_RESULT	POINTS
1571617	12	22	1	295436	12	1	0	0	0	702	FG Missed	0
1571617	12	22	1	3235	12	1	0	0	0	702	FG Missed	0
1571617	12	22	1	398142	12	1	0	0	0	702	FG Missed	0
1571617	12	22	1	172643	12	1	0	0	0	702	FG Missed	0
1571617	12	22	1	229598	12	1	0	0	0	702	FG Missed	0
1571617	12	22	1	457688	22	1	0	0	0	702	FG Missed	0
1571617	12	22	1	786397	22	1	0	0	0	702	FG Missed	0
1571617	12	22	1	510790	22	1	0	0	0	702	FG Missed	0
1571617	12	22	1	548610	22	1	0	0	0	702	FG Missed	0
1571617	12	22	1	463121	22	1	0	0	0	702	FG Missed	0

Figure 4: Screenshot of the raw possession data I am provided with.

From this data, I perform brief data manipulation to extract the necessary information for each possession.

GAME_CODE	POSSESSION_ID	START_TIME	END_TIME	POINTS	OFF_1	OFF_2	OFF_3	OFF_4	OFF_5	DEF_1	DEF_2	DEF_3	DEF_4	DEF_5
1570441	1	720	695	0	599809	399612	458730	3494	550991	65834	214152	253997	552679	395374
1570441	2	695	683	2	65834	214152	253997	552679	395374	599809	399612	458730	3494	550991
1570441	3	683	671	0	599809	399612	458730	3494	550991	65834	214152	253997	552679	395374
1570441	4	671	663	2	65834	395374	253997	552679	214152	599809	399612	458730	550991	3494
1570441	5	663	649	0	599809	399612	458730	550991	3494	65834	395374	253997	552679	214152
1570441	6	649	628	2	65834	395374	253997	552679	214152	599809	399612	458730	550991	3494
1570441	7	628	604	2	599809	399612	458730	550991	3494	65834	395374	253997	552679	214152
1570441	8	604	587	0	65834	395374	253997	552679	214152	599809	399612	458730	3494	550991
1570441	9	587	578	2	599809	3494	399612	458730	550991	65834	395374	253997	552679	214152
1570441	10	578	555	2	65834	395374	253997	552679	214152	599809	3494	399612	458730	550991

Figure 5: Screenshot of cleaned data, after calculating the start and end time of each possession, along with the number of points scored and who was on the court at the time.

I am now left with $N = 236,428$ observations in my dataset, where each observation corresponds to a single possession in the 2015-16 NBA season. $M = 431$ different players are on the court for at least one possession.

I am also provided with a sample of the player tracking data. Due to the size of that data, I am not granted access to raw player tracking data for every game in the 2015-16 season but rather only to data for 81 of the 82 games played by the Boston Celtics. The Celtics played one game abroad, in an arena without the SportVu tracking system set up, and thus there exists no data for that game. I am initially given the data in XML format, which means that it is more difficult to visualize than the other raw data I am provided with. However, I can convert the player tracking data to a tabular format so I can store it in a SQLite database. A screenshot of the tabularized data (exported to Excel for aesthetic purposes) is shown in Figure 6.

gameclock	time	shotclock	team_id	player_id	x_loc	y_loc	z_loc	game_id	quarter_id
720	1447632621474	24	25	263899	51.43823	16.45027	0	1571688	1
720	1447632621474	24	25	329830	76.07631	24.64747	0	1571688	1
720	1447632621474	24	25	456450	66.22842	25.27837	0	1571688	1
720	1447632621474	24	25	548635	49.83855	32.83601	0	1571688	1
720	1447632621474	24	25	679478	53.55625	23.96939	0	1571688	1
720	1447632621474	24	2	292407	45.25862	23.73938	0	1571688	1
720	1447632621474	24	2	457605	19.51164	22.94087	0	1571688	1
720	1447632621474	24	2	552381	45.836	34.0113	0	1571688	1
720	1447632621474	24	2	552403	36.99511	25.40845	0	1571688	1
720	1447632621474	24	2	697132	48.94563	15.8758	0	1571688	1

Figure 6: Screenshot of player tracking data I am provided with, after having converted it from XML format into a more database friendly format. Displayed are the x- and y-coordinates for all ten players on the court for a single frame.

The units for the x- and y-coordinates are in feet. The size of the court in the NBA is

required to be 94 by 50 feet, so the x-coordinates range mostly between 0 and 94 (with 47 being at half court) and the y-coordinates between 0 and 50. Note that some observations fall outside of that range, as occasionally players run or fall out bounds. By connecting these coordinates on a player level, each player's path throughout the game can be tracked. Figure 7 shows the path of five defensive players during a single possession, plotted in different colors so one can distinguish between the players.

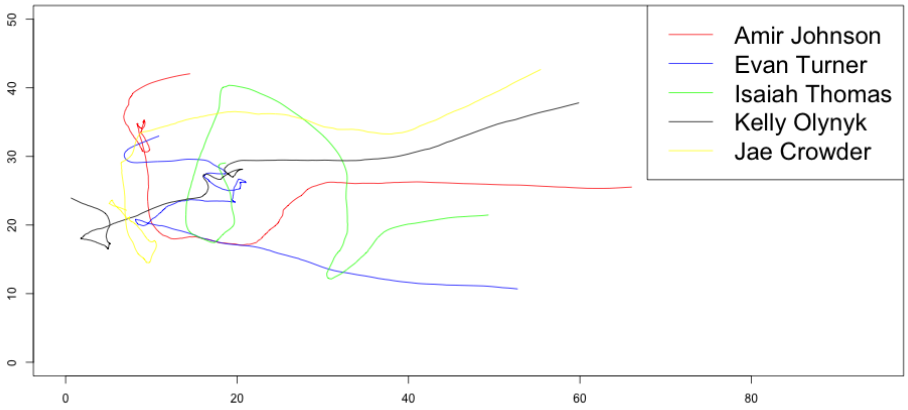

Figure 7: Traces of player movement for a single defensive possession, where each colored path represents a different player.

3.2 Variables

I now discuss the different variables that I create in my attempt to model player performance as a function of energy expenditure. These fall into two broad categories. The first I refer to as *measures of energy expenditure*, which are variables I derive from the player tracking data to be a proxy for energy expenditure. These measures of energy expenditure are computed on a possession level. Theoretically, these measures could be computed on a more granular level, such as per second or per half second. Such a level of granularity is unnecessary for my analysis, however, as I am exploring the relationship between energy expenditure and player performance on a possession-by-possession basis. The second type of variables that I create are transformations of these measures of energy expenditure for use in the final regression,

referred to in this paper as *regression variables*. These variables are created based on my domain knowledge and what I hypothesize to be most predictive.

3.2.1 Measures of Energy Expenditure

Using the player tracking data, I compute four different measures of energy expenditure. The reason for working with multiple measures is to determine which measures allow for the closest fitting relationship between energy expenditure and player performance. Although these four measures of energy expenditure are not the only measures one could compute, they are the ones that are most commonly used in existing literature when attempting to estimate energy expenditure from locational tracking data.

3.2.1.1 Time

An extremely simple measure of energy expenditure is the time a player has spent on the court, or played so far in the game, hereafter referred to as *Time*. It should be the case that energy expenditure is positively correlated with Time. For a given game with N possessions, let s_n be the start time of the nth possession (in terms of the gameclock) and e_n be the end time of the nth possession. I create two separate N by M matrices to encode which players are on the court, differentiating between offense and defense, such that one (I^O) reflects the status of players in the following way:

$$I^O_{n,m} = \left\{ \begin{array}{ll} 1, & \text{if player } m \text{ on court and on offense for possession } n \\ 0, & \text{if player } m \text{ not on court and on offense for possession } n \end{array} \right\} \tag{7}$$

and the other (I^D) reflects the status of players in the following way:

$$I^D_{n,m} = \left\{ \begin{array}{ll} 1, & \text{if player } m \text{ on court and on defense for possession } n \\ 0, & \text{if player } m \text{ not on court and on defense for possession } n \end{array} \right\}. \tag{8}$$

Creating two matrices, one for offense and the other for defense, may seem unnecessary

here but are useful later on.

Player m is on the court for possession n when $I^O_{n,m} + I^D_{n,m} = 1$, and if he is not on the court then $I^O_{n,m} + I^D_{n,m} = 0$. Using this notation I defined the time player m spent on the court during the nth possession (written $t_{n,m}$) as:

$$t_{n,m} = (s_n - e_n) \times (I^O_{n,m} + I^D_{n,m}) \tag{9}$$

Time is intended to be a baseline measure, as it is easily computable from play-by-play and thus player tracking data is not needed.

3.2.1.2 Distance

A slightly more complex measure of energy expenditure is distance run by a player when he is on the court, hereafter referred to as *Distance*. For each player I am given x_t and y_t, their x and y coordinates on the court at time t. Rather than working with the provided locational data points, a moving average with a window of Δ is taken, similar to the work done by Maymin [2013]. Later in this paper an appropriate value of Δ is calculated. New vectors x' and y' are created, defined as:

$$\begin{aligned} x'_t &= \frac{1}{\Delta} \sum_{i=t-\Delta}^{t} x_t \,, \\ y'_t &= \frac{1}{\Delta} \sum_{i=t-\Delta}^{t} y_t \,. \end{aligned} \tag{10}$$

The change in distance from time t to $t+1$ is then written as D_t and defined as:

$$D_t = \sqrt{(x'_{t+1} - x'_t)^2 + (y'_{t+1} - y'_t)^2} \,. \tag{11}$$

These changes in distance are summed up for each possession. For a possession that started at s_n and ended at e_n, the distance run for that possession (written as D'_n) is defined as:

$$D'_n = \sum_{t=s_n}^{e_n} D_t \,. \tag{12}$$

3.2.1.3 Vector Dynamic Body Acceleration

As described in Section 2, Vector Dynamic Body Acceleration, hereafter referred to as *VeDBA*, is simply the vector sum of accelerations [Qasem et al., 2012]. In order to calculate acceleration, I first calculate velocity along both the x and y axes at time t, written as $v_{x,t}$ and $v_{y,t}$ and defined as:

$$\begin{aligned} v_{x,t} &= (x'_t - x'_{t-1}) \times 25 \,, \\ v_{y,t} &= (y'_t - y'_{t-1}) \times 25 \,. \end{aligned} \tag{13}$$

The multiplication by 25 is necessary as the time between observations is $\frac{1}{25}$ of a second and velocity should be converted to a more intuitive ft/s scale.

Acceleration along both the x and y axes at time t is then calculated, written as $a_{x,t}$ and $a_{y,t}$ and defined as:

$$\begin{aligned} a_{x,t} &= (v_{x,t} - v_{x,t-1}) \times 25 \,, \\ a_{y,t} &= (v_{y,t} - v_{y,t-1}) \times 25 \,. \end{aligned} \tag{14}$$

Again, note the multiplication by 25 to convert it to a common ft/s^2 scale.

From there, the VeDBA at time t, written as V_t is defined as:

$$V_t = \sqrt{a_{x,t}^2 + a_{y,t}^2} \tag{15}$$

.

These measures of VeDBA are summed up for each possession. For a possession that started at s_n and ended at e_n, the VeDBA for that possession (written V'_n) is defined to be:

$$V'_n = \sum_{t=s_n}^{e_n} V_t \tag{16}$$

.

3.2.1.4 Metabolic Power

Metabolic power was originally derived as an estimate of energy expenditure for sprinters running in a straight line [di Prampero et al., 2005]. Although it is used by Osgnach et al. [2010] to estimate energy expenditure among soccer players, they failed to properly account for the two dimensional aspect of acceleration in sports. They calculate a singular velocity vector, v_t, as:

$$v_t = \sqrt{(x_t - x_{t-1})^2 + (y_t - y_{t-1})^2} \tag{17}$$

from which they calculate a single acceleration vector, a_t, as:

$$a_t = v_t - v_{t-1}\,. \tag{18}$$

However, this methodology is inaccurate in certain situations. If a player transitions from moving at velocity V in the x direction and velocity 0 in the y direction to moving at velocity 0 in the x direction and velocity V in the y direction, the above method would calculate an acceleration of 0, while clearly a substantial amount of acceleration is being used along both axes. For sports than involve a large amount of cutting and changing velocity along multiple axes at the same time, the methodology used by Osgnach et al. [2010] is imprecise. While soccer does not involve a lot of these actions, basketball is sport where these types of movements occur much more frequently and thus this issue should be addressed.

Therefore, I have modified the method used by Osgnach et al. [2010] to account for the duality of dimensions. For a given measurement at time t, the current accelerations along the x- and y-axes are equal to running at constant speed up slopes of angle $i_{x,t}$ and $i_{y,t}$

respectively. These angles are calculated as:

$$\begin{aligned} i_{x,t} &= \tan(90 - \arctan(g/a_{x,t}))\,, \\ i_{y,t} &= \tan(90 - \arctan(g/a_{y,t}))\,. \end{aligned} \tag{19}$$

The energy costs of accelerating along the x- and y-axes ($c_{x,t}$ and $c_{y,t}$) are then calculated as:

$$\begin{aligned} c_{x,t} &= 155.4i_{x,t}^5 - 30.4i_{x,t}^4 - 43.3i_{x,t}^3 + 46.3i_{x,t}^2 + 19.5i_{x,t} + 3.6\,, \\ c_{y,t} &= 155.4i_{y,t}^5 - 30.4i_{y,t}^4 - 43.3i_{y,t}^3 + 46.3i_{y,t}^2 + 19.5i_{y,t} + 3.6\,. \end{aligned} \tag{20}$$

Furthermore, for each of the calculated inclines equivalent mass ratios much be calculated, and thus $m'_{x,t}$ and $m'_{y,t}$ are written as:

$$\begin{aligned} m'_{x,t} &= \sqrt{a_{x,t}^2/g^2 + 1}\,, \\ m'_{y,t} &= \sqrt{a_{y,t}^2/g^2 + 1}\,. \end{aligned} \tag{21}$$

Finally, the metabolic power used during one time period can be written as P_t and defined as:

$$P_t = P_{x,t} + P_{y,t} \tag{22}$$

where

$$\begin{aligned} P_{x,t} &= c_{x,t}m'_{x,t}v_{x,t}\,, \\ P_{y,t} &= c_{y,t}m'_{y,t}v_{y,t}\,. \end{aligned} \tag{23}$$

These bursts of metabolic power are summed up for each possession. For a possession that started at s_n and ended at e_n, the Metabolic Power used for that possession (written P'_n) is:

$$P'_n = \sum_{t=s_n}^{e_n} P_t\,. \tag{24}$$

3.2.2 Regression Variables

After creating four measures of energy expenditure for all players during all possessions of the game, I create variables of which I can model player performance as a function of. For a given possession, the variables for which I am attempting to relate to player performance should only include information known prior to that possession. When a coach is deciding whether to substitute a player in for another one, at the very most he can only know energy expenditure data up to that point in time. Therefore, introducing variables that utilize energy expenditure information that occurs after the fact is not be helpful as one could not know that before making a decision.

There are many variables one can create from energy-expenditure information prior to a possession. The ones considered in this paper are listed below. In order to create these features, it is helpful for the sake of notation to create a new matrix $I' = I^O - I^D$, where $I'_{n,m}$ indicates a player m's involvement during each possession n:

$$I'_{n,m} = \left\{ \begin{array}{ll} 1, & \text{if player } m \text{ on court and on offense during possession } n \\ 0, & \text{if player } m \text{ not on court during possession } n \\ -1, & \text{if player } m \text{ on court and on defense during possession } n \end{array} \right\}. \tag{25}$$

As there are four different measures of energy expenditure, these regression variables are calculated four times, once for each measure of energy expenditure. In the following formulas for the transformed variables, the placeholder variable $X_{n,m}$ is used as a generic representation of the energy expenditure by player m during possession n.

3.2.2.1 Total Energy Expenditure

The first featured considered is perhaps the most intuitive. **Total Energy Expenditure** is simply the summation of energy expenditure in the game. Specifically, for a player m his total energy expenditure at possession n of a game can be computed as:

$$T_{n,m} = \sum_{i=1}^{n-1} X_{i,m} \, . \tag{26}$$

Note that Total Energy Expenditure corresponds to how much energy a player has used prior to possession n. The energy expenditure during possession n should not be included as that would not be known prior to the possession.

3.2.2.2 Consecutive Energy Expenditure

There are several flaws with Total Energy Expenditure, primarily that it does not capture the order in which energy expenditure occurred. For example, a player who expends a lot of energy and then is subbed out and sits on the bench would be expected to have a different energy level from a player who starts off on the bench and then is subbed in. Therefore, a second feature is created: **Consecutive Energy Expenditure**, which is the sum of continuous energy expenditure, where continuous is defined as being without breaks. The most obvious break in action that players have is when they are substituted out, and thus Consecutive Energy Expenditure should reset whenever a player is taken out of the game. Other natural breaks to consider are halftime and the break in between quarters. However, the break in between quarters is only two minutes long, much shorter than the break in between halves, which is fifteen minutes long. Thus, it is unclear whether those breaks should be treated the same. Due to the uncertainty here, two variables are created, one which resets only at halftime and the other which resets between all quarters. To distinguish between these two, they are referred to as **Consecutive Energy Expenditure (Halftime)** and **Consecutive Energy Expenditure (Quarters)** and written as $C^H_{n,m}$ and $C^Q_{n,m}$ respectively. I define them as follows:

$$C^H_{n,m} = \left\{ \begin{array}{ll} C^H_{n-1,m} + X_{n-1,m}, & \text{if } I'_{n-1,m} \neq 0 \text{ and } n \text{ is not the first possession of a half} \\ 0, & \text{if } I'_{n-1,m} = 0 \text{ or } n \text{ is the first possession of a half} \end{array} \right\} \tag{27}$$

$$C^Q_{n,m} = \left\{ \begin{array}{ll} C^Q_{n-1,m} + X_{n-1,m}, & \text{if } I'_{n-1,m} \neq 0 \text{ and } n \text{ is not the first possession of a quarter} \\ 0, & \text{if } I'_{n-1,m} = 0 \text{ or } n \text{ is the first possession of a quarter} \end{array} \right\}. \tag{28}$$

Note that for the creation of this variable, the break before overtime periods is treated the same as the break between quarters.

3.2.2.3 Rest

One component that is currently missing is the amount of rest a player has before they are subbed on. Intuitively, I expect that players on shorter rest perform worse than those on longer rest, all else being equal.

I define **Rest** ($R_{n,m}$) as the amount of rest player m got before the shift that possession n is part of. Rest is defined as time elapsed from the game clock in seconds. During halftime an extra 900 seconds (15 minutes) are added on, and during quarters 130 seconds are, as these are the lengths of breaks between the quarters. The Rest of each player at the start of the game is defined as ∞. As $R_{n,m} \in (0, \infty)$, it is more apt to work with the transformation $\exp(-\frac{R_{n,m}}{60})$, which lies in the range from 0 to 1. The division by 60 was performed to transform the units from seconds to minutes.

3.2.2.4 Exponential Moving Average of Energy Expenditure

It is reasonable to hypothesize that recent possessions influence a player's energy level more than previous ones. As mentioned in Section 2, there is a paucity of research examining

in-game physical exhaustion, and thus there is not much precedent to fall back on. As such, I rely on the research regarding workloads on a day-to-day or week-to-week basis. Research has shown that when attempting to model "acute" and "chronic" work loads inbetween days, exponential weighted moving averages are more appropriate than simple moving averages [Williams et al., 2017]. Thus, a new variable called **Exponential Moving Average of Energy Expenditure** is created. For a given player m, his Exponential Moving Average of Energy Expenditure before a particular possession n ($E_{n,m}$) can be written recursively as a function of $X_{n,m}$, a player's energy expenditure during a given possession, and ρ, the decay parameter as follows:

$$\begin{aligned} E_{1,m} &= 0 \\ E_{n,m} &= X_{n-1,m} + \rho E_{n-1,m} \,. \end{aligned} \tag{29}$$

Additionally, a variable (H_n) is created to reflect whether possession n is the first possession of the second or fourth quarter or the first possession of the second half. The inclusion of H_n seems appropriate as one would expect possessions that occur in another quarter (or half) to have less of an effect than ones in the same in the same quarter, even after the exponential discounting (due to the rest that occurs during those breaks). The breaks between quarters are treated differently than the break at halftime since halftime is longer than the breaks between quarters. H_n is defined as follows:

$$H_n = \left\{ \begin{array}{ll} \gamma_1, & \text{if } n \text{ is the first possession of the second or fourth quarter, or any overtime periods} \\ \gamma_2, & \text{if if } n \text{ is the first possession of the second half} \\ 0, & \text{if if } n \text{ is not the first possession of the second or forth quarter or the second half} \end{array} \right\} . \tag{30}$$

The variable H_n then incorporated in the recursion as follows:

$$
\begin{aligned}
E_{1,m} &= 0 \\
E_{n,m} &= X_{n-1,m} + \rho^{H_n+1} E_{n-1,m} \, .
\end{aligned}
\tag{31}
$$

One potential flaw of defining Exponential Moving Average of Energy Expenditure as in Equation 31 is that not all possessions are the same length. Therefore, I am losing information by discounting possessions that occurred the same number of possessions but different number of seconds ago the same way. I can incorporate the length of the possession (written as $s_n - e_n$) by altering the recursion in the following way:

$$
\begin{aligned}
E_{1,m} &= 0 \\
E_{n,m} &= X_{n-1,m} + \rho^{H_n+s_{n-1}-e_{n-1}} E_{n-1,m} \, .
\end{aligned}
\tag{32}
$$

To find optimal values of ρ, γ_1 and γ_2, a grid search can be performed.

3.2.2.5 Score Differential

A final variable is included that is not a transformation of energy expenditure. The margin of the game during possession n, from the point of view of the team that has the ball, is written as δ_n and referred to as **Score Differential**. Previous work has shown the margin of the game to be a significant predictor of the outcome of the possession [Engelmann, 2011]. A nearly perfectly linear effect is found between the margin and outcome of the possession, with teams that are ahead performing worse than expected and teams that are behind performing better than expected. An argument can be made that δ_n should be included in the initial ridge regression that is used to estimate player skill, as is described in Section 3.3. However, I am concerned that part of the effect of energy expenditure may be captured by Score Differential, as when teams are ahead they may put in their substitutes while when behind they may keep their starters in. Thus, Score Differential is included in the final regression as opposed to the initial one.

3.3 Methods

The first task is to construct estimates of player skill in a RAPM framework. A similar approach to the methodology used in Ilardi and Barzilai [2007] is taken. Each possession corresponds to a singular observation in the dataset. The response variable Y is an N-by-1 vector, where the entry Y_n corresponds to the points scored on the nth play in my dataset.

For the independent variables, each player is encoded in two variables, as I wish to estimate the effect of offense and defense separately. One of these variables is set equal to 1 if the player is on the court and on offense, and zero otherwise, while the other is set equal to 1 if the player is on the court and on defense, and zero otherwise. In order to control for home court advantage, an additional independent variable h_n, is created, set equal to 1 if the team on offense is also the home team, and zero otherwise. For each observation, there are a total of $2M + 1$ independent variables: two independent variables for each of the M players, and an additional one for home court advantage. For N observations, the data matrix containing the independent variables for all observations is written as I and has dimensions N-by-$2M + 1$. Note that I can be constructed as the column concatenation of the previously defined I^O and I^D (in Equations 7 and 8 respectively) and an additional column encoding home court advantage.

As I am using ridge regression, I am seeking to minimize the function:

$$f(\beta, \lambda) = \sum_{n=1}^{N} (I_n \beta - Y_n)^2 + \lambda \sum_{j=1}^{2M+1} \beta_j^2 \,. \tag{33}$$

$\beta^\star$ is the set of coefficients that minimize Equation 33, where $2M$ of them are estimates of player skill and the final coefficient is an estimate of home court advantage. Residuals are then calculated for each data point, where the residual for possession n (r_n) is written as:

$$r_n = Y_n - I_n \beta^\star \,. \tag{34}$$

These r_ns are then used as the dependent variable for the regression involving the various measures of energy expenditure. The hope is that one or more of those measures explains the variation of the residuals.

Through the ridge regression a static measurement of a player's skill is computed as the coefficient of the ridge regression ($\beta^\star$). That is to say, if I define S_j to be the skill of player m on either offense or defense (as offense and defense skill are estimated separately), then, as currently defined,

$$S_j = \beta_j^\star \,. \tag{35}$$

Of interest is how S_j varies with energy expenditure. However, I have yet to define the relationship between S_m and energy expenditure. The first simplifying assumption I make is that the relationship between energy expenditure and skill is the same for all players, which is necessary because for most of the players (all but those on the Boston Celtics) I have only four games worth of data on them.

As my primary interest lies in quantifying the relationship of energy expenditure and player performance, the only modeling techniques that are considered are ones where the relationships of independent variables to the dependent variable can be easily summarized. For this reason, models that may allow for more predictive accuracy (like gradient boosted decision trees and neural networks) are eschewed in favor of regression based models.

As currently defined, I have four separate measures of energy expenditure, computed for all possessions: Time, Distance, VeDBA and Metabolic Power. I also have five variables created from those measures: Total Energy Expenditure (T), Consecutive Energy Expenditure (Halftime) (C^H), Consecutive Energy Expenditure (Quarters) (C^Q), Rest (R) and Exponential Moving Average of Energy Expenditure (E), along with one variable not derived from energy expenditure: Score Differential (δ). One might also hypothesize that the relationship between energy expenditure and player performance may depend on whether the player is on offense or defense, in which case the I' matrix created earlier is useful. In

considering a theoretical model for how a player's skill at possession n ($S_{n,j}$) would depend on energy expenditure, two approaches are considered. Note that even though a player's skill is denoted with a subscript j, measures of energy expenditure still uses subscript m. This is because a player's skill depends on whether he is on offense or defense, while measures of energy expenditure do not (although their relationship with skill do).

$$S_{n,j} = \beta_j^\star + f(T_{n,m}, C_{n,m}^H, C_{n,m}^Q, R_{n,m}, E_{n,m}, I'_{n,m}, \delta) \tag{36}$$

and

$$S_{n,j} = \beta_j^\star \times f(T_{n,m}, C_{n,m}^H, C_{n,m}^Q, R_{n,m}, E_{n,m}, I'_{n,m}, \delta)\,. \tag{37}$$

For ease of notation, $f(T_{n,m}, C_{n,m}^H, C_{n,m}^Q, R_{n,m}, E_{n,m}, I'_{n,m}, \delta)$ can be written as $f(\omega)$.

The first approach described in equation 36 is more appropriate given the hypothesized nature of $f(\omega)$: since I am assuming that $f(\omega)$ is the same for all players, the first derivative of $f(\omega)$ with respect to ω is the same for all players. However, $\beta_j^\star$ not only varies for players but also takes on both positive and negative values. Using the relationship $S_{n,j} = \beta_j^\star \times f(\omega)$ would imply that the derivative of $S_{n,j}$ with respect to ω would act in opposite directions for players with positive $\beta_j^\star$ compared to ones with negative $\beta_j^\star$. However, this is not consistent with my approach, as I make the simplification that all players have similar relationships with energy expenditure. Thus, the first approach is taken, and skill is modeled as:

$$S_{n,j} = \beta_j^\star + f(\omega)\,. \tag{38}$$

As mentioned above, I am forced to attempt to model the residuals of the initial regression (r_n) rather than points scored directly, as I do not have player tracking data for all games. Recall that

$$r_n = Y_n - I_n\beta^\star\,. \tag{39}$$

One can recognize that theoretically, as points scored should be modeled by a player's skill

and any effect from home court advantage,

$$Y_n = \sum_{j=1}^{2M} S_{n,j} I_{n,j} + \beta^{\star}_{2M+1} h_n. \tag{40}$$

Substituting in our equation for player skill from Equation 38:

$$Y_n = \sum_{j=1}^{2M} (\beta^{\star}_j + f(\omega)) I_{n,j} + \beta^{\star}_{2M+1} h_n. \tag{41}$$

Simplifying:

$$Y_n = \sum_{j=1}^{2M} \beta^{\star}_j I_{n,j} + \sum_{j=1}^{2M} f(\omega) I_{n,j} + \beta^{\star}_{2M+1} h_n. \tag{42}$$

One can realize that $\beta^{\star} I_n = \sum_{j=1}^{2M} \beta^{\star}_j I_{n,j} + \beta^{\star}_{2M+1} h_n$ by vector multiplication. Substituting in, one gets:

$$Y_n = \beta^{\star} I_n + \sum_{j=1}^{2M} f(\omega) I_{n,j} + \beta^{\star}_{2M+1} h_n. \tag{43}$$

Rearranging terms:

$$\begin{aligned} Y_n - \beta^{\star} I_n &= \sum_{j=1}^{2M} f(\omega) I_{n,j} \\ \rightarrow r_n &= \sum_{j=1}^{2M} f(\omega) I_{n,j} \, . \end{aligned} \tag{44}$$

This justifies my approach of estimating $f(\omega)$ based on the residuals from the prior regression.

4 Results

In section I discuss the results I find from working with the data. I first show why using a ridge regression as opposed to an OLS regression is necessary when estimating player skill. I then find an optimal smoothing window for the player tracking data and estimate various hyperparameters. I then perform brief data exploration, showing the correlations between the four measures of energy expenditure and how player performance varies with respect to time in the game. After that, I detail how parametric nonlinear relationships between energy expenditure and player performance are estimated using a technique known as binning. Final models are created for the four measures of energy expenditure, and then cross validation is used to estimate prediction errors. Finally, individual players' relationships between energy expenditure and performance are estimated for the players for whom enough data exists. Discussion of the significance of my findings is withheld until the end of the section, where I show that:

- Player performance is significantly affect by energy expenditure.
- Player tracking data can be used to calculate more predictive measures of energy expenditure than before.
- All measures of energy expenditure show similar relationships between energy expenditure and player performance, from which the actionable insight that players should warm up more before entering the game can be gleaned.

4.1 Estimating Player Skill

Following the methodology laid out in Engelmann [2017], I use generalized cross validation to find the optimal ridge constant in Python. For the data I am given for the the 2015-16 season, I find the ridge constant that leads to the lowest cross validation error to be 3600,

consistent with previous literature which found it be around 3000 [Engelmann, 2017]. The error curve from the generalized cross validation is plotted in Figure 8.

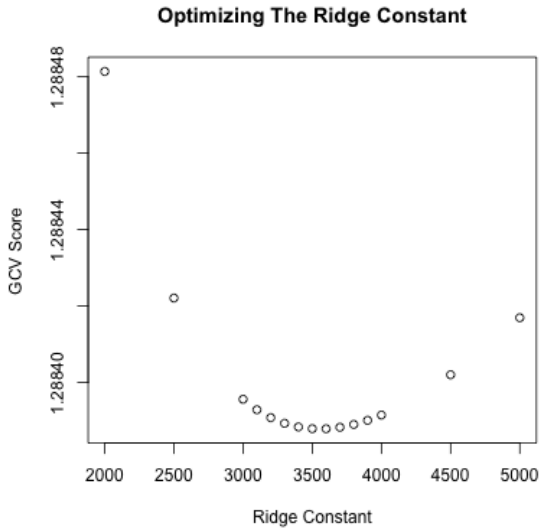

Figure 8: A plot of GCV scores by the ridge constant, which reaches a minimum around 3600.

The coefficients from the ridge regression give estimates for player skill on both sides of the ball. Table 1 lists the top five and bottom five players for both offense and defense.

Offensive Rankings

Top Five	Bottom Five
Draymond Green	Jahlil Okafor
Stephen Curry	Wayne Ellington
Russell Westbrook	Roy Hibbert
LeBron James	Marco Belinelli
Kevin Durant	Kevin Seraphin

Defensive Rankings

Top Five	Bottom Five
Tony Snell	Julius Randle
Kawhi Leonard	Markel Brown
Steven Adams	Brandon Rush
Jeremy Lamb	Terrence Jones
Draymond Green	Sean Kilpatrick

Table 1: Rankings of player skill as estimated as the coefficients a ridge regression.

As an example of why ridge regression is necessary, compare the results listed in Table 1 to the results of an ordinary least squares (OLS) regression. The top and bottom five players on each side of the ball, as estimated from the OLS regression, are reported in Table 2.

Offensive Rankings

Top Five	Bottom Five
Coty Clarke	Jimmer Fredette
Thanasis Antetokounmpo	Chuck Hayes
Rakeem Christmas	Luis Montero
Jorge Gutierrez	Nazr Mohammed
Jordan McRae	Bruno Caboclo

Defensive Rankings

Top Five	Bottom Five
Keith Appling	Rakeem Christmas
Pat Connaughton	Joe Harris
Elton Brand	Dahntay Jones
Shayne Whittington	Kevon Looney
Steve Novak	Thanasis Antetokounmpo

Table 2: Rankings of player skill as estimated as the coefficients an OLS regression.

Ideally, it would be possible to know the true skill level of each player and compare the accuracy of rankings generated from the two regression methods. However, as it is impossible to know the true skill level of each player, I instead compare these rankings to human-based rankings to see which of the regression technique generates rankings that better line up with what a knowledgeable viewer of the game might expect. The best form of human-based rankings are the Most Valuable Player (MVP) and Defensive Player of the Year (DPOY) voting results. There is unfortunately not an Offensive Player of the Year award and thus there is no voting results judging solely offensive skill. However, the MVP award is typically biased towards strong offensive players and thus can be thought of as a reflection of player's offensive skill [Johnson, 2015].

The MVP and DPOY awards are voted on by a panel of journalists and broadcasters who cover the sport. For MVP, each voter casts five votes, weighted to give more weight to higher ranked players on a voter's ballot (the votes are worth ten points, seven points, five points, three points, and one point respectively). For the DPOY, each voter casts three votes, which are also weighted in a similar way (the votes are worth five points, three points, and one point respectively). The ten players with the most MVP and DPOY points for the 2015-16 are listed in Table 3 (along with their point total).

MVP	DPOY
Stephen Curry (1,310)	Kawhi Leonard (547)
Kawhi Leonard (634)	Draymond Green (421)
LeBron James (631)	Hassan Whiteside (83)
Russell Westbrook (486)	DeAndre Jordan (50)
Kevin Durant (147)	Paul Millsap (21)
Chris Paul (107)	Avery Bradley (14)
Draymond Green (50)	Rudy Gobert (13)
Damian Lillard (26)	Tony Allen (5)
James Harden (9)	Anthony Davis (4)
Kyle Lowry (6)	Andre Drummond (3)

Table 3: Voting results for the 2015-16 NBA season, as reported by nab.com.

It is clear based on the above information that the rankings obtained via ridge regression more closely track human judgement than those obtained from OLS. The top five offensive players from ridge regression are all among the top seven vote getters for the MVP award, while none of the top five from OLS even received a vote. Meanwhile, two of the top five defensive players from ridge regression are among the top ten in DPOY voting, while none of the top five from OLS are.

As further proof that ridge regression produces more reasonable results, plots of the coefficient estimates from each regression against the number of possessions played are displayed in Figure 9. In ridge regression, the estimates for players who have played a small number of possessions are shrunk towards zero, while the estimates for players who have played more possessions are more variable and generally stronger. Shrinkage in the estimates for players who have played minimal possessions is desirable as there is more information on players who have played many possessions, so one can be more certain in estimates of larger magnitude. Meanwhile, the OLS regression produces highly variable coefficient estimates for those players who have played a small number of possessions, but it is unlikely that any player playing so few possessions is really the best in the league.

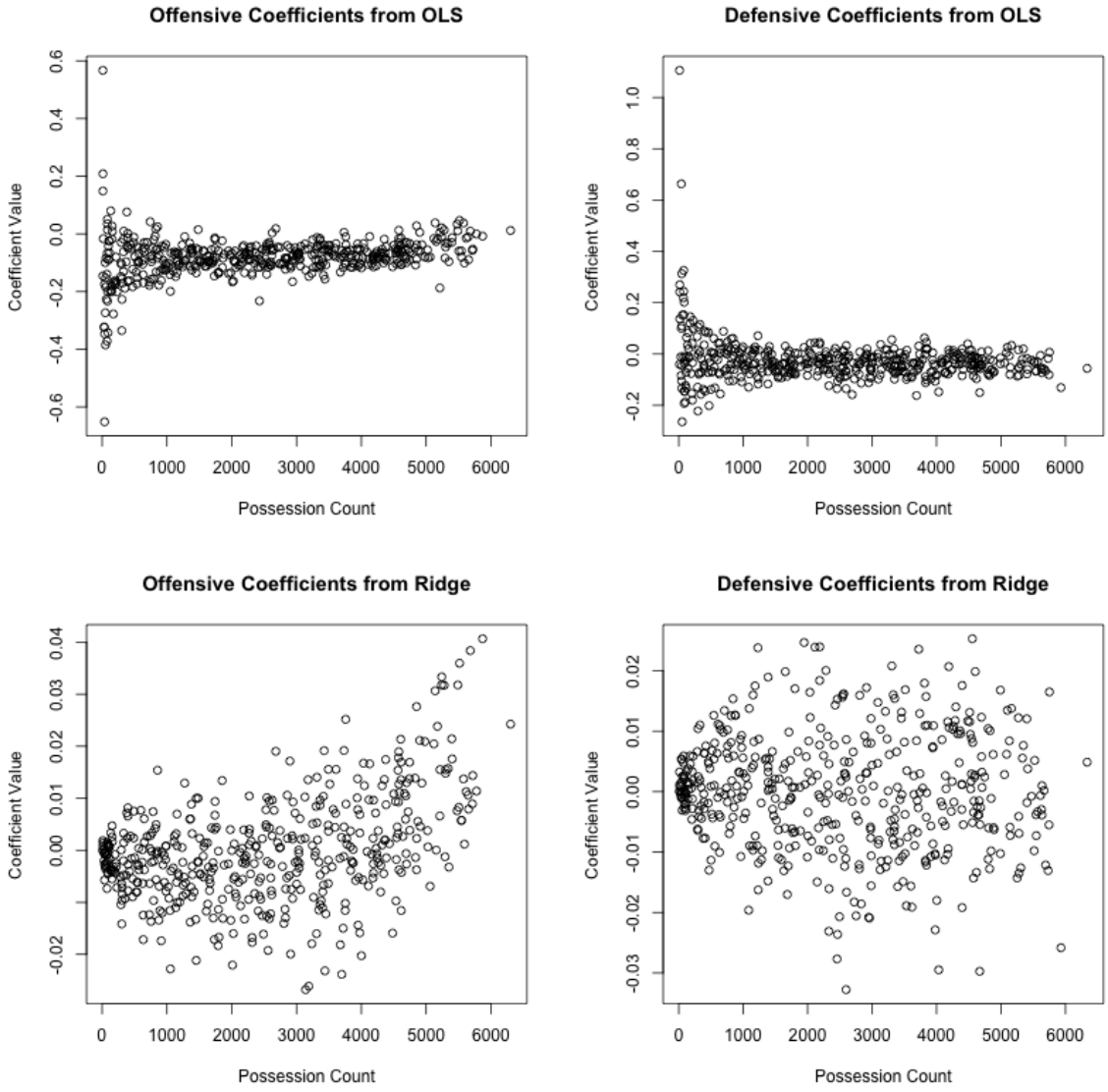

Figure 9: Plots of coefficient estimates I find from doing an OLS regression (top row) compared to a ridge regression (bottom row).

4.2 Smoothing of Player Tracking Data

Earlier work smooths the player tracking data to eliminate the noise in the data [Maymin, 2013]. A similar approach is taken in this paper. In order to determine the proper window over with data should should be smoothed (hereafter refered to as the *smoothing constant*) a game is chosen at random to examine the percent of acceleration values that are within a certain threshold of reasonableness for different smoothing constants. The threshold of

reasonableness is determined to be $9.5m/s^2$ (or $32.144ft/s^2$), which was the measured acceleration of Usain Bolt coming off of the starting blocks in his record setting 100-m dash in 2009 [Dvorsky, 2013]. It is unlikely that any NBA players are accelerating faster than Usain Bolt.

Figure 10 shows a plot of the percent of acceleration values that are considered valid after smoothing the locational data with various smoothing constants.

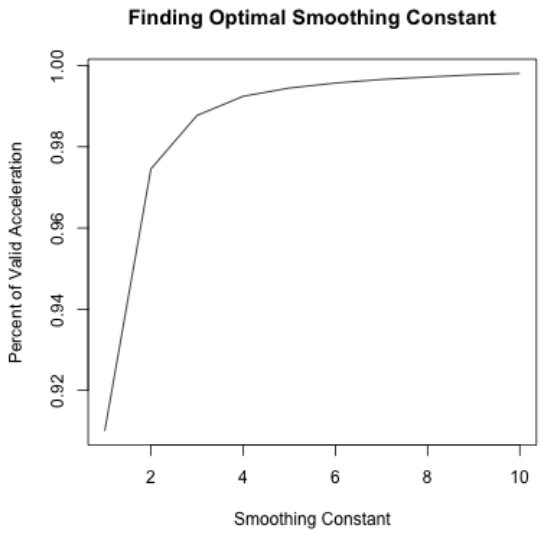

Figure 10: In an attempt to find the optimal smoothing constant, the percentage of realistic acceleration values was plotted as a function of the smoothing window.

The percentage of acceleration vectors deemed to be reasonable begins to asymptote around a smoothing constant of five, and thus five is chosen to be the smoothing constant in this paper in the hope that it would be large enough to eliminate most of the unreasonable values but at the same time capture most of players' true movement.

4.3 Dealing With Outliers

The smoothing of the player tracking data eliminates most of the unreasonable acceleration values, but there is still a small percentage of values that are above what could be considered a reasonable value. These values are curtailed down to be equal to $32.144ft/s^2$, earlier

determined to be an upper bound on reasonable acceleration values.

Additionally, the regression Minetti et al. [2002] performed in order to model the energy cost of running as a function of incline of the slope was only for slopes ranging in incline from -.45 to .45. To avoid extrapolating from the regression line that was fit in that model, slopes that are calculated to be outside of that boundary are curtailed to be either -.45 or .45, depending on which extreme of the boundary they fall.

4.4 Correlation between measurements

While four measures of energy expenditure are calculated, I expect them all to be relatively similar as they are attempting to capture the same information. Figure 11 shows a correlation matrix of these four measures of energy expenditure.

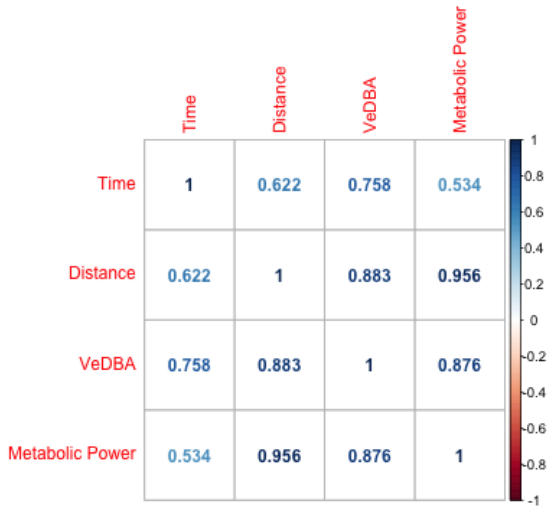

Figure 11: Correlation matrix of the four measures of energy expenditure that are considered in this paper.

All four of these measures are highly correlated, but it should stand out that Time is significantly less so than the other three. The magnitudes of these correlations both make sense and are encouraging because the other three measures are all derived from the tracking data, while Time can be computed solely from play by play. Therefore, measures of energy

expenditure that require player tracking data to calculate differ from previously computable measures of energy expenditure.

4.5 Optimization of Exponential Moving Average of Energy Expenditure

There are three parameters to optimize with respect the Exponential Moving Average of Energy Expenditure, as detailed in Section 3.2.2.4. They are: ρ, the discounting weight; γ_1, the discounting multiplier for quarter breaks; and γ_2, the discounting multiplier for halftime. These parameters are optimized through a grid search that tries to optimize the fit of a model containing only the Exponential Moving Average of Energy Expenditure variables. Although a grid search, which utilizes a discrete and limited state space, is not the most accurate method of estimating hyperparameters, this method was used due to computational and time constraints. Offense and defense effects are treated differently, and a model is created with these two variables. The R^2 values for models fit for all combinations of $(\rho, \gamma_1, \gamma_2)$ are recorded, and the combination that produced the highest R^2 is chosen. A grid search is performed for all four estimates of energy expenditure. The optimized values for the three variables for all types of energy expenditure are recorded in Table 4.

	Time	Distance	VeDBA	Metabolic Power
ρ	0.985	0.94	0.985	0.9475
γ_1	0	0	400	0
γ_2	500	500	700	500

Table 4: Optimized values of hyperparameters for Exponential Moving Average of Energy Expenditure via CV.

The optimal values of the hyper-parameters provide insight to the relationship between energy expenditure and player performance. Time and VeDBA arrive at the same estimate of ρ, while Distance and Metabolic Power arrive at estimates that are nearly the same. Equivalent values within these two groups are not unexpected as Figure 11 shows that

the variable Time is most correlated with VeDBA, while Distance is most correlated with Metabolic Power. Three of the four measures arrive at an optimal estimate of zero for γ_1, suggesting that the break between quarters provides little to no rest. On the other hand, the estimates for γ_2 are estimated to be quite large, suggesting halftime provides a long rest, essentially resetting a player's energy. The large estimate for γ_2 relative to γ_1 is an intuitively expected result as the halftime break is substantially longer than the quarter break.

4.6 Data Exploration

It is helpful to explore the data visually in order to get an idea of how performance might be related to with energy expenditure. As I expect player performance to vary over time, and my measure of player performance is derived from points scored, I visualize the relationship between points per possession (PPP) and time in the game. To create this visualization, a generalized additive model (GAM) is fit using the *gam* package in R, and the effect curve of time in game is plotted. The curve is fit for only regulation period possessions (i.e. not overtime) and first for all games, and then for just the subset of games for which I have the tracking data.

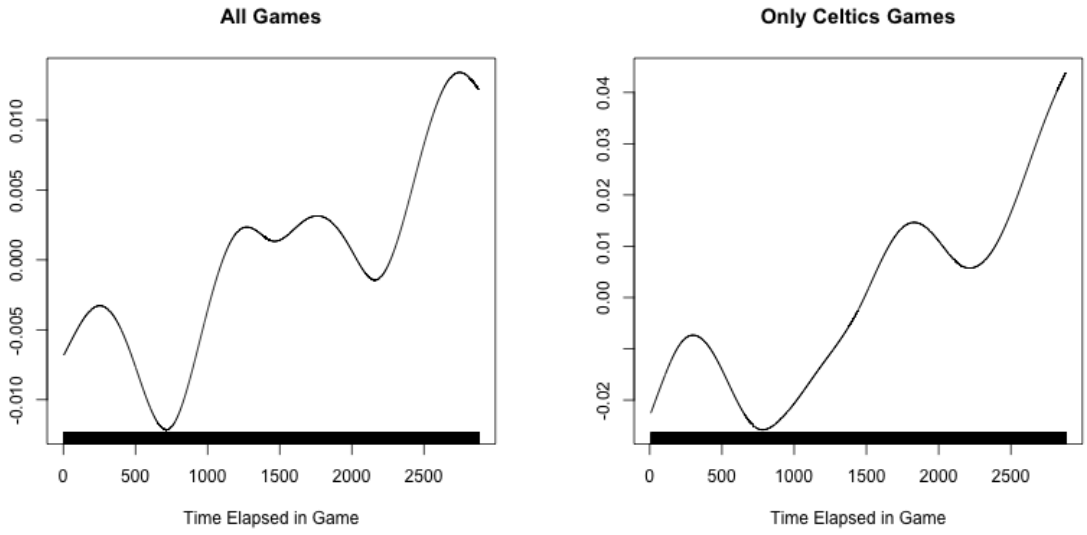

Figure 12: Relationship between points scored and time in game, before controlling for player skill.

Points scored are dependent on who is in the game. In an attempt to isolate the effect of just the time in the game, the residuals from the ridge regression are used rather than PPP, and once again the effect curve is plotted.

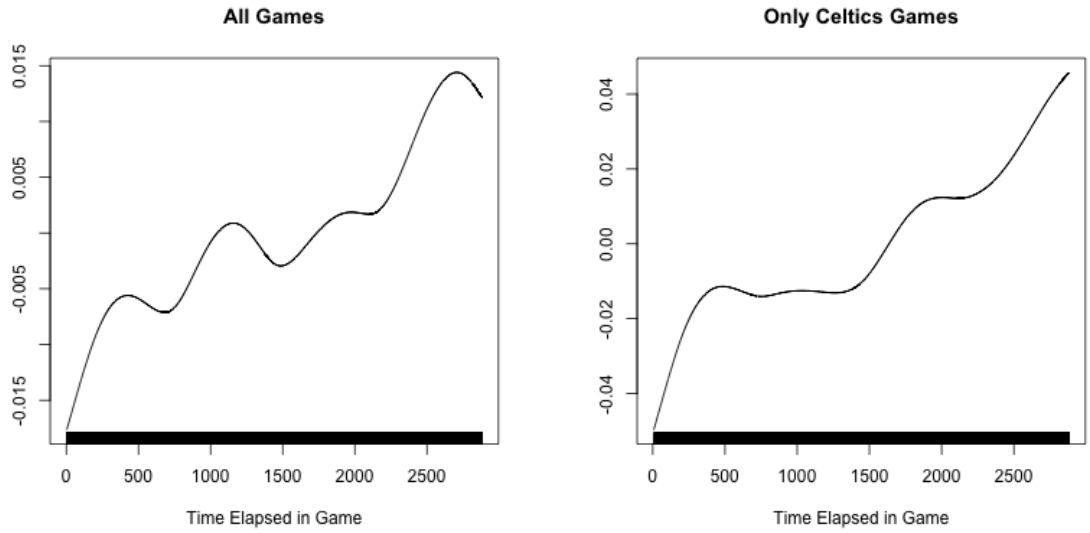

Figure 13: Relationship between points scored and time in game, after controlling for player skill.

The two curves tell similar stories. Even after controlling for the players on the court, scoring rates increase throughout the game. There are, however, noticeable exceptions to this trend: towards the end of the halves the scoring takes a downswing, and towards the ends of the quarters it levels out or even decreases. Additionally, the period of time where there is the sharpest increase in points scored per possession is right at the beginning of the game. Ideally, the final model is able to explain both the continuous increase in PPP throughout the game as well as the sharp increase in PPP at the start of the game.

4.7 The Model

The model regresses the residuals from the ridge regression on a function of five variables derived from player tracking data (T, C^H, C^Q, R and E) and one that is not (δ). It is simplest to first attempt to model them as being additive:

$$f(\omega) = f_1(T_{n,m}) + f_2(C^H_{n,m}) + f_3(C^Q_{n,m}) + f_4(R_{n,m}) + f_5(E_{n,m})\,. \tag{45}$$

It is unrealistic to assume that the relationship between the individual variables and player performance is linear. Normally, a generalized additive model (GAM) would be used to estimate a nonlinear relationship. However, this approach is not feasible when assuming that all players have the same relationship between energy expenditure and player performance, as a GAM would estimate a separate $f(\omega)$ for each player - I want all the $f(\omega)$ to be the same. Thus, in order to estimate the nonlinear relationships, a different approach known as binning is used.

4.7.1 Using Binning to Estimate Nonlinear Relationships

Binning involves transforming a quantitative variable into several categorical variables. For example, if I was to transform total minutes played into a categorical variable with 4 bins, instead of representing the total minutes played by a player as a number, I would represent it as a categorical variable in a bin of either 0-12 minutes, 12-24 minutes, 24-36 minutes or 36-48 minutes. This procedure can be done with any number of bins. One downside of binning is that information is lost, but this practice is still often performed on large datasets for computational purposes [Blower and Kelsall, 2002]. It is common practice for, once a variable is assigned to a bin, each bin to be represented by a discrete 1/0 indicator variable [Eriksson et al., 2013]. I treat offensive and defensive effects separately, and as there are five offensive players I create bins ranging not from zero to one but rather from zero to five, where the value of the bin is equal to the number of offensive players whose energy expenditure places them in that bin. I repeat the same process for the five defensive players. Binning is performed for the five different regression variables. The values of the bins are used collectively as the independent variables in a regression attempting to model the residuals from the initial model. With five variables of energy expenditure, considering two relationships for each variable (offense and defense), if each binning procedure divides

the data into n bins then the number of independent variables totals $5 \times 2 \times n = 10n$.

Bins are created such that across all observations all bins have as equal a number of players in them as possible. One issue that arises is that there are always a number of entries that are zero. When the number of bins are high enough, there may be a disproportionate number of zero entries compared to other bins. To account for this issue, a separate bin is set aside for zero entries. This bin still suffers from the same issue of having a disproportionate number of entries in certain cases, but the non-zero bins now have an equal number of observations.

Multiple numbers of bins are tried out. The number of bins used shows the tradeoff between bias and variability. Using less bins decreases the variability but increases the bias, while more bins increases the variability but decreases the bias. As an example, Figure 14 graphs the relationships discovered using 10 bins and using 100 bins for Consecutive Energy Expenditure (Halftime) for offense. Additionally, as the coefficient estimates of the bins still have significant variance, I use GAMs to calculate a smoothed curve estimating the relationship. To deal with the issue that the bin reflecting the zero entries may have a different number of observations than others, the GAM uses weights proportional to the number of observations in each bin.

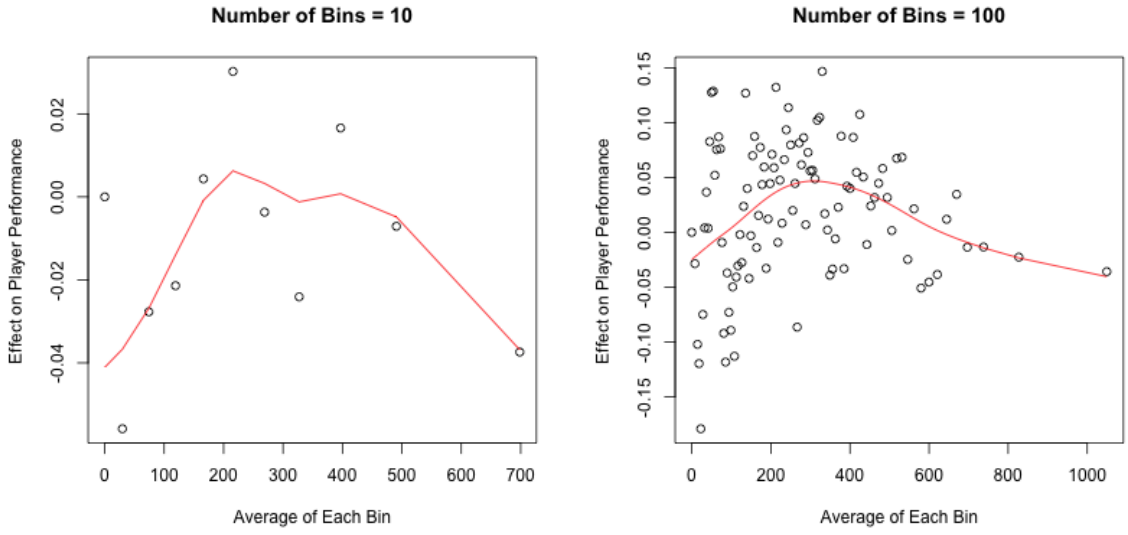

Figure 14: Comparing the relationship of Consecutive Energy Expenditure (Halftime) for offensive player performance that I find when using a different number of bins.

Using a different number of bins produces similar, but distinct, results. In order to estimate what the optimal number of bins would be, the Adjusted R^2 of the models fit by binning are recorded and plotted in Figure 15.

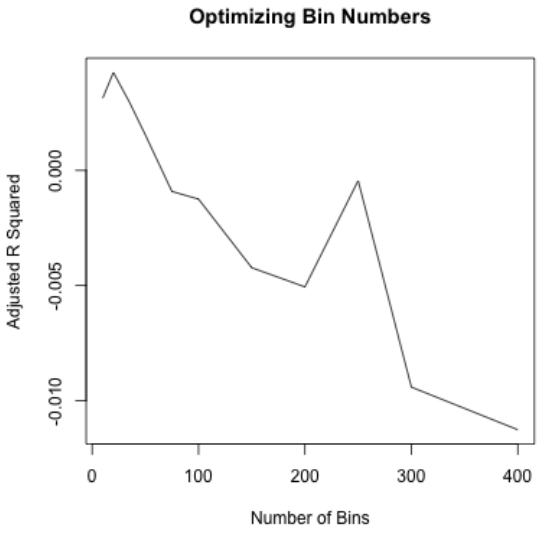

Figure 15: Plotting the fit of the model, as measured by adjusted R^2, by the number of bins used.

After a small increase, increasing the number of bins appears to sharply lower the Adjusted R^2 of the model. For this paper I set the number of bins equal to 20, as that is the number that produced the optimal fit.

I estimate the individual relationships for every variable, both on offense and defense, using binning. As an example, Figure 16 shows plots of the coefficients calculated for the bins of Consecutive Energy Expenditure (Quarters) on offense and defense.

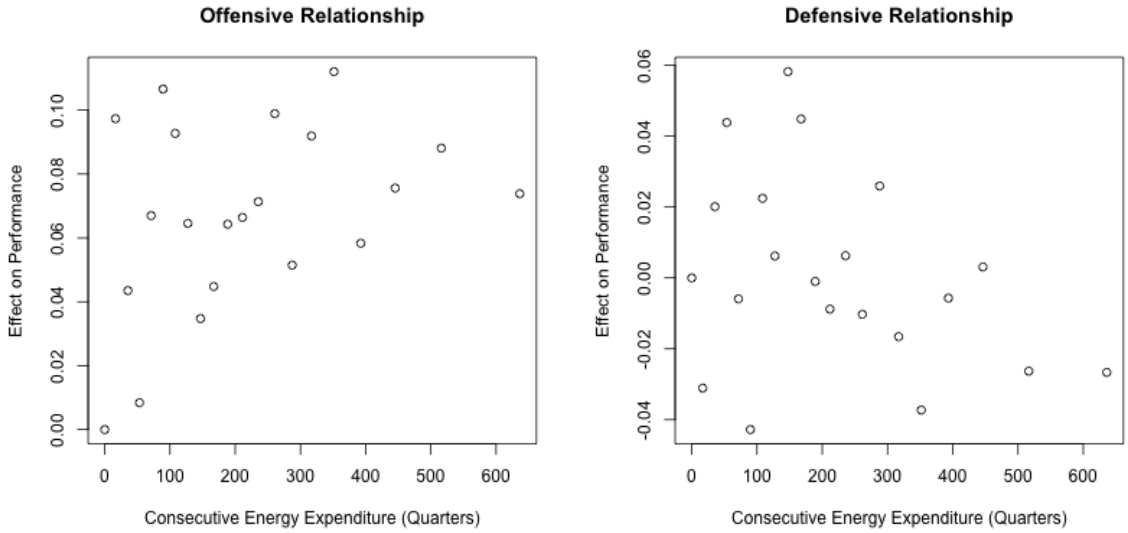

Figure 16: Plotting coefficients for the bins of Consecutive Energy Expenditure (Quarters) when using Time as the measure of energy expenditure.

I fit a combination of a linear, quadratic, and logarithmic terms to each relationship in order to parametrically model the general relationship shown in Figure 16. To find the most appropriate combination, a series of F tests are used to compare the fit of various models. To continue the example above, Figure 17 plots the best fitting models over the plots of the coefficients. For offense, the best model consists of solely a logarithmic term, while for defense it consists of solely a quadratic term.

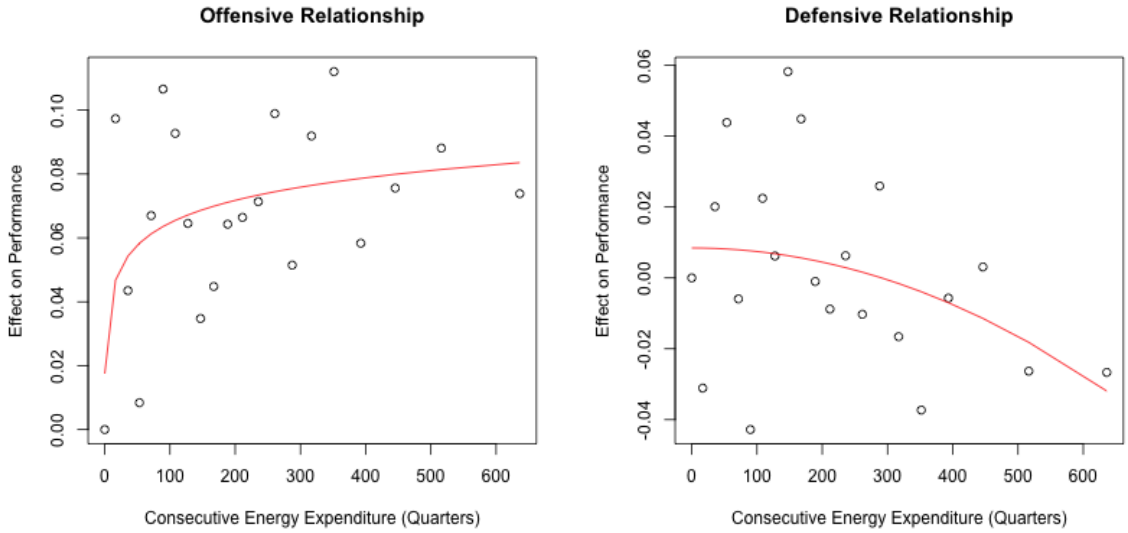

Figure 17: Attempting to model the relationship shown by the coefficients of the bins parametrically.

An identical procedure is repeated for the four other variables, treating offense and defense separately. The graphs for those relationships can be found in Appendix A. Table 5 marks which transformations (linear, quadratic, and logarithmic) I find to best model the coefficients for all variables. A check signifies the transformation of the column is included when modeling the variable in that row.

Variable	Linear	Quadratic	Logarithmic
C^H (Offense)		✓	
C^H (Defense)		✓	✓
T (Offense)	✓	✓	✓
T (Defense)		✓	
C^Q (Offense)			✓
C^Q (Defense)		✓	
E (Offense)	✓	✓	✓
E (Defense)	✓	✓	✓
R (Offense)			✓
R (Defense)			✓

Table 5: Which transformations I find to best model the relationship shown by the coefficients from binning when using Time as the measure of energy expenditure.

I repeat this process but using the three other measures of energy expenditure. The

graphs showing the coefficients when using Distance as the measure of energy expenditure can be found in Appendix B, while those created when using VeDBA as the measure of energy expenditure can be found in Appendix C, and finally those created when using Metabolic Power as the measure of energy expenditure can be found in Appendix D. The tables summarizing which transformations are included when attempting to model the coefficients parametrically for each measure of energy expenditure can be found in Appendix E. The transformations that best model each relationship are very similar across measures of energy expenditure, which is encouraging as it suggests all measures tell a similar story.

4.7.2 Model Selection

After the work above, I now have estimates of what parametric relationships best describe the relationship of each variable to player performance. I can now attempt to use these relationships on the quantitative, unbinned data to create my final model. One additional trivial transformation of the data was performed: a value of 1 is added to all variables so that in the case of a logarithmic transformation the variable would be defined. Without this transformation, logarithms of zero would have been taken, producing values of $-\infty$, which are not suitable for regression. To determine which of the variables are needed to accurate model player performance, a model selection process is performed.

In order to select the final model, a stepwise selection procedure initialized at the null model is performed. In addition to Score Differential (δ), there are five variables of energy expenditure (C^H, T, C^Q, E, R) being considered, of which the offensive and defensive effects are considered separately, leading to a total of 11 variables to consider. In Section 4.7.1, I find many of the relationships for these variables to best be described by a combination of transformations: i.e. both a quadratic and logarithmic transformation best describe the relationship with C^H on defense. I perform the stepwise selection process considering those two transformations jointly - either both would be included or neither would be.

The criterion used to perform the stepwise selection procedure is the Akaike Information

Criterion (AIC), which seeks to minimize the following equation:

$$2k - 2ln(\hat{L}) \tag{46}$$

where

- $\hat{L}$ = the maximum likelihood.
- k = the number of free parameters to be estimated.

Under this criterion, the inclusion of a variable faces the tradeoff of increasing the maximum likelihood while also increasing the product $2k$. This tradeoff means that variables that do not increase the maximum likelihood by a significant enough amount are not included.

This procedure is performed for all four measures of energy expenditure. The coefficients for each model are listed in Table 6, along with the p-value from the F-statistic for each model (the result of an ANOVA test comparing the constructed model to the null model). Note that for the reported coefficients $(*)$ represents significance at the 0.05 level, $(**)$ represents significance at the 0.01 level, and $(***)$ represents significance at the 0.001 level.

Variable	Time	Distance	VeDBA	Metabolic Power
Margin	-2.768e-03**	-2.738e-03**	-2.832e-03**	-2.806e-03**
C^H (Offense)		-7.174e-06		
$(C^H)^2$ (Offense)	-7.736e-08**		-7.736e-08**	-4.724e-14
$\log(C^H)$ (Offense)		-4.498e-03		
C^H (Defense)		1.309e-05*		9.041e-08*
$(C^H)^2$ (Defense)	4.939e-08		4.939e-08	
$\log(C^H)$ (Defense)	-1.375e-02 ***	-1.023e-02***	-1.375e-02***	-1.475e-02***
T (Offense)	-1.013e-04***	-1.164e-05*	-1.013e-04***	
T^2 (Offense)	3.815e-08***	8.837e-10*	3.815e-08 ***	
$\log(T)$ (Offense)	2.494e-02 ***	1.575e-02***	2.494e-02***	5.724e-03***
T (Defense)				
T^2 (Defense)				
$\log(T)$ (Defense)				
C^Q (Offense)				
$(C^Q)^2$ (Offense)				
$\log(C^Q)$ (Offense)	1.141e-02***	1.141e-02***	5.352e-03**	4.540e-03**
C^Q (Defense)		-1.697e-05		
$(C^Q)^2$ (Defense)	-9.323e-08*		-9.323e-08*	-1.179e-13**
$\log(C^Q)$ (Defense)				
E (Offense)	1.177e-04	4.154e-04*	1.177e-04	1.185e-06
E^2 (Offense)	1.003e-05	-3.390e-07	1.003e-05	-2.693e-12
$\log(E)$ (Offense)	-3.097e-02*	-2.500e-02**	-3.097e-02*	-1.355e-02***
E (Defense)				
E^2 (Defense)				-3.406e-12
$\log(E)$ (Defense)				1.388e-02**
R (Offense)				
R^2 (Offense)				
$\log(R)$ (Offense)				
R (Defense)				
R^2 (Defense)				
$\log(R)$ (Defense)				
p-value from F-statistic	3.55e-12	1.121e-10	2.91e-10	5.472e-12

Table 6: A coefficient table for the final models fit for the four different measures of energy expenditure.

I find that all four models include some transformation of the same six variables, and then the model created when using Metabolic Power as the measure of energy expenditure contains an additional variable. In addition to simply reporting the coefficients of the variables, I can also plot each relationship individually, to help visualize the different relationships that I find in the four different models.

Plotting the offensive relationships first:

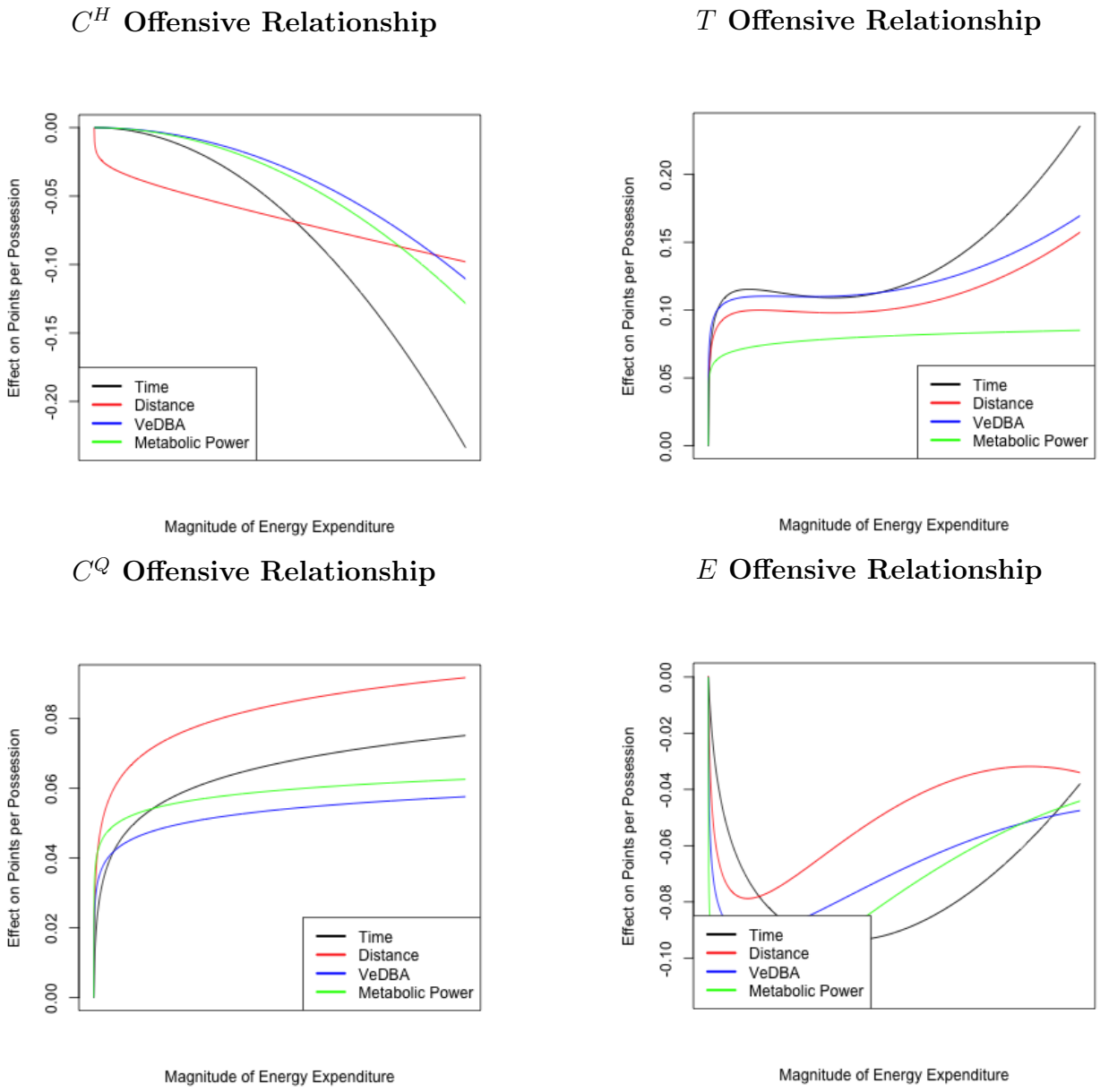

Figure 18: Visualizing the relationships that I find best relate offensive player performance to different transformations of energy expenditure.

I refrain from commenting on the implications of the relationships shown in Figure 18 until Section 4.10 of the paper, but of note is that I find very similar relationships when using the four different measures of energy expenditure. Similarity among relationships is expected as the measures of energy expenditure are positively correlated, but is also reassuring as this means I can be more confident in the general direction of my findings.

Now, plotting the defense relationships:

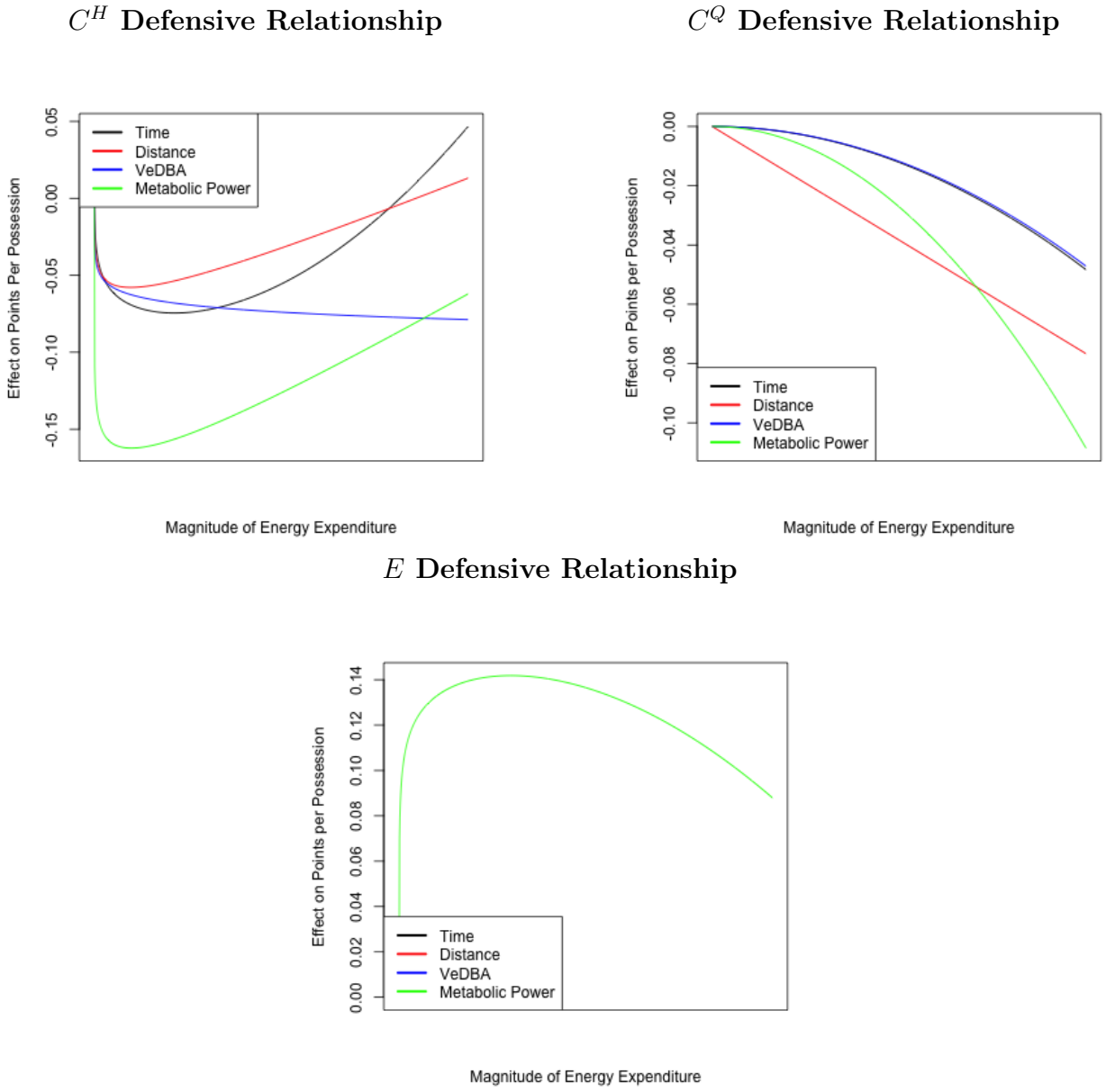

Figure 19: Visualizing the relationships I find to best relate defensive player performance to different transformations of energy expenditure.

Again, elaborate discussion is withheld until Section 4.10, but once again all measures of energy expenditure reflect the same general relationship with player performance.

4.8 Estimating Prediction Errors

One of the primary purposes of this work is to determine whether it is possible to better estimate player fatigue with variables captured by the tracking data than with preexisting data, which is why I calculate four measures of energy expenditure. One of these (Time) is available from play-by-play data, and does not require the player tracking data. The other three (Distance, VeDBA and Metabolic Power) are only able to be calculated with the player tracking data.

To determine if measures of energy expenditure derived from player tracking data allows one to more accurately model player performance, I estimate prediction errors for all four measures. To do this, the data is split into two subsets, a learning set and a test set. When splitting the data into a test and training set, the data is not split by individual observation but rather by game, due to both practical and theoretical considerations. Practically, one would get a game's worth of data altogether, as the data is processed after the conclusion of each game and then delivered to teams. Theoretically, each split would have an equivalent number of observations taken early and late in the game, as some variables (like T) are strictly increasing throughout the game. A simple method of assuring an even distribution of observations from the early game and the late game across the learning and test set is to keep all the data points from the same game together. Thus, the dataset is split by games. The number of games assigned to the test set is chosen to be 9, which is $\frac{9}{81} = 11.11\%$ of the available games.

The same model selection process as outlined in Section 4.7.2 is performed on the data in the learning set, and the model chosen is then used to predict the test set. The mean square error (MSE) of the predictions is recorded. This process is performed once using each of the four measurements of energy expenditure (Time, Distance, VeDBA and Metabolic Power),

making sure the same learning set and test set are used with each measure. This procedure is then repeated 1000 times. The average prediction errors using each measure of energy are listed in Table 7, along with the average prediction error from a baseline model using only an intercept.

Measurement of Energy Expenditure	Average Prediction Error
None (Base Model)	1.2719
Time on Court	1.2685
Distance Run	1.2707
VeDBA	1.2698
Metabolic Power	1.2680

Table 7: Prediction errors for different measures of energy expenditure.

In addition to reporting the average MSE, I am also interested in how the errors compare to each other. To examine whether one set of errors is significantly lower than another I used a paired t-test, as the errors are correlated with each other, since the same learning/test splits are being used. Although the nominal differences between prediction errors are small, they all differ from each other at a highly significant ($< 10^{-10}$) level due to the highly correlated nature of them, as well as the number of iterations that are performed. The implications of these prediction errors are discussed later in the Section 4.10.

4.9 Individual Player Curves

The major simplifications made in the final model are that all players have the same relationship regarding energy expenditure and player performance and that these relationships can be parametrically defined. I attempt to relax this assumption and quantify non-parametric relationships on an individual player level by using GAMs to estimate the nonlinear relationship for players. One of the complications of this approach is that tracking data is only available for Celtics games and thus I have at most four games worth of data on non-Celtics players. Modeling relationships for players for whom I do not have a significant amount of data would lead to highly variable relationships. To address this issue, I only attempt to

model individual relationships for the twelve Celtics players for whom I have the most data (each with more than 1200 possessions played) I can attempt to get individual player curves.

When fitting individual curves, Metabolic Power is the only measure of energy expenditure used, as in Section 4.8 I find it to produce the best fit and be the most predictive. I model individual relationships for the twelve Celtics players one variable at a time, while for the other variables, and all other players, I assume the same relationship as found in Section 4.7.2. Figure 20 shows the individual player curves for the C^H variable for offense.

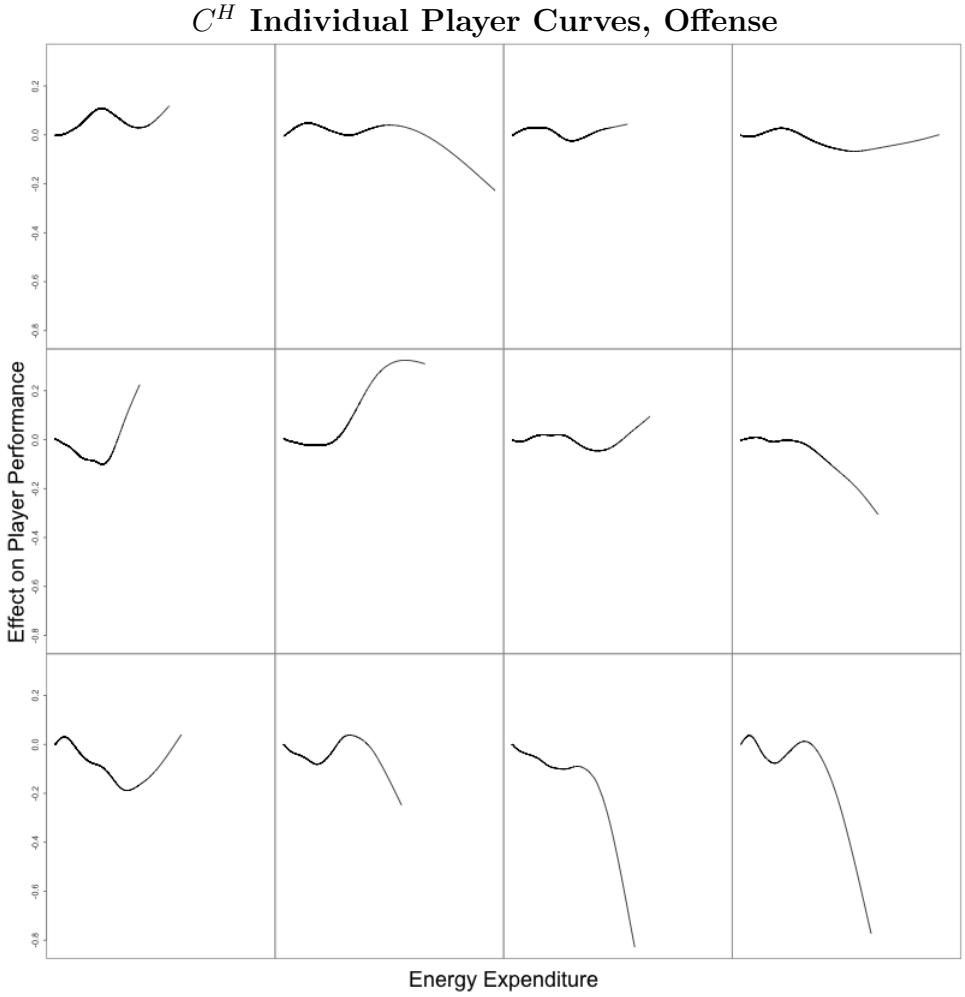

Figure 20: Individual player relationships to Consecutive Energy Expenditure (Halftime) when on offense for the twelve Celtics players that played the most minutes.

I then take the average of those curves and plot that. The average of those curves is taken by aggregating the fitted data points for the individual curves and then fitting another

GAM through those points. This method is used to ensure that players with more possessions played (and thus whose relationship I am more confident about) are weighted more heavily.

C^H Aggregate Curve, Offense

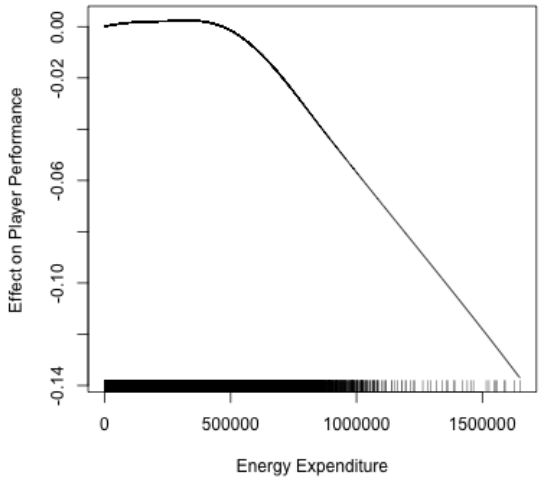

Figure 21: The average of the individual players curves for Consecutive Energy Expenditure (Halftime) on offense.

The individual player curves, as plotted in Figure 20, reveal a distinct relationship for each player, which is not that surprising given the highly variable nature of points being scored. However, similarities do exist between most of the curves. Most of them show an increase in the beginning, followed by a decrease later on, which is partially consistent with the parametric relationships found in Section 4.7.1. The average of the individual player curves resembles a quadratic curve with a negative coefficient, which I find to be the parametric relationship for C^H. I calculate individual player curves and aggregated curves for the six other variables found to be significant in the final model. These can be found in Appendix F.

4.10 Discussion

Even with access to only a small subset of the available data, several key insights towards the relationship between energy expenditure can be made. Although I find marginally different

relationships across the different measures of energy expenditure, all tell a similar story. Specifically, two insights that I would like to highlight are:

- It takes players a certain amount of time to "warm up".
- In some cases, it does not appear that players ever tire in a way that effects their performance.

I find ample evidence that players are significantly worse earlier in the game, both on offense and defense. This claim is backed up by evidence from the model selection process, as seen in plots shown in Figure 18 and Figure 19. Three of the four offensive relationships, along with two of the three defensive relationships, are best described by a combination of transformations that involve a logarithmic transformation. The defining aspect of the logarithmic relationship is a relatively quick increase in the beginning that eventually levels off, which appears to be important for capturing the discrepancy between a player's performance when first entering the game relative to his usual performance.

What does this discrepancy represent? Practically, it means a player performs well below his usual level when he first checks onto the court. It is important to note that the plots in Figure 19 show the relationship between points per possession and energy expenditure on defense. The finding that players typically allow more points per possession therefore means that they are performing below their usual level. This decreased performance early on supports the idea that players require a certain amount of time to "warm up".

It has been found in several other sports (primarily individual sports, where performance is more simple to define) that an active warm up can lead to better performance, as concluded by a recent meta-analysis [Fradkin et al., 2010]. The warm ups performed by players before entering the game in the NBA are limited and not very active. Most players warm up by shooting, but rarely do they do anything close to practicing full plays that might be called during the game [Dubin, 2015]. Furthermore, there is not much opportunity to warm up after the game has started. When a player is substituted in during the game they do not

warm up at all, as they typically go from their seat on the bench to a squatting position next to the scorers table while they wait for an opportunity to enter the game. This inactivity prior to entering the game contrasts sharply with other sports like soccer, where players often run sprints up and down the sideline before entering the game. Between quarters, players huddle around their coach and do not have time to stay active, as the break is rather short. At halftime, players go back to the locker room, reemerging only briefly at the end of the break for a quick shoot around before the second half starts.

Given that players warm up so little in the NBA, it seems reasonable that the first possession or two would serve as their warm up. Looking more closely at the relationships that I find, the logarithmic transformation is particularly noticeable in the C^Q relationship and the T relationship. The C^Q relationship suggests that players require time to warm up when first entering the game at the beginning of a quarter as well as whenever they come off the bench. The T relationship suggests that there is a further need to warm up at the beginning of the game - not in the way that it takes a player longer to warm up but rather in that that his performance suffers even more during his initial time on the court. The lack of a logarithmic relationship in the C^H variable suggests a player's performance after halftime is no different than after a regular break between quarters.

The finding that players require time to warm up or their play is worse, especially at the beginning of the game, is an actionable insight for NBA teams. The theoretical fix would be to have players warm up more substantially before entering the game than they currently do. Realistically, this fix is only practical for the beginning of the game, when the players have plenty of space and time to warm up. During the game there is little to no space between the bench and the sideline for players to warm up in any manner without running the risk of accidentally stumbling onto the court. Therefore, while it may not be practical for a team to have players warm up completely before entering the game, it is definitely actionable for teams to have players warm up more rigorously before the game, either on the court or in a separate workout space in the arena.

Moving onto the second insight, one of the more surprising conclusions that I reach is that in several cases player performance is strictly increasing with energy expenditure. The results from the fitted model as graphed in Figure 18 and Figure 19 show that on offense the effect observed between both Total Energy Expenditure and Consecutive Energy Expenditure (Quarters) and player performance is never decreasing. Likewise, on defense, the effect observed between Consecutive Energy Expenditure (Quarters) and player performance is also never decreasing (note that for defense it is always decreasing with respect to points scored, but as this relationship is with respect to defense that reflects positively on player performance). The fact that performance never decreases with energy expenditure feels fairly counterintuitive - why is this happening?

There are several possible explanations. First, for both offense and defense there are variables that show a negative relationship between energy expenditure and player performance. For offense, the transformation of Consecutive Energy Expenditure (Halftime) that best describe the data is a negative quadratic term - thus, player performance is always decreasing. For defense, the transformation of Consecutive Energy Expenditure (Halftime) that best describe the data is a negative logarithmic term and a positive linear term. After a certain a period of decrease, the linear term becomes more important than the logarithmic term and points scored per possession increases (i.e. their defensive performance weakens). In many cases, the effect of Consecutive Energy Expenditure (Halftime) is acting on a player at the same time as Consecutive Energy Expenditure (Quarters) is - the only time it isn't is in the second or fourth quarter after a player finishes the first or third quarter respectively and stays on for the start of the next quarter. Thus, it is reasonable to plot the combination of these two effects to see how they interact.

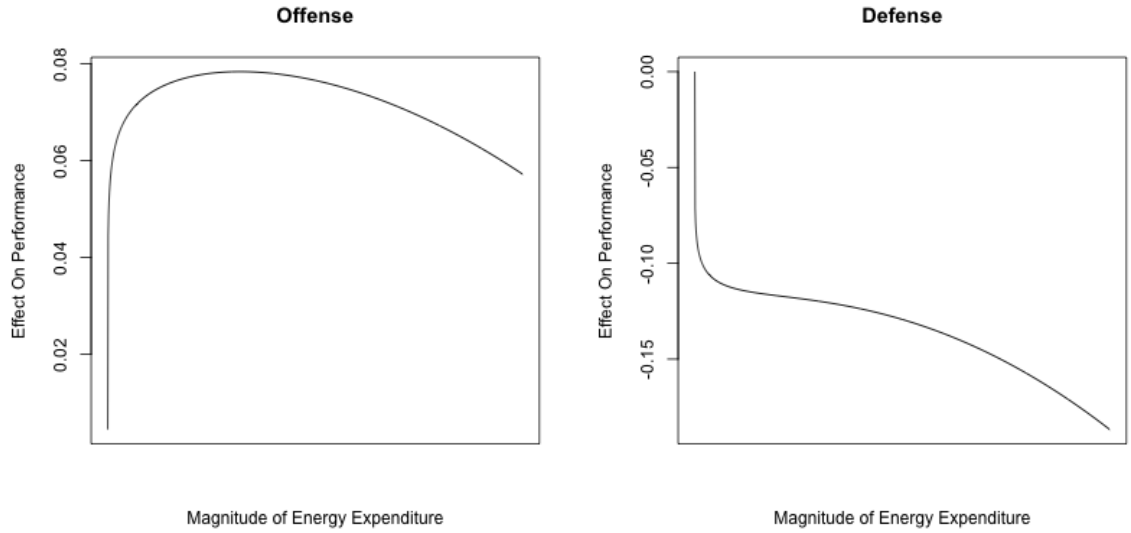

Figure 22: Combining the effects of C^H and C^Q, which often act together, when using Metabolic Power as the measure of energy expenditure.

I find a decreasing effect for offense past a certain point, which is more intuitive. However, the same counterintuitive effect still exists on defense, which is not ideal but does highlight the fact that my analysis is based on a relatively small dataset. With data for only 81 games, that is approximately only 6% of the available possessions played in the NBA last season. The relative scarcity of data is reflected in plots for the average of the individual player curves, as found in Appendix F. I find that for both Consecutive Energy Expenditure (Quarters) and Consecutive Energy Expenditure (Halftime) high measures of energy expenditure are associated with better performance on defense. However, the number of observations at high consecutive energy expenditure is fairly low (the number is denoted by the frequency of the black ticks at the bottom of the plots). Although evidence exists linking high energy expenditure to better performance, this evidence is driven by a small number of observations.

Finally, the prediction errors shown in Table 7 provide two key insights. Every measure of energy expenditure has more predictive power than the null model, implying that attempting to account for energy expenditure when predicting the outcome of a possession is a worthwhile pursuit. Second, Metabolic Power proves to be the best predictor, which is expected as it is the most theoretically complex measure of energy expenditure to compute

and is regarded as the best measure by existing literature. However, Time proved to be a better predictor than both Distance and VeDBA, which implies that while it is possible to use measures of energy expenditure demand from player tracking data (like Metabolic Power) to model player performance more accurately, simple tracking measures (like distance and VeDBA) don't provide any information gain.

5 Conclusion

This paper is intended to serve as the first step towards quantifying the relationship between energy expenditure and player performance in the National Basketball Association. Estimating player skill through ridge regression has been a common practice for years now, and thus this framework is used to initially obtain estimates of player skill. Since I am only granted access to a subset of the data, including energy expenditure variables directly into the model is not an option. Therefore, the residuals from the initial ridge regression are used as the dependent variables for a second regression, in which the independent variables are various transformations of energy expenditure variables.

Estimating energy expenditure itself is no easy task, and several measures are considered in this paper. One of these measures is derivable from play-by-play data and is included as a baseline measure: the time a player spends on the court (Time). The other three measures require player tracking data to calculate: distance run (Distance), Vector Dynamic Body Acceleration (VeDBA) and Metabolic Power. The first two of these measures have been calculated in previous papers with the same player tracking data, while Metabolic Power has been computed only for tracking data in other sports, most commonly soccer.

With these measures of energy expenditure during a play on the court, multiple variables are computed to serve as independent variables in the final regression - Total Energy Expenditure (T), Consecutive Energy Expenditure (Quarters) (C^Q), Consecutive Energy Expenditure (Halftime) (C^H), Rest (R) and an Exponential Moving Average of Energy Expenditure (E) - along with a singular variable for which player tracking data is not necessary to calculate - Score Differential (δ). It is hypothesized that the relationship between these variables and player performance is nonlinear, an thus a technique known as binning is used to approximate these relationships, of which a parametric form is estimated. Due to the number of subsets of possible models, it is not possible to perform an exhaustive search over all potential models and so a stepwise model selection process is performed to choose the

final model.

After performing this process, I compute final models for each of the four measures of energy expenditure, which all include nearly the exact same variables. A key insight from these models is that players take time to warm up immediately after entering the game, both on offense and defense. This finding suggests that teams could maximize a player's performance while he is on the court by having him warm up more rigorously before checking in.

Prediction errors for the models using the four different measures of energy expenditure are computed, and I find that the models created from all four measures of energy expenditure allow for more accurate predictions than the the null model. This finding is important as it signifies that energy expenditure is predictive of player performance. Metabolic Power, which is derived from player tracking data, allows for the creation of the model associated with the lowest predictive MSE. This finding is significant as it demonstrates that information can be gained from player tracking that allows one to better understand how player performance varies with energy expenditure.

This paper is novel in its attempt to analyze energy expenditure as computed from the player tracking data from the NBA. It is also one of the first to attempt to correlate energy expenditure with player performance in any sport, and no previous papers examine this relationship at such a granular level (every possession) as this paper does. Still, there is much work that can be done. Foremost, utilizing data from every NBA game, as opposed to the subset of Celtics games that I was granted access to, not only would surely provide more stable estimates but also would allow for the inclusion of energy expenditure variables directly into the ridge regression that estimates player skill. This procedure would be more methodologically sound and eliminate any confounding between the two sets of variables that may exist now. Interaction effects between the five transformations of energy expenditure could be considered. Specifically, while Rest was not found to be a significant predictor of player performance, it is reasonable to hypothesize that it could be when interacted

with other variables. Initial findings support this hypothesis, although were not included in this paper due to time constraints. Additionally, with estimates of the link between player performance and energy expenditure computed, attention could be directed towards the best ways to utilize this relationship: specifically, how to optimize rotations given this information. As an extension of that, it would also be possible to judge coaches on the effectiveness of their rotations.

While there is undoubtably a lot of work to be done, it is the my hope that this paper has both introduced an appropriate method for and clearly presented initial findings from quantifying the relationship between energy expenditure and player performance in the National Basketball Association.

Appendices

A

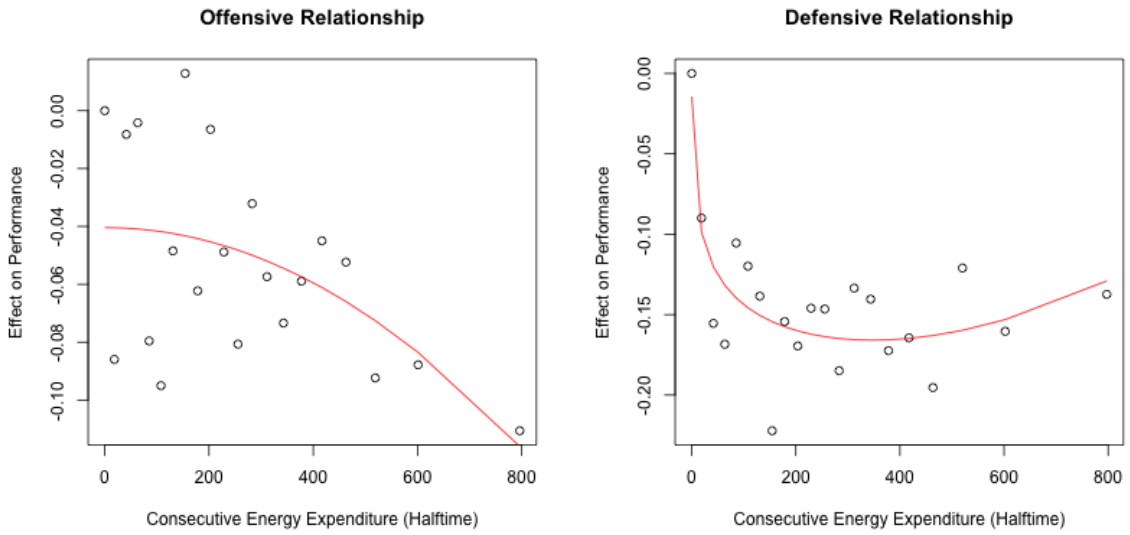

Figure 23: Plotting coefficients for the bins of Consecutive Energy Expenditure (Halftime) and attempting to model the relationship shown parametrically when using Time as the measure of energy expenditure.

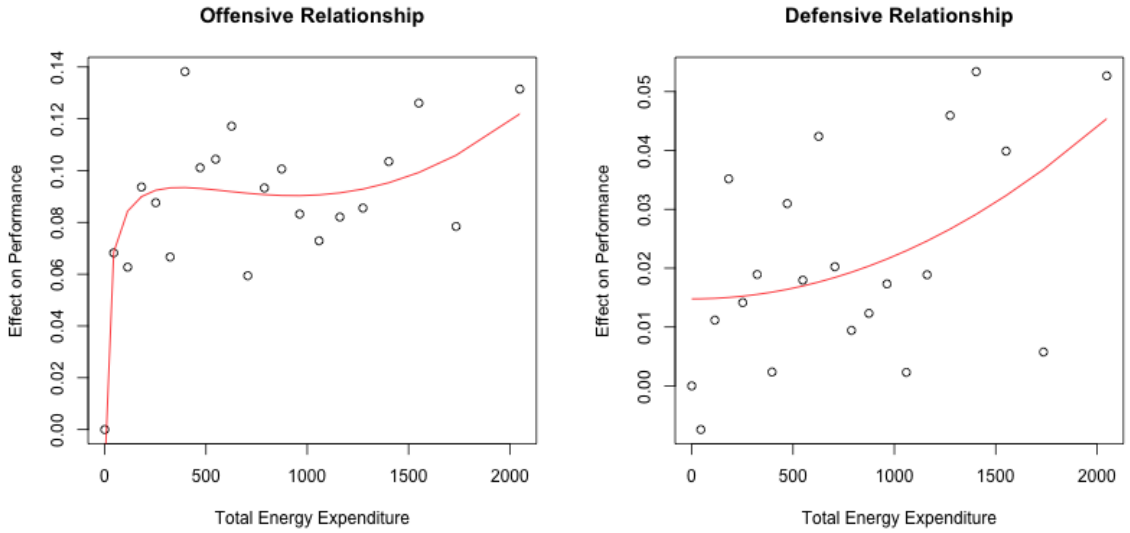

Figure 24: Plotting coefficients for the bins of Total Energy Expenditure and attempting to model the relationship shown parametrically when using Time as the measure of energy expenditure.

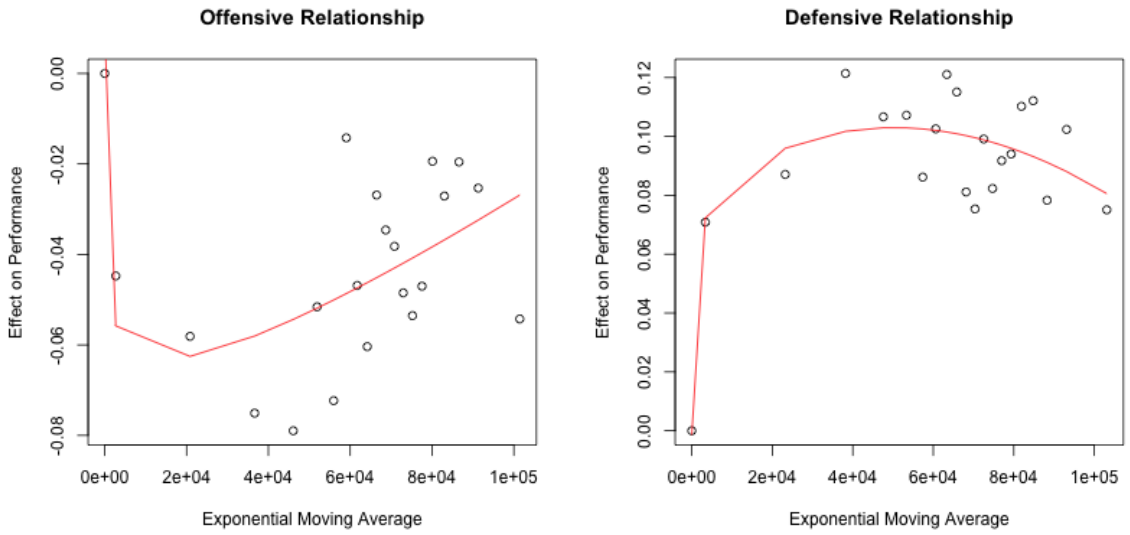

Figure 25: Plotting coefficients for the bins of Exponential Moving Average of Energy Expenditure and attempting to model the relationship shown parametrically when using Time as the measure of energy expenditure.

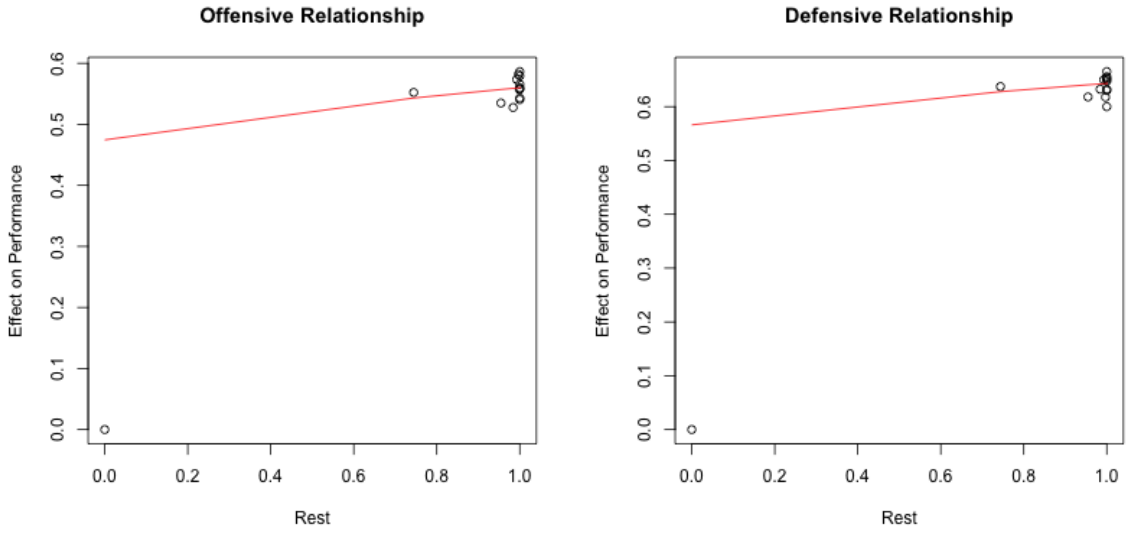

Figure 26: Plotting coefficients for the bins of Rest and attempting to model the relationship shown parametrically when using Time as the measure of energy expenditure.

B

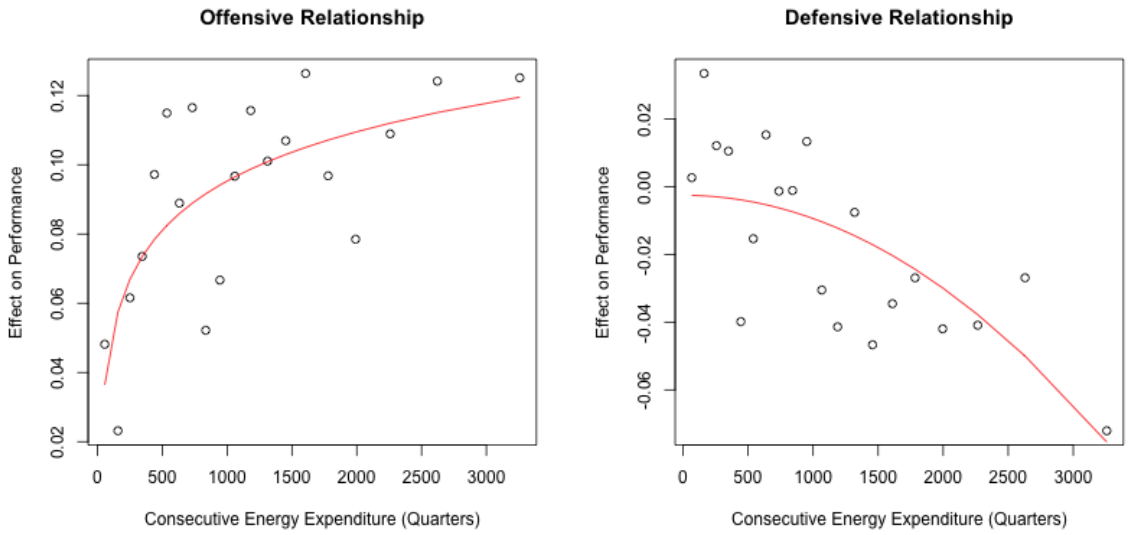

Figure 27: Plotting coefficients for the bins of Consecutive Energy Expenditure (Quarters) and attempting to model the relationship shown parametrically when using Distance as the measure of energy expenditure.

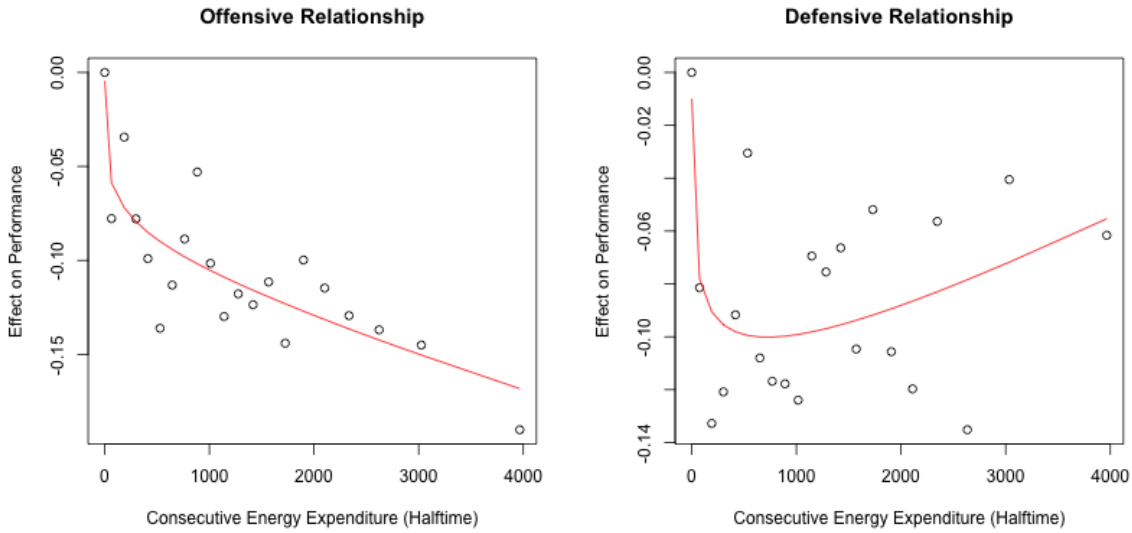

Figure 28: Plotting coefficients for the bins of Consecutive Energy Expenditure (Halftime) and attempting to model the relationship shown parametrically when using Distance as the measure of energy expenditure.

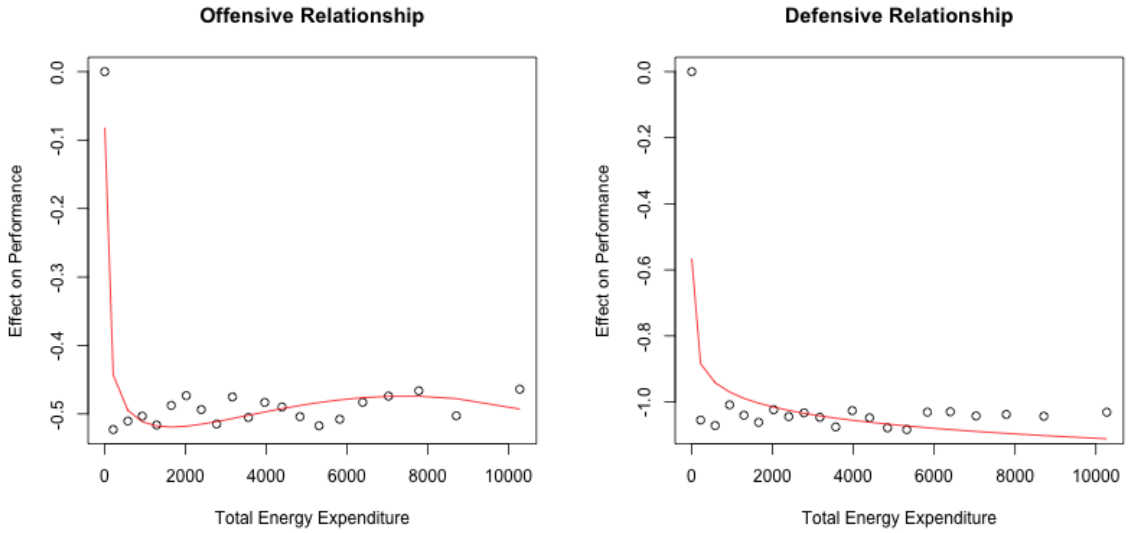

Figure 29: Plotting coefficients for the bins of Total Energy Expenditure and attempting to model the relationship shown parametrically when using Distance as the measure of energy expenditure.

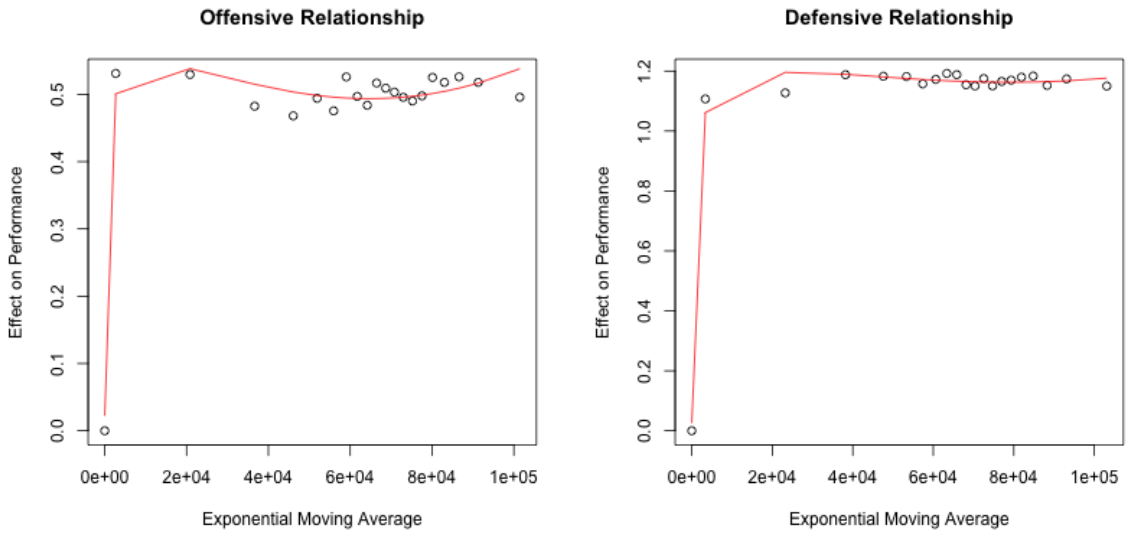

Figure 30: Plotting coefficients for the bins of Exponential Moving Average of Energy Expenditure and attempting to model the relationship shown parametrically when using Distance as the measure of energy expenditure.

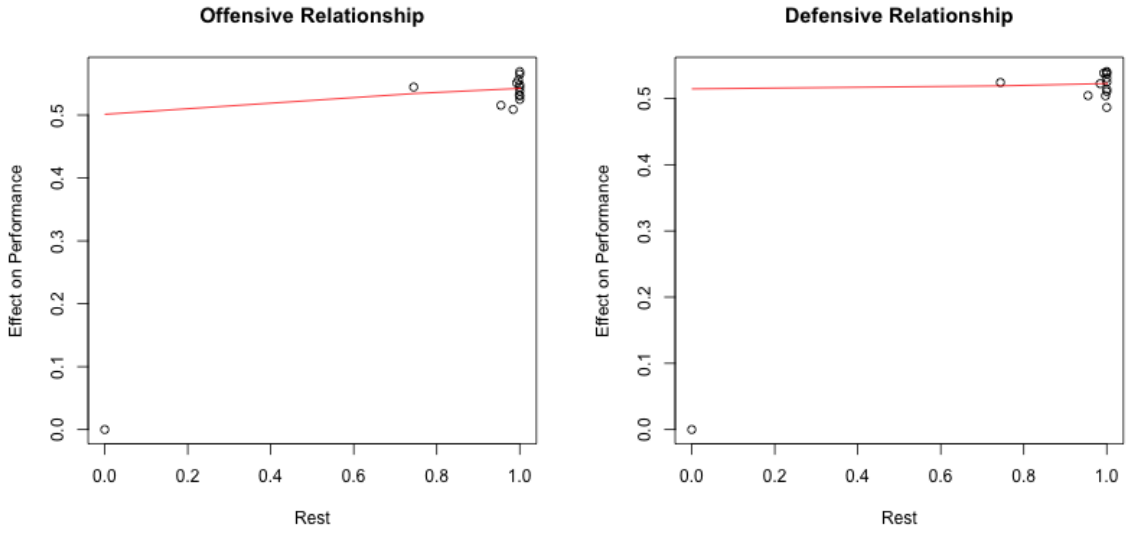

Figure 31: Plotting coefficients for the bins of Rest and attempting to model the relationship shown parametrically when using Distance as the measure of energy expenditure.

C

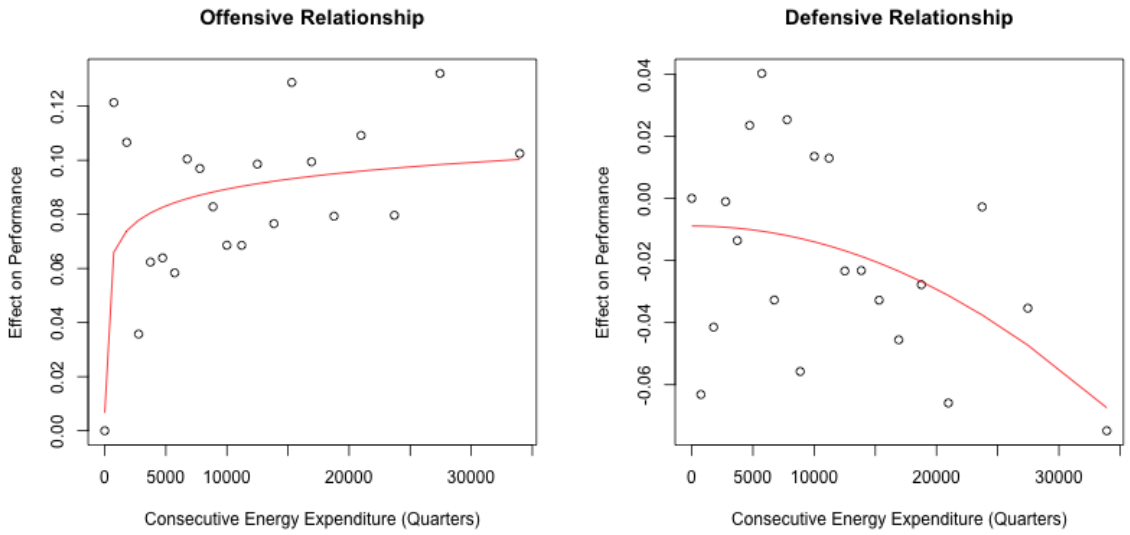

Figure 32: Plotting coefficients for the bins of Consecutive Energy Expenditure (Quarters) and attempting to model the relationship shown parametrically when using VeDBA as the measure of energy expenditure.

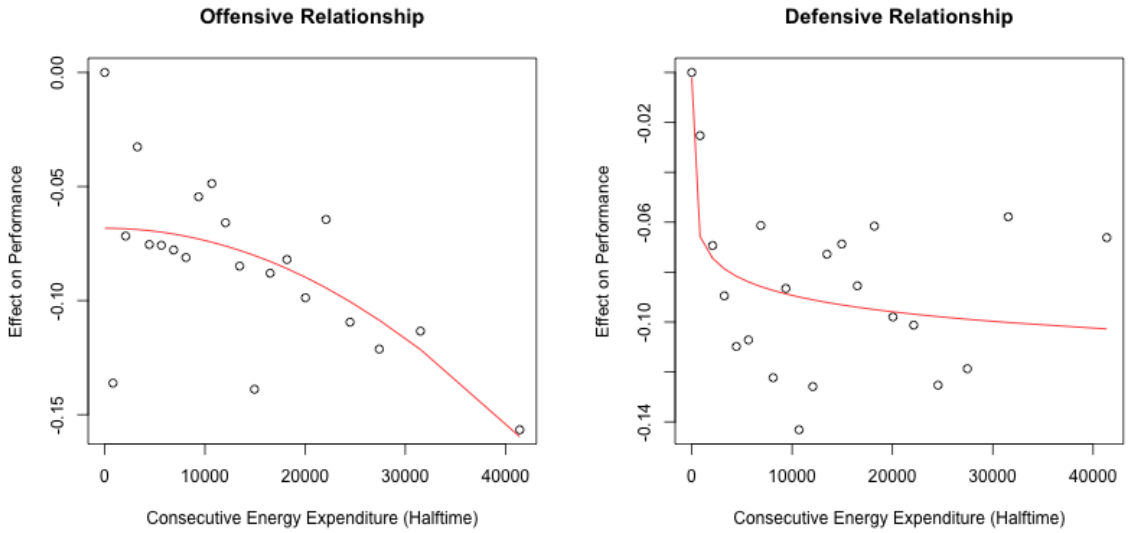

Figure 33: Plotting coefficients for the bins of Consecutive Energy Expenditure (Halftime) and attempting to model the relationship shown parametrically when using VeDBA as the measure of energy expenditure.

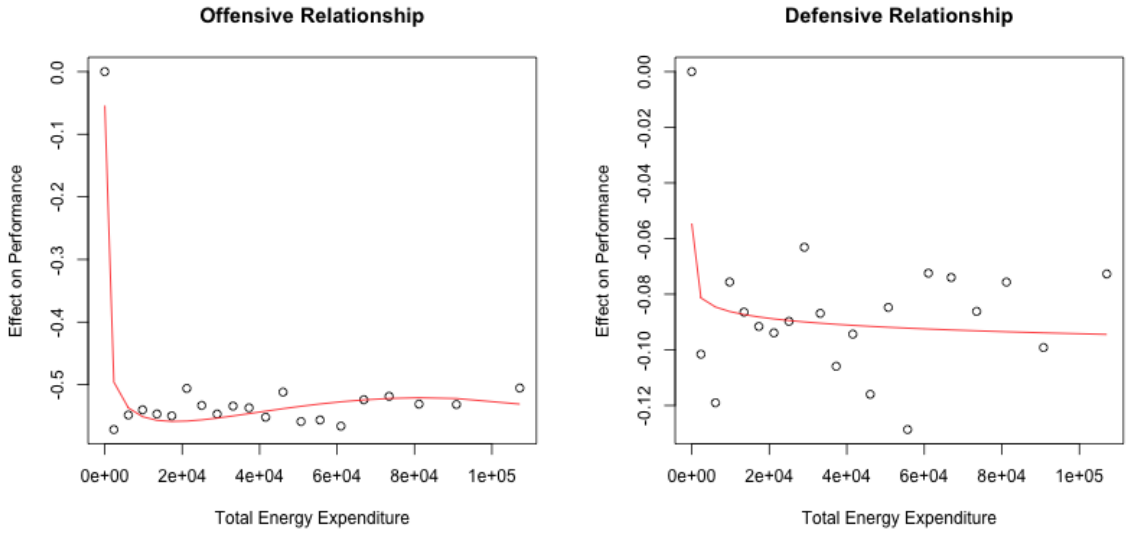

Figure 34: Plotting coefficients for the bins of Total Energy Expenditure and attempting to model the relationship shown parametrically when using VeDBA as the measure of energy expenditure.

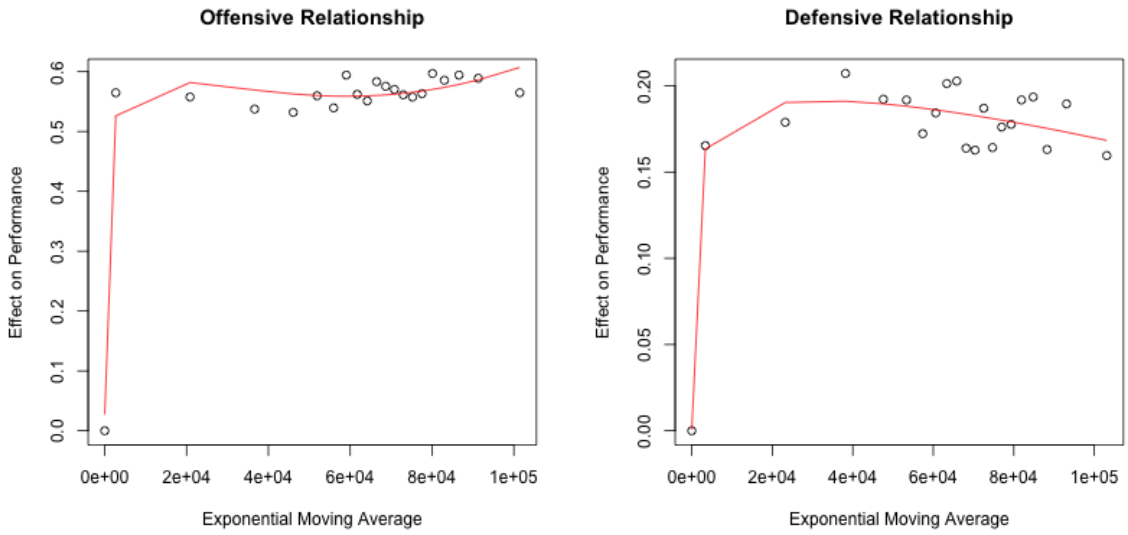

Figure 35: Plotting coefficients for the bins of Exponential Moving Average of Energy Expenditure and attempting to model the relationship shown parametrically when using VeDBA as the measure of energy expenditure.

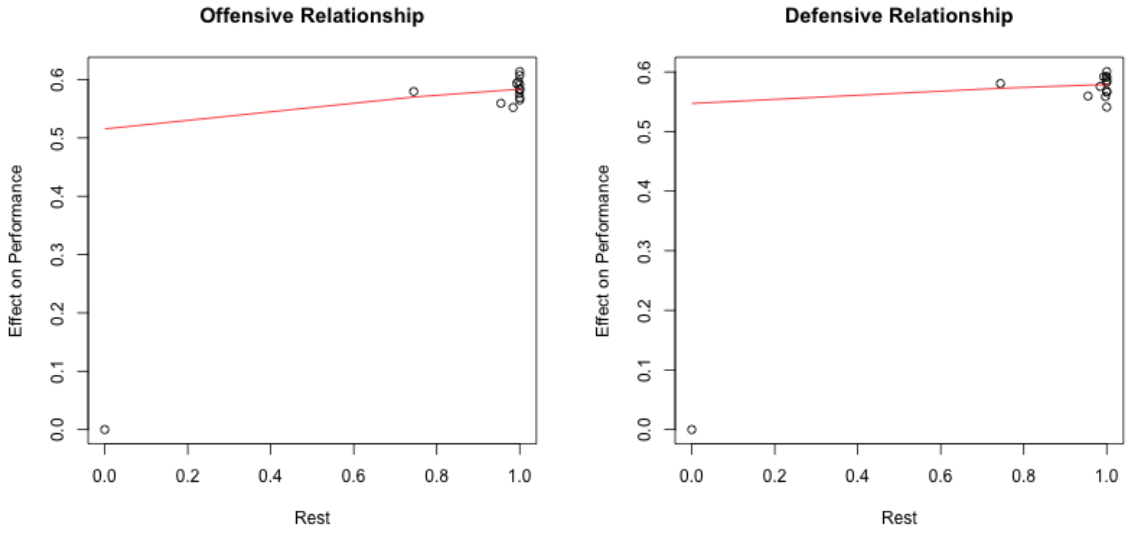

Figure 36: Plotting coefficients for the bins of Rest and attempting to model the relationship shown parametrically when using VeDBA as the measure of energy expenditure.

D

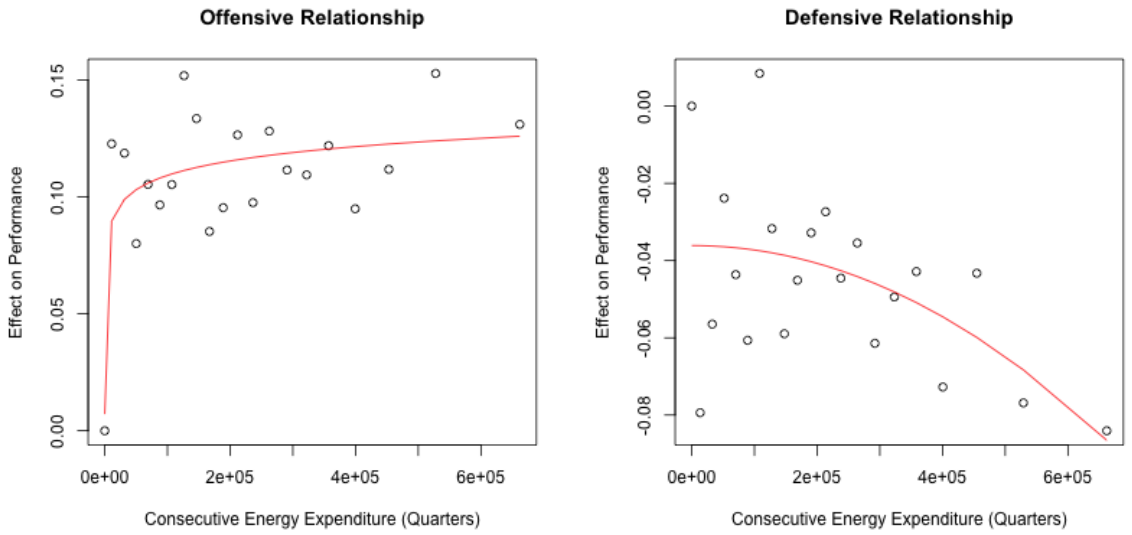

Figure 37: Plotting coefficients for the bins of Consecutive Energy Expenditure (Quarters) and attempting to model the relationship shown parametrically when using Metabolic Power as the measure of energy expenditure.

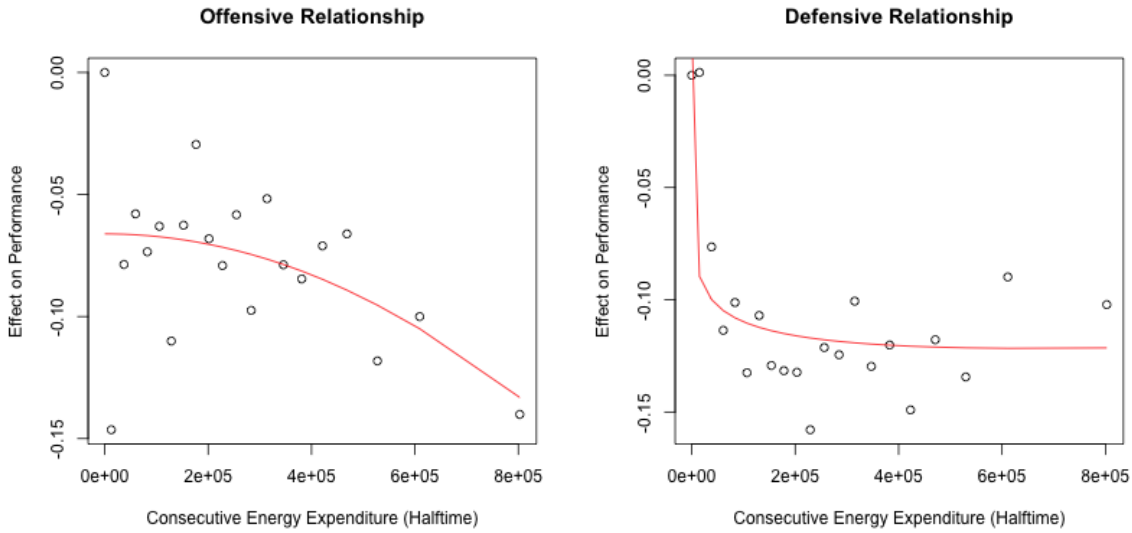

Figure 38: Plotting coefficients for the bins of Consecutive Energy Expenditure (Halftime) and attempting to model the relationship shown parametrically when using Metabolic Power as the measure of energy expenditure.

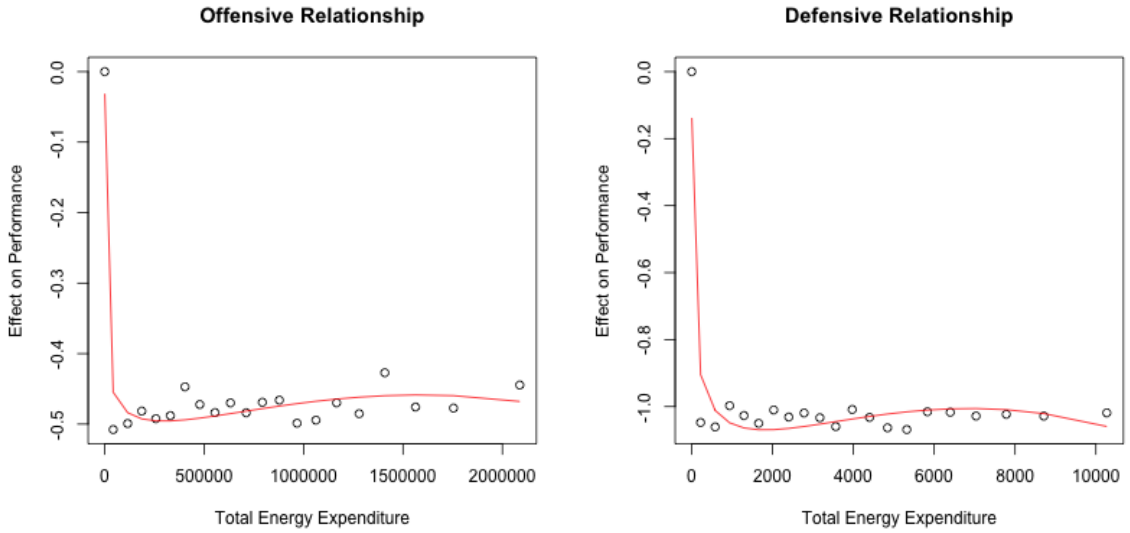

Figure 39: Plotting coefficients for the bins of Total Energy Expenditure and attempting to model the relationship shown parametrically when using Metabolic Power as the measure of energy expenditure.

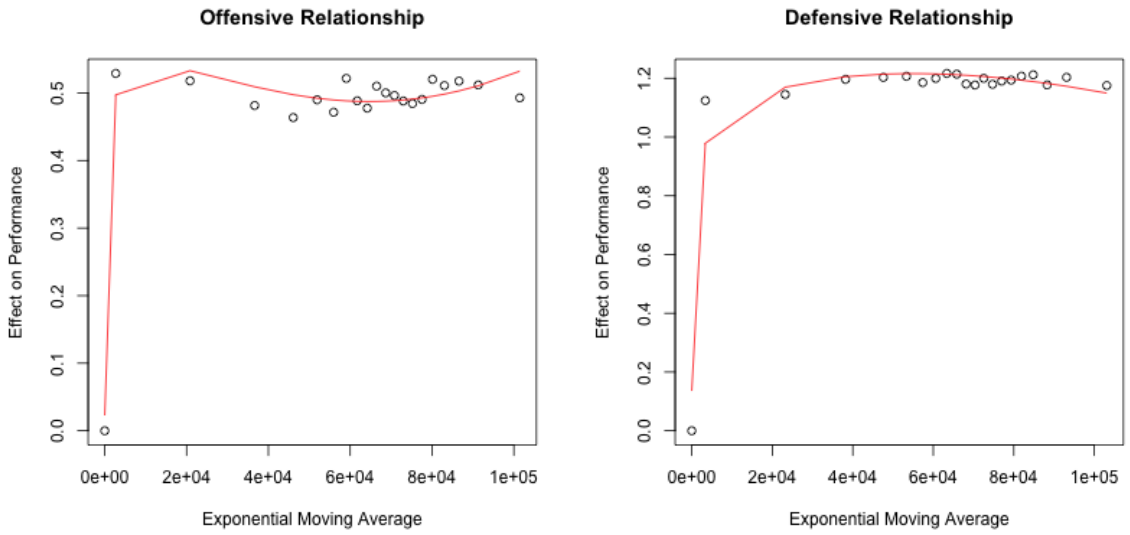

Figure 40: Plotting coefficients for the bins of Exponential Moving Average of Energy Expenditure and attempting to model the relationship shown parametrically when using Metabolic Power as the measure of energy expenditure.

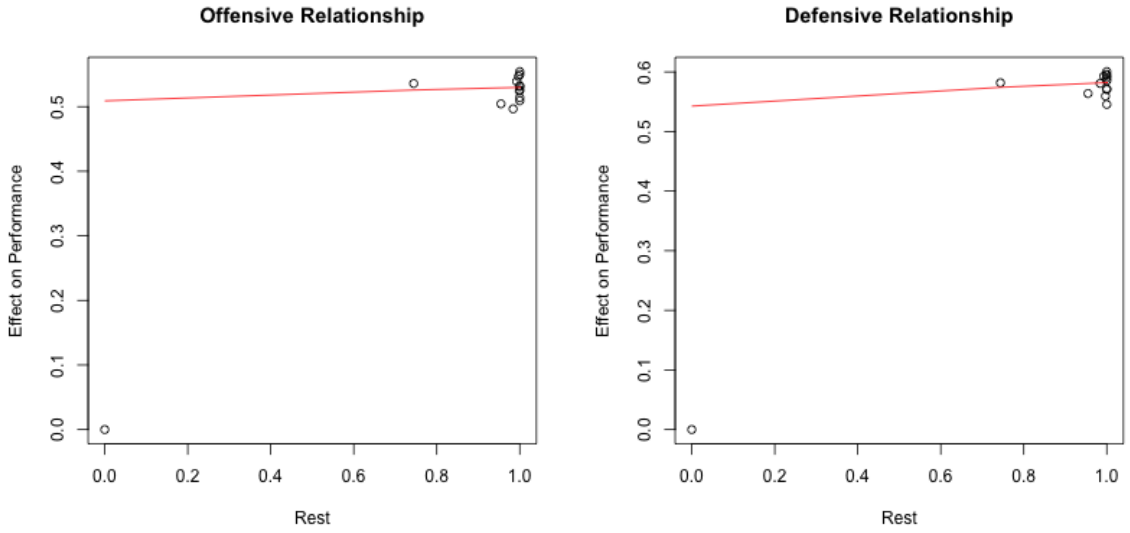

Figure 41: Plotting coefficients for the bins of Rest and attempting to model the relationship shown parametrically when using Metabolic Power as the measure of energy expenditure.

E

Variable	Linear	Quadratic	Logarithmic
C^H (Offense)	✓		✓
C^H (Defense)	✓		✓
T (Offense)	✓	✓	✓
T (Defense)	✓	✓	✓
C^Q (Offense)			✓
C^Q (Defense)	✓		
E (Offense)	✓	✓	✓
E (Defense)	✓	✓	✓
R (Offense)			✓
R (Defense)		✓	

Table 8: Which transformations I find to best model the relationship shown by the coefficients from binning when using Distance as the measure of energy expenditure.

Variable	Linear	Quadratic	Logarithmic
C^H (Offense)		✓	
C^H (Defense)			✓
T (Offense)	✓	✓	✓
T (Defense)			✓
C^Q (Offense)			✓
C^Q (Defense)		✓	
E (Offense)	✓	✓	✓
E (Defense)	✓	✓	✓
R (Offense)			✓
R (Defense)			✓

Table 9: Which transformations I find to best model the relationship shown by the coefficients from binning when using VeDBA as the measure of energy expenditure.

Variable	Linear	Quadratic	Logarithmic
C^H (Offense)		✓	
C^H (Defense)	✓		✓
T (Offense)			✓
T (Defense)			✓
C^Q (Offense)			✓
C^Q (Defense)		✓	
E (Offense)	✓	✓	✓
E (Defense)		✓	✓
R (Offense)			✓
R (Defense)			✓

Table 10: Which transformations I find to best model the relationship shown by the coefficients from binning when using Metabolic Power as the measure of energy expenditure.

F

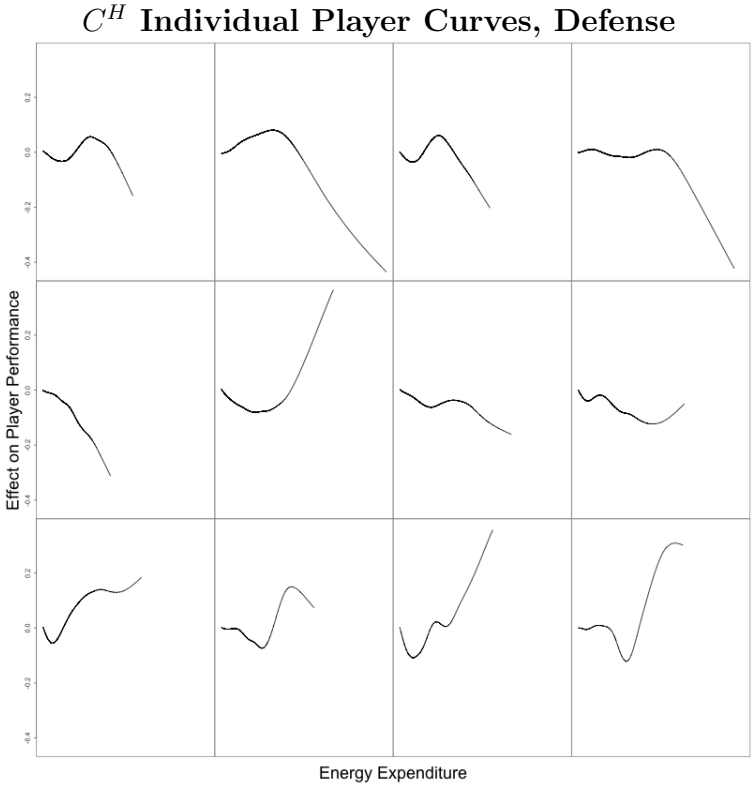

Figure 42: Individual player relationships to Consecutive Energy Expenditure (Halftime) when on defense for the twelve Celtics players that played the most minutes.

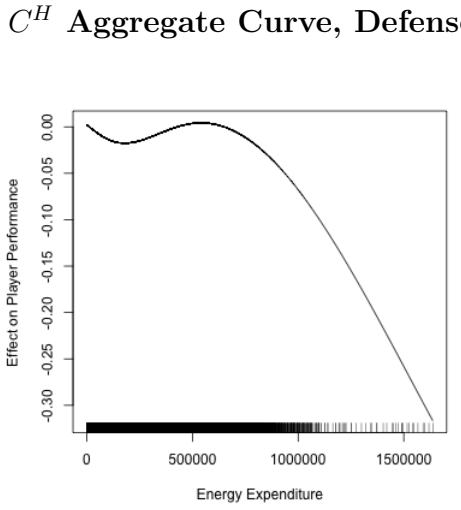

Figure 43: The average of the individual players curves for Consecutive Energy Expenditure (Halftime) on defense.

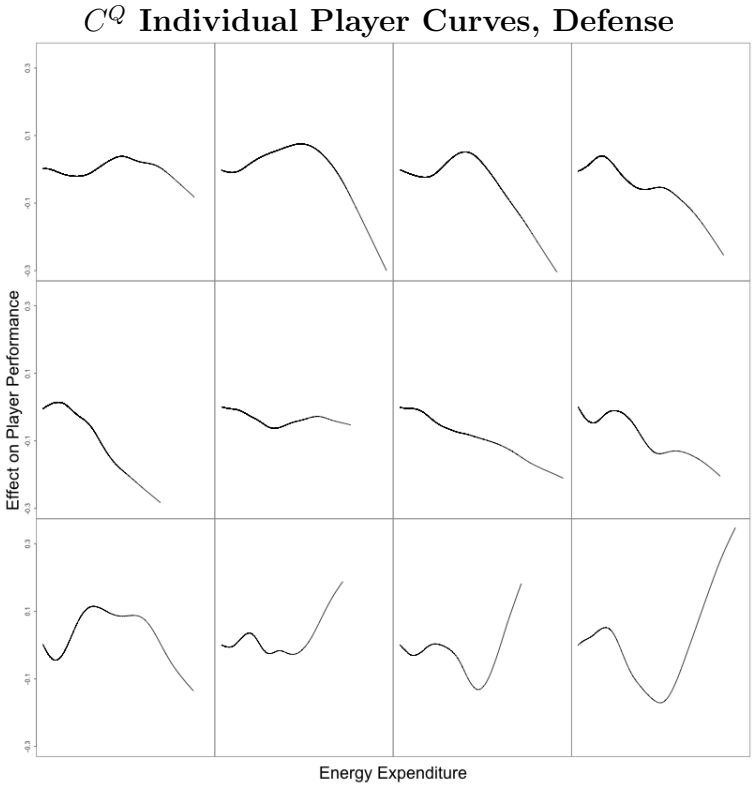

Figure 44: Individual player relationships to Consecutive Energy Expenditure (Quarters) when on defense for the twelve Celtics players that played the most minutes.

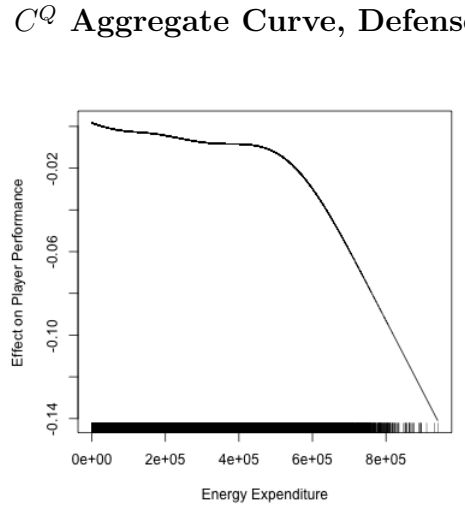

Figure 45: The average of the individual players curves for Consecutive Energy Expenditure (Quarters) on defense.

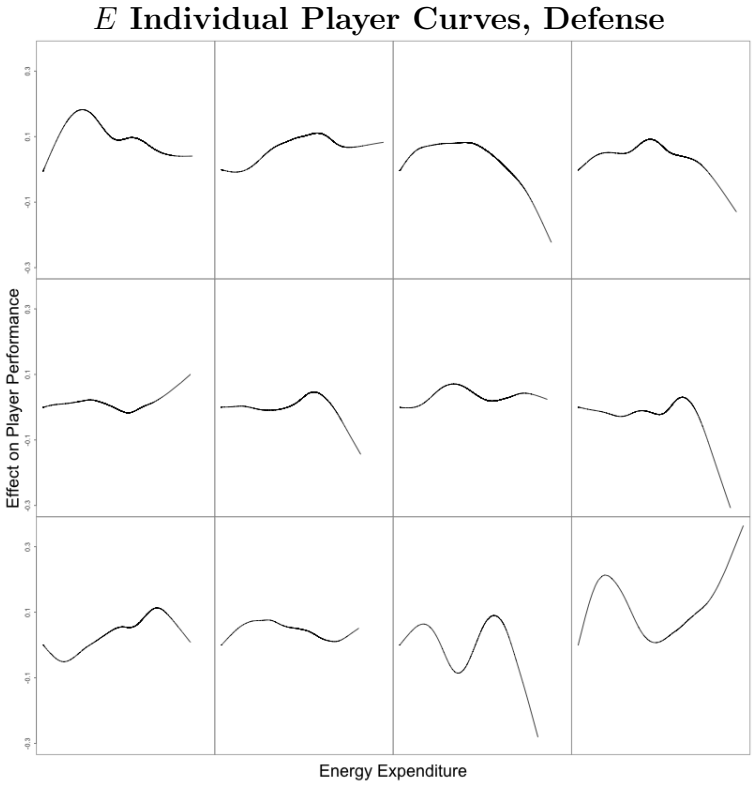

Figure 46: Individual player relationships to Exponential Moving Average of Energy Expenditure when on defense for the twelve Celtics players that played the most minutes.

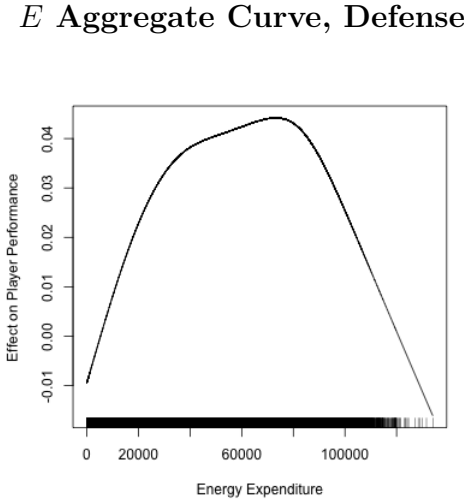

Figure 47: The average of the individual players curves for Exponential Moving Average of Energy Expenditure on defense.

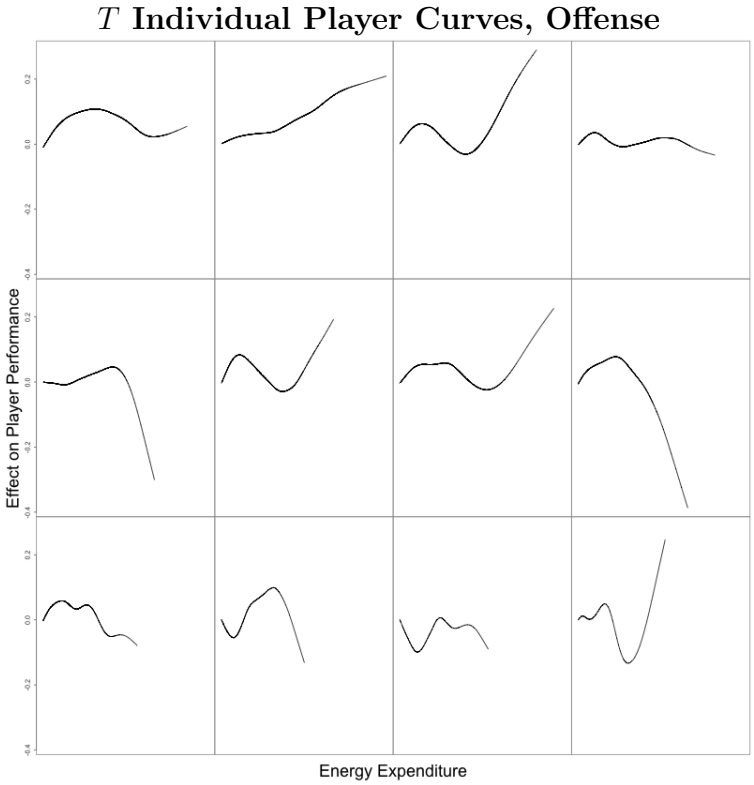

Figure 48: Individual player relationships to Total Energy Expenditure when on offense for the twelve Celtics players that played the most minutes.

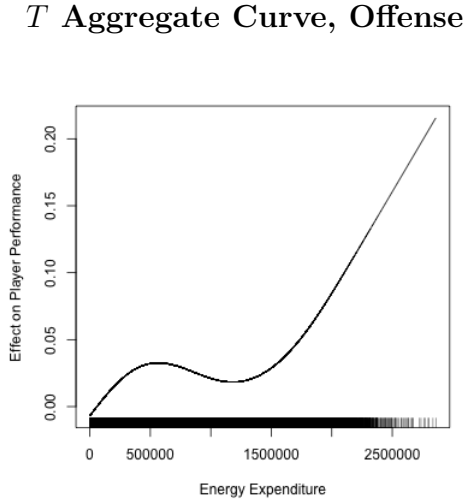

Figure 49: The average of the individual players curves for Total Energy Expenditure on offense.

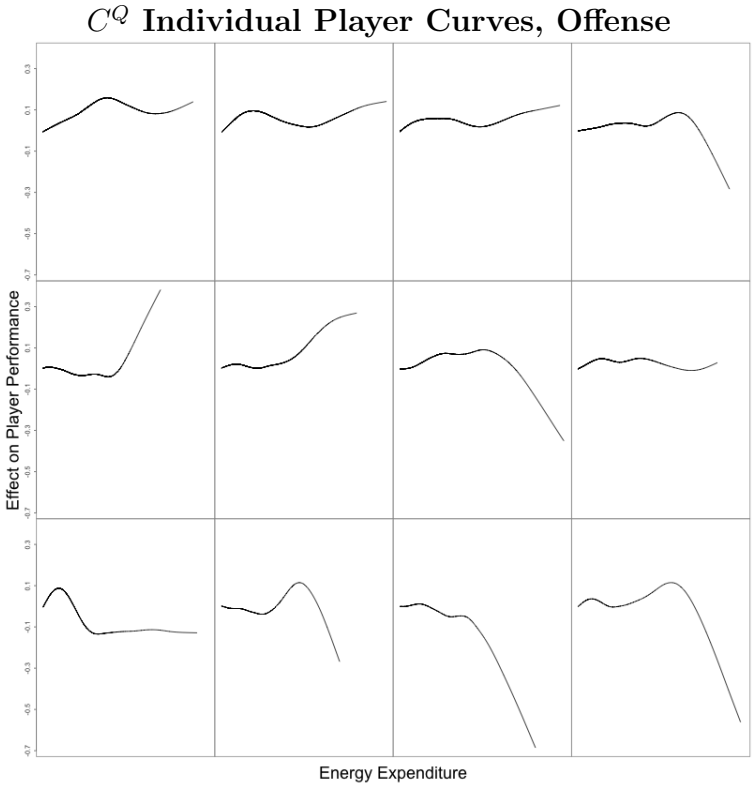

Figure 50: Individual player relationships to Consecutive Energy Expenditure (Quarters) when on offense for the twelve Celtics players that played the most minutes.

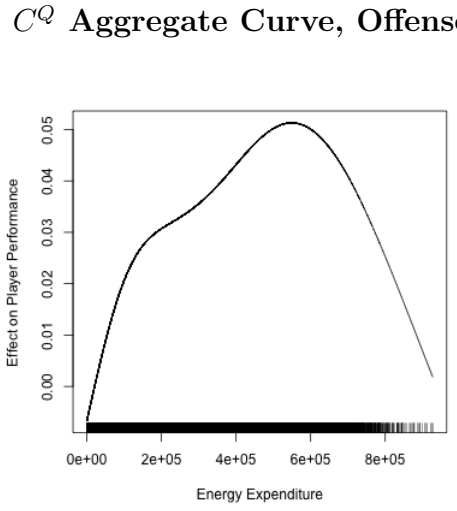

Figure 51: The average of the individual players curves for Consecutive Energy Expenditure (Quarters) on offense.

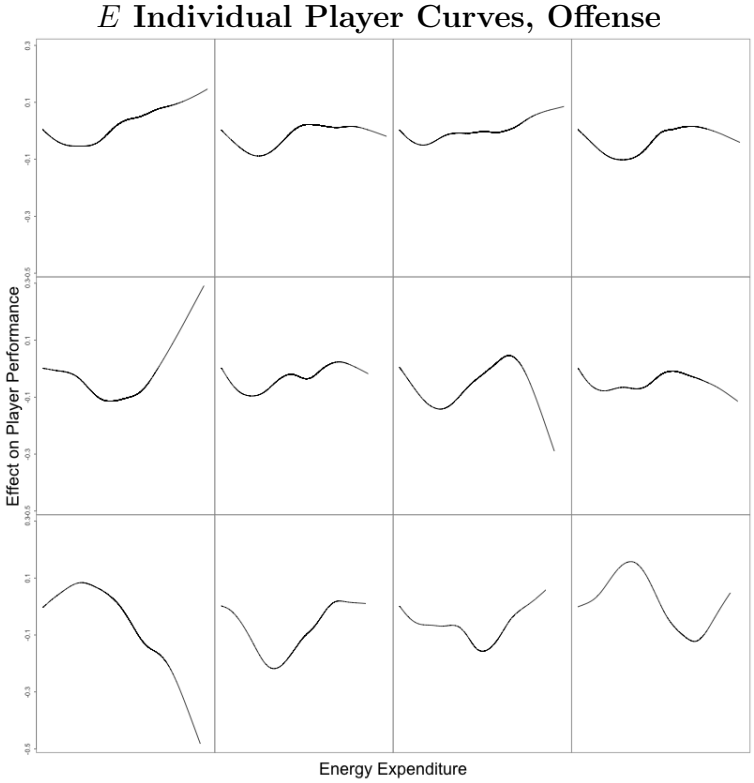

Figure 52: Individual player relationships to Exponential Moving Average of Energy Expenditure when on offense for the twelve Celtics players that played the most minutes.

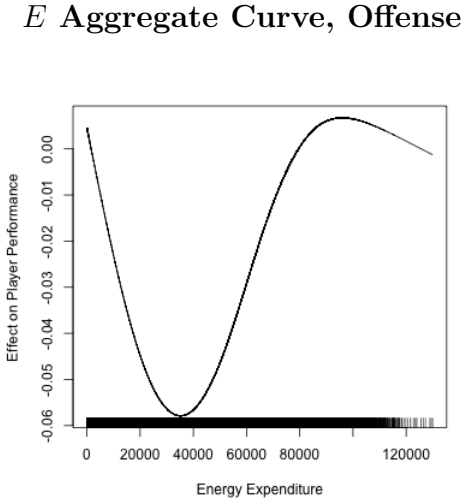

Figure 53: The average of the individual players curves for Exponential Moving Average of Energy Expenditure on offense.

References

Chris Anderson and David Sally. *The Numbers Game: Why Everything You Know about Football is Wrong*. Penguin Books, London, 2014.

Gordon Blower and Julia E. Kelsall. Nonlinear kernel density estimation for binned data: Convergence in entropy. *Bernoulli*, 8(4):423–449, 2002. ISSN 13507265.

Andrew Bocskocsky, John Ezekowitz, and Carolyn Stein. The hot hand: A new approach to an old fallacy? Paper presented at the MIT Sloan Sports Analytics Conference, 2014.

Luke Bornn, Kirk Goldsberry, Dan Cervone, and Alex D'Amour. Pointwise: Predicting points and valuing decisions. Paper presented at the MIT Sloan Sports Analytics Conference, 2014.

P. E. di Prampero. The energy cost of human locomotion on land and in water. *International Journal of Sports Medicine*, 7(2):55–72, 1986.

P. E. di Prampero, S. Fusi, L. Sepulcri, J. B. Morin, A. Belli, and G. Antonutto. Sprint running: a new energetic approach. *Journal of Experimental Biology*, 208(14):2809–2816, 2005. ISSN 0022-0949. doi: 10.1242/jeb.01700.

Jared Dubin. Rhythm of the night: Nba players talk about their pregame warm-ups, 2015.

George Dvorsky. The physics of usain bolt's world record 100-meter dash, July 2013. [Online; posted 26-July-2013].

Jeremias Engelmann. The effect of leading, 2011. URL `http://apbr.org/metrics/viewtopic.php?f=2&t=8501&start=0`.

Jeremias Engelmann. Possession based player performance analysis in baseketball. In Jim Albert, Mark E. Glickman, Tim B. Swartz, and Ruud H. Koning, editors, *Handbook of Statistical Methods and Analyses in Sports*. Chapman and Hall/CRC, 2017.

L. Eriksson, T. Byrne, E. Johansson, J. Trygg, and C. Vikström. *Multi- and Megavariate Data Analysis Basic Principles and Applications:*. MKS Umetrics, 2013. ISBN 9789197373050.

Andrea Fradkin, Tsharni Zazryn, and James Smoliga. Effects of warming-up on physical performance: a systematic review with meta-analysis, 2010.

Alexander Franks, Andrew Miller, Luke Bornn, and Kirk Goldsberry. Counterpoints: Advanced defensive metrics for nba basketball. Paper presented at the MIT Sloan Sports Analytics Conference, 2015.

Tim Gabbett. Physical, technical and tactical factors affect final ladder position in semiprofessional rugby league. *International Journal of Sports Physiology and Performance*, 9 (4):680–688, 2013.

Kirk Goldsberry and Eric Weiss. The dwight effect:a new ensemble of interior defense analytics for the nba. Paper presented at the MIT Sloan Sports Analytics Conference, 2013.

Steven Ilardi and Aaron Barzilai. Adjusted plus-minus ratings, 2007.

Jared Johnson. Are offense and defense created equal in the nba?, 2015.

Rajiv Maheswaran, Yu-Han Chang, Aaron Henehan, and Samantha Danesis. Deconstructing the rebound with optical tracking data. Paper presented at the MIT Sloan Sports Analytics Conference, 2012.

Rajiv Maheswaran, Yu-Han Chang, Jeff Su, Sheldon Kwok, Tal Levy, Adam Wexler, and Noel Hollingsworth. The three dimensions of rebounding. Paper presented at the MIT Sloan Sports Analytics Conference, 2014.

R. Margaria, P. Cerretelli, P. Aghemo, and G. Sassi. Energy cost of running. *Journal of Applied Physiology*, 18(2):367–370, 1963. ISSN 8750-7587.

Philip Maymin. Acceleration in the nba: Towards an algorithmic taxonomy of basketball plays. Paper presented at the MIT Sloan Sports Analytics Conference, 2013.

Avery McIntyre, Joel Brooks, John Guttag, and Jenna Wiens. Recognizing and analyzing ball screen defense in the nba. Paper presented at the MIT Sloan Sports Analytics Conference, 2016.

Armand McQueen, Jenna Wiens, , and John Guttag. Automatically recognizing on-ball screens. Paper presented at the MIT Sloan Sports Analytics Conference, 2014.

Alberto E. Minetti, Christian Moia, Giulio S. Roi, Davide Susta, and Guido Ferretti. Energy cost of walking and running at extreme uphill and downhill slopes. *Journal of Applied Physiology*, 93(3):1039–1046, 2002. ISSN 8750-7587. doi: 10.1152/japplphysiol.01177.2001.

Cristian Osgnach, Stefano Poser, Riccardo Bernardini, Roberto Rinaldo, and Pietro Enrico di Prampero. Energy cost and metabolic power in elite soccer: a new match analysis approach. *Medicine and science in sports and exercise*, 42 1:170–8, 2010.

R. Passmore and J. V. G. A. Durnin. Human energy expenditure. *Physiological Reviews*, 35 (4):801–840, 1955.

John F Patton. Measurement of oxygen uptake with portable equipment. *Emerging Technologies for Nutrition Research: Potential for Assessing Military Performance Capability*, page 297, 1997.

Ted Polglaze, Brian Dawson, and Peter Peeling. Gold standard or fool's gold? the efficacy of displacement variables as indicators of energy expenditure in team sports. *Sports Medicine*, 46(5):657–670, 2016. ISSN 1179-2035. doi: 10.1007/s40279-015-0449-x.

Lama Qasem, Antonia Cardew, Alexis Wilson, Iwan Griffiths, Lewis G. Halsey, Emily L. C. Shepard, Adrian C. Gleiss, and Rory Wilson. Tri-axial dynamic acceleration as a proxy

for animal energy expenditure; should we be summing values or calculating the vector? *PLOS ONE*, 7(2):1–8, 02 2012. doi: 10.1371/journal.pone.0031187.

Ermanno Rampinini, Franco M Impellizzeri, Carlo Castagna, and Ulrik Wislff. Technical performance during soccer matches of the italian serie a league. *Journal of Science and Medicine in Sport*, 12(1):227–233, 2009.

Jaime Sampaio, Bruno Gonalves, L. Rentero, Catarina Abrantes, and Nuno Leite. Exploring how basketball players? tactical performances can be affected by activity workload. *Science and Sports*, 29(4):23–30, 2014.

Joseph Sill. Improved nba adjusted +/- using regularization and out-of-sample testing. Paper presented at the MIT Sloan Sports Analytics Conference, 2010.

Matthew C. Varley, Tim Gabbett, and Robert J. Aughey. Activity profiles of professional soccer, rugby league and australian football match play. *Journal of Sports Sciences*, 2013.

Jenna Wiens, Guha Balakrishnan, Joel Brooks, and John Guttag. To crash or not to crash. Paper presented at the MIT Sloan Sports Analytics Conference, 2013.

Sean Williams, Stephen West, Matthew J Cross, and Keith A Stokes. Better way to determine the acute:chronic workload ratio? *British Journal of Sports Medicine*, 51(3):209–210, 2017. ISSN 0306-3674. doi: 10.1136/bjsports-2016-096589.

W.L. Winston. *Mathletics: How Gamblers, Managers, and Sports Enthusiasts Use Mathematics in Baseball, Basketball, and Football (New in Paper)*. Princeton University Press. Princeton University Press, 2012. ISBN 9780691154589.