Foul Accumulation in the NBA

by

Dani Chu

B.Sc., Simon Fraser University, 2018

Project Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
Department of Statistics and Actuarial Science
Faculty of Science

SIMON FRASER UNIVERSITY
Spring 2020

Approval

Name:	**Dani Chu**
Degree:	**Master of Science (Statistics)**
Title:	**Foul Accumulation in the NBA**
Examining Committee:	**Chair:** Jinko Graham Professor
	Tim B. Swartz Senior Supervisor Professor
	Derek Bingham Internal Examiner Professor
	Aaron Danielson External Examiner Postdoctoral Research Scholar
Date Defended:	**January 13th 2020**

Abstract

This project investigates the fouling time distribution of players in the National Basketball Association. A Bayesian analysis is presented based on the assumption that fouling times follow a Gamma distribution. Various insights are obtained including the observation that players accumulate their nth foul more quickly for increasing n. Methods are developed that will allow coaches to better manage playing time in the presence of fouls such that key players are available in the latter stages of matches.

Keywords: Bayesian analysis; censoring; constraints; failure time distributions; predictive inference

Acknowledgements

“I am because we are" is the South African concept of Ubuntu. It is a concept that was introduced to our basketball team by Coach Colin Macdonald during my senior year at Dr. Charles Best Secondary School. It meant everything to our team then and it could not be more true now. I am because of everyone around me. I am eternally grateful to those that I have worked with, mentored me and gave me the privilege of coaching them. This project would not be possible without a mountain of support and love.

First and foremost, thank you to my parents (my mom always told me if I ever won anything on TV I had to thank her first). Thanks Mom. Thank you for your endless support, sacrifices, lessons and laughter. Thank you for everything. Thank you to my grandparents for teaching me about art, hard work and love. To my cousins and brothers, Omar, Sammy, Jamal, Noah, Luc, Leo, Nikolas, Nadia, Jasper and Jonathan, thank you for all of the uniqueness you bring to the family roster. There isn't a more well rounded group out there. For my Aunts, Reema, Yaz, and Tasha and Uncles, Abe, Marv and Pete, who have shaped so much of who I am, thank you. Thank you to my extended families the Uzelacs and the McFarlens.

While the statistics program at SFU is different from the basketball court in many ways it is the same in a very important way. No one succeeds without teammates. To Lucas, Matt and James thank you for being the most incredible teammates I could ever ask for. The rest of our cohort is also incredible, shoutout to Coco, Barinder, Dylan, Jason, Michael, Megan, Matt and Gabe for all of the laughs in the office and your help getting through the program. Thank you to the grad students who came before me for all of their mentorship, Sarah, Will, Kevin, Jacob, Nate and Maude I hope one day to be even close to as smart as you. To my previous classmates who have moved on from SFU Abe, Forrest, Kristen, Sophia and Andrew, it is incredible to follow your successes.

In the same way as a team is directionless if not for the head coach, I am forever grateful to Dr. Tim Swartz for all of your support and guidance. This would not be possible without you. Thank you for your tireless work and sacrifices. I've learned so much from Dr. Luke Bornn, Dr. Mike Lopez, Dr. Dave Clarke and Dr. Katherine Evans. Thank you for all of your advice, teachings and opportunities. To that end, thank you to the entire Statistics department at SFU. Specifically, to Dr. Derek Bingham and Dr. Aaron Danielson for being

on my committee, and to Charlene, Sadika and Kelly for their consistent help with anything and everything.

I've been a part of 3 other major projects in my time at SFU. To the SFU Basketball research group, Coach Denis Beausoleil, Dr. Aaron Danielson, Kevin, Lucas thank you for the weekly chats and the great work trying to give our team every advantage possible. Thank you to Coach Hanson for being so receptive and engaged with our ideas. To the group, Dr. Ming-Chang Tsai, Dr. Jack Davis, Eli, Aaron and Dr. Dave Clarke, working with Canadian Sport Institute Pacific. It has been an absolute pleasure and I am excited to see Canada bring home gold medals in Japan in 2020. Thank you to Andy Peat, Stephen Jeske, Luke Summers and Drew Foster from the Vancouver Whitecaps for your support with our Club and our VanSASH event.

I've had the immense privilege of coaching basketball at the youth level throughout my studies. I'd like to thank the players at Dr. Charles Best Secondary School and TC North Basketball for their continuous work and love of the game. To my coaching mentors Coach Parkins, Coach Sokol, Coach Scott and Coach Van Os thank you for all of your lessons, support, and stories over the years.

My professional development has been huge due to great managers and mentors such as Matt St. John, Jorge Vasquez, Josh Orenstein, Arup Sen and Evan Wasch. To that same point, thank you to all of my incredible coworkers Domenic Fayad, Rhythm Tang, Emma Rong, Kareem Shouhdy, Jaime Rodriguez, Sam Garofalo, Rebecca Higgit, Patrick Harrel, Ally Blake, Lindell Galvan and Jolie Katz. Thank you to Ron Yurko and Asmae Toumi for their continuous and unwavering support.

Finally, thank you to my incredible girlfriend Juliana Garrone. Your determination, and strength each and every day is inspiring. Thank you for your constant love. Thank you to your wonderful family, Adrianne, Luigi, Nic, Lisa, Tracey and Rae for welcoming me with open arms. Thank you to my best friends Alecia, Rebecca, Aly, Tianna, Jen, Tegan, Marisa, Eileen, Chelsea, Kharis, Rachel, Nicole, Drew and Ness. Thank you for the sushi, the wine, the dancing, the improv, the cactus, the gossip and the camping. I am so thankful to have such an incredible group of friends.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

In the National Basketball Association (NBA), a player fouls out of the game and is disqualified from play after committing their sixth personal foul. Coaches, in an effort to maintain the availability of star players at the end of the game, will "sit" a player who has accumulated too many fouls in the early stages of the game.

NBA coaches have often followed the $Q + 1$ guideline when managing a player's foul trouble. That is when their number of fouls reaches one more than the quarter of the game the player is considered to be in foul trouble and is often subbed out of the game for the remainder of the quarter. For example, when a player attains two fouls in the first quarter of a game, the coach will remove the player for the rest of the quarter. Similarly, when a player attains three fouls during the second quarter of a game, the coach will remove the player from the game.

However, there may be an issue with the above coaching tradition. The above rule does not account for player level differences in fouling rates. Additionally by removing a player from the game, the coach is voluntarily limiting the playing time of the player. The benefit the coach does receive is the choice of when to play their player in exchange for how much the player will play. By removing the player from the game the coach is giving more weight to the player's value in end of game situations over the rest of the game.

The decision to treat all players the same and value their abilities more at the end of game does not appear to be informed by data. While the rule is widely followed, there is room for coaches to optimize their decisions further to get more out of their best players at the most valuable times throughout the game. This can be done by accounting for specific player tendencies. For example, it may be possible that a particular player is less foul prone than another and can continue to avoid fouling even when in perceived foul trouble.

In sporting practice, there exist traditions that are on the level of folklore, and upon closer inspection, do not appear optimal. For example, in hockey, the tradition had been for a team to pull its goalie when trailing with about one minute remaining in a match. However, it has been suggested through statistical modelling and simulation that goalies should be pulled with approximately three minutes remaining (Beaudoin and Swartz 2010). The new

recommendations appear to have made an impact in NHL practice (Davis and Lopez 2015). Another example of a misguided sporting tradition involves the over-reliance on popular baseball statistics such as batting average. As is well known, the moneyball phenomenon (Lewis 2013) highlighted alternative baseball measures such as on-base percentage which serve as better predictors of success. In the sport of football, Yam and Lopez (2018) use methods of causal inference to assess the impact of punting on fourth down in the National Football League.

Many cases where traditions do not appear optimal arise when teams are exhibiting risk-averse behaviours. For example, if a coach breaks from the traditional viewpoint and plays a star player in foul trouble we can imagine a series of possible outcomes. Firstly, the star player may foul out early and be unavailable at the end of the game. If the team loses by 3 points for example, blame will be placed on the coach for mismanaging the player's playing time under foul trouble. Secondly, the star player may not foul out early and be available at the the end of the game. The player will be credited for avoiding fouls. If the coach were to follow tradition we can imagine one more scenario. Since the star player was not on the court during large parts of the game the team ends up losing by 15 for example. The player will be blamed for picking up early fouls and the coach will be absolved of blame since they followed tradition. We can see that the coach will often not get credit for the decision that made the game closer but blamed for it. Even though the decision helped the team's chance of winning the credit is placed elsewhere. Following tradition however, has less risk for the coach since they are not breaking the status quo.

In this paper, we use data and statistical modelling to investigate various questions associated with foul accumulation in the NBA. For example, do all players foul at the same rate? Does the fouling time distribution between the $(n-1)$st foul and the nth foul, $n = 1, \ldots, 6$ depend on n? Are there differences in the fouling time distributions according to playing position? With respect to foul accumulation, can we advise coaches when to sit their players?

The topic of NBA substitution patterns is a topic that has been mostly discussed on blog sites. For example, Rochford (2017) uses item response theory and Bayesian modelling to draw various insights with respect to NBA fouls. In particular, Rochford (2017) draws attention to the relationship between fouling and salary with the suggestion that higher paid players are treated preferentially with respect to foul calls. Klobuchar (2018) investigates the impact on win shares from the "early" substitution of players due to fouls. Falk (2018) examines playing minutes when a coach employs a foul management strategy. The investigation suggests different strategies, including changing a player's defensive assignment to decrease a player's foul rate and rearranging the player's minutes. Pomeroy (2016) examines how players and coaches define foul trouble in college basketball. He finds self preservation behaviour in players who are considered to be in foul trouble. Partnow (2019) lists a number of factors to be considered when deciding on when and for how long players

in foul trouble should be removed from the game. He notes that "coaches in effect 'foul out' their own players early in the game, possibly taking the team out of the running for crunch time, or putting the squad in a closer game, with greater chances to lose, than would have been the case had the star played close to [their] normal minutes." An example of the other factors that he suggests considering are the foul rates of the player in question, how well the player plays while in perceived foul trouble, does the game itself have an elevated foul rate due to physicality or refereeing, and finally how good is the backup player. In the journal article by Maymin, Maymin and Shen (2012), the impact of early foul trouble is assessed using tools from finance. Of note, Maymin, Maymin and Shen (2012) suggest that teams exhibit poorer performance if they continue to play foul-plagued starters. Evans (2017b) proposes a conditional risk set model for ordered events to model a player's time to foul while including covariates such as the point differential, and time remaining in the game among others.

In Chapter 2, we begin with an exploratory data analysis where we assess the conjecture that the fouling time distribution is Exponential. We suggest that the Gamma distribution provides a more realistic fouling time distribution. We also investigate the impact of player position and the impact of foul level on the fouling time distribution. In Chapter 3, we use the Gamma distribution to build a stochastic model which incorporates unknown parameters. The model is Bayesian and requires the specification of prior distributions and computational strategies to assess the parameters. The exploratory analysis in Chapter 2 helps us specify the prior distribution in Chapter 3. A predictive distribution is then introduced which may be used by coaches. The models are implemented on NBA data in Chapter 4 where interesting insights are obtained with respect to the fouling tendencies of players. Additionally, we validate the results of the predictive inference and compare them to other models. We then demonstrate how our model can be useful in real in-game scenarios We conclude with a short discussion of the implications of our work and areas for further reasearch in Chapter 5. The material in this MSc project is an extension of Chu and Swartz (2019). As such, large passages from Chu and Swartz (2019) are included in this project.

Chapter 2

Exploratory Data Analysis

2.1 The Fouling Time Distribution

It has been suggested that the Exponential distribution may be appropriate for the distribution of fouling times (Evans 2017a).

Some simple thought experiments reveal that the validity of this assumption is too simplistic. For example, if the time between fouls is Exponential, then fouling time satisfies the *memoryless property*. This implies that a player who has just stepped onto the court has the same probability of fouling within a period of time compared to the situation where the player had been on the court for a longer period of time. The memoryless property seems suspect as a tired basketball player may have difficulty moving his feet into a good defensive stance, and is therefore more likely to commit a foul than a fresh player. On the other hand, it may be argued that a fresh player may not be in the flow of the game and may become overly aggressive, and more prone to foul than a player who has been on the court for a while.

In order to test the suitability of the Exponential distribution, we introduce some standard failure time notation. Consider then a specific player who has committed his $(n-1)$st foul, and this occurs in a match which we label the jth match. We denote $X_j^{(n)}$ as the time played between the $(n-1)$st and nth foul, $n = 1, \ldots, 6$ and $j = 1, \ldots, m_n$. It is therefore apparent that $m_1 \geq m_2 \geq \cdots \geq m_6$. It is possible that the time to foul $X_j^{(n)}$ is unobserved and there is a potential censoring time $C_j^{(n)}$. In this case, the corresponding observed dataset for the player at the nth *foul level* is given by $(Y_1^{(n)}, \delta_1), \ldots, (Y_{n_m}^{(n)}, \delta_{n_m})$ where $Y_j^{(n)} = \min(X_j^{(n)}, C_j^{(n)})$ and

$$\delta_j = \begin{cases} 0 & X_j^{(n)} \leq C_j^{(n)} \quad \text{(uncensored)} \\ 1 & X_j^{(n)} > C_j^{(n)} \quad \text{(censored)} \end{cases} .$$

In this application, it is important to note that the censoring mechanism involves random right censoring rather than fixed right censoring. Should a player not commit the nth foul,

$n = 1, \ldots, 6$, we can think of the nth foul as randomly censored. In a medical application with fixed right censoring, this corresponds to an experiment which concludes at the same time for all subjects. A detailed treatment of the statistical analysis of failure time data is given by Kalbfleisch and Prentice (2002).

In our investigation of the Exponential as a fouling time distribution, we first note that the Exponential is a special case of the Gamma distribution. Here, we use the parameterization $X \sim \text{Gamma}(\alpha, \beta)$ such that $\text{E}(X) = \alpha/\beta$. We consider alternative data to that used in Chapter 4 to avoid the perils of "double use of the data". Specifically, we consider data from the 2012/2013 NBA regular season involving players at the jth foul level who have $m_j \geq 30$ observations. This provides 1,010 player-foul combinations involving 376 unique players. To test the fit of the Exponential distribution, we carry out likelihood ratio tests of $\text{H}_0 : \alpha = 1$ at the 0.05 level of significance for each of the 1,010 datasets. Expressions for the maximum likelihood estimators of α and β under right random censoring are given by Harter and Moore (1965) and are estimated via an iterative procedure. We reject 20% of the null hypotheses; this provides evidence that the Exponential may not generally be appropriate as a fouling time distribution. More powerful goodness-of-fit tests such as tests based on the empirical distribution function (see D'Agostino and Stephens 1986) would likely result in higher rates of rejection of the null hypotheses.

The question then arises as to whether the Gamma is appropriate as a fouling time distribution. We have investigated the 2012/2013 data by comparing the survival curves (based on maximum likelihood estimation under right random censoring) versus the associated non-parametric Kaplan-Meier estimate. In the cases that we have studied, the fit appears adequate. In Figure 2.1, we provide the corresponding plot for Paul Pierce at foul level $n = 1$. Pierce's data is based a sample size of $m_1 = 77$ where only 3 of the observations were censored. The survival curve for Pierce agrees nicely with Pierce's Kaplan-Meier estimate.

Returning to our original conundrum involving the Exponential distribution, we found that $\hat{\alpha} > 1$ for most player-foul combinations (821 out of the 1,010 cases). This indicates an increasing hazard function. In basketball terms, this means that a player is more likely to foul when they are on the court for longer periods of time. For the remainder of our investigation, we will use the Gamma as the fouling time distribution. As a two-parameter distribution which includes the Exponential, the Gamma is a safer choice in the sense that it has more flexibility to accommodate various distributional shapes.

2.2 The Impact of Player Position

There is a perception that the accumulation of fouls may depend on player position. To investigate this notion, we considered NBA data from four recent seasons, 2013-2014 through 2016-2017. We restrict the analysis to players who have accumulated more than five fouls at

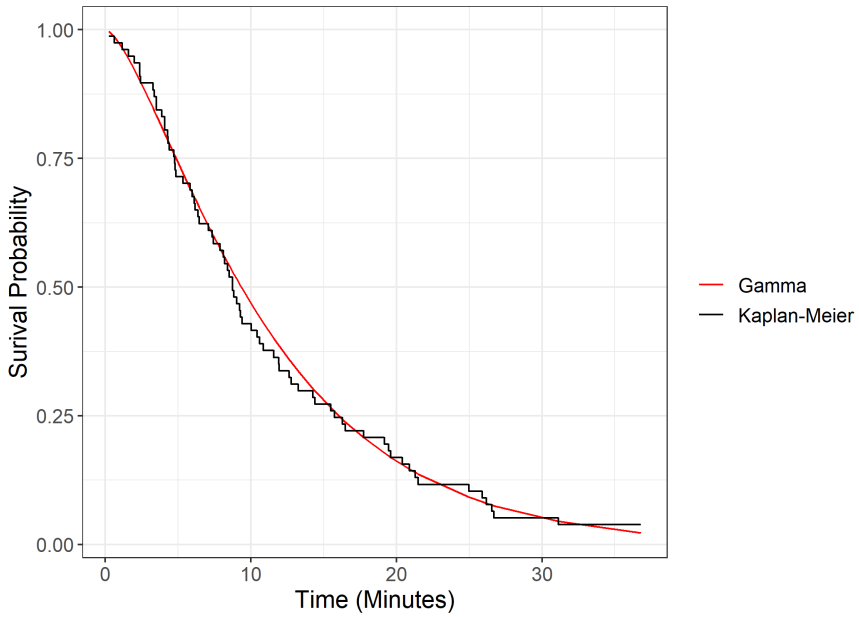

Figure 2.1: Estimated survival function (based on the Gamma distribution) and the Kaplan-Meier estimate corresponding to the first foul level for Paul Pierce based on data from the 2012/2013 regular season.

each foul level over the three seasons. For the ith player, we used the maximum likelihood procedure previously discussed to estimate the Gamma parameters α_i and β_i. The boxplots in Figure 2.2 are based on the estimated mean fouling times $\hat{\alpha}_i/\hat{\beta}_i$ where the boxplot categories correspond to playing positions. Each player was classified according to one of three standard positions, namely bigs, forwards and guards. From Figure 2, we observe that bigs foul the most quickly, followed by forwards, and then followed by guards.

2.3 The Impact of Foul Level

There is a second perception that the accumulation of fouls may depend on the foul level. We again used the 2013-2014 through 2016-2017 NBA data but for each foul level $n = 1, 2, \ldots, 6$, we restricted the analysis to players who accumulated more than five fouls and played at least 1640 minutes in the 2016-2017 season. This restriction was carried out so that the maximum likelihood estimates of α_i and β_i are reliable. The boxplots in Figure 2.3 are based on the estimated mean fouling times $\hat{\alpha}_i/\hat{\beta}_i$ where the boxplot categories correspond to foul levels. From Figure 2.3, we observe a clear trend that fouls occur more quickly for increasing foul levels n.

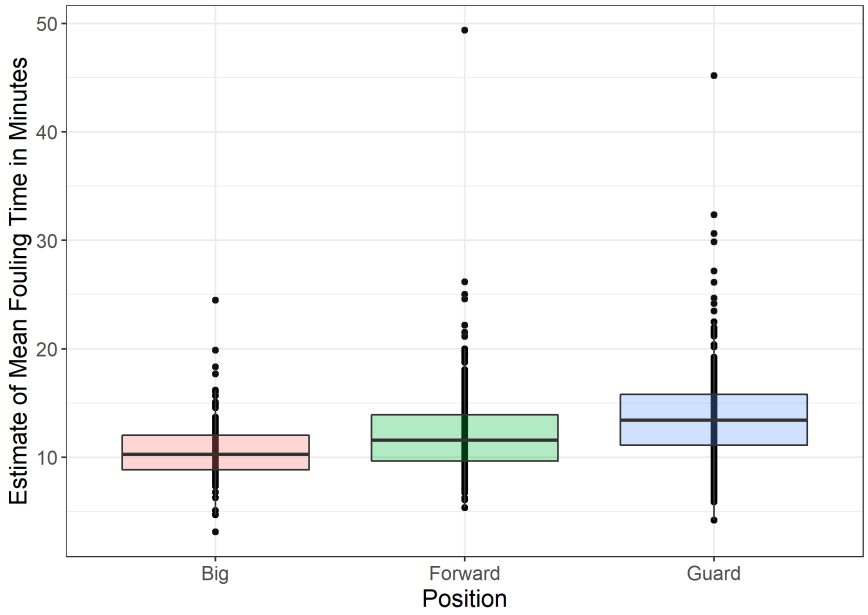

Figure 2.2: Boxplots of the estimated mean fouling times for each of the standard basketball playing positions.

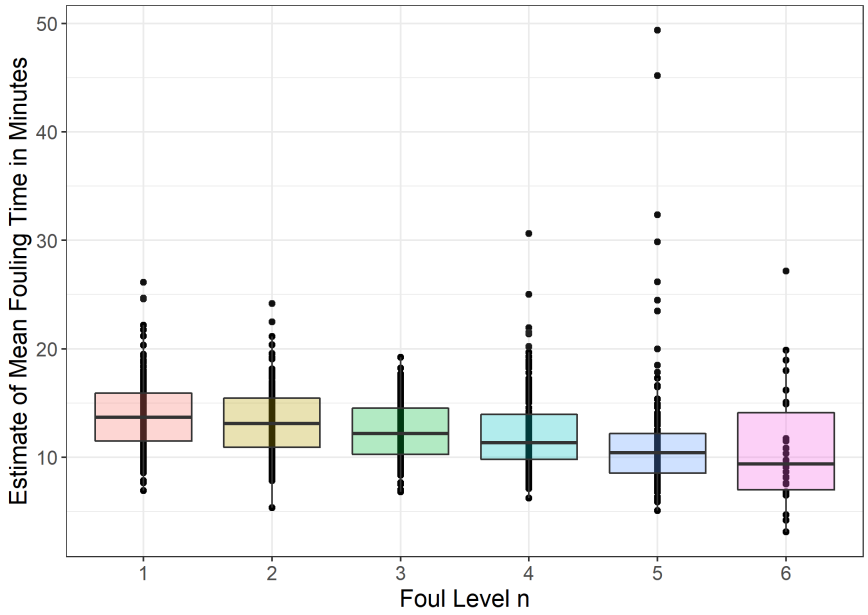

Figure 2.3: Boxplots of the estimated mean fouling times for foul levels $n = 1, 2, \ldots, 6$.

Chapter 3

Modelling

We expand the notation of Chapter 2 where we let $X_{ij}^{(n)}$ denote the time played between the $(n-1)$st and nth foul for the ith player in the jth match, $n = 1, \ldots, 6$. Similarly, we define $Y_{ij}^{(n)} = \min(X_{ij}^{(n)}, C_{ij}^{(n)})$ where $C_{ij}^{(n)}$ and δ_{ij} are the corresponding potential censoring times and indicator variables, respectively. With $X_{ij}^{(n)} \sim \text{Gamma}(\alpha_{in}, \beta_{in})$, this leads to the posterior density

$$\pi(\alpha, \beta \mid y) \propto \prod_i \prod_j \prod_n f(y_{ij}^{(n)} \mid \alpha_{in}, \beta_{in})^{1-\delta_{ijn}} \; [1 - F(y_{ij}^{(n)} \mid \alpha_{in}, \beta_{in})]^{\delta_{ijn}} \; \pi(\alpha, \beta) \tag{3.1}$$

where vector notation is utilized, f and F are the density and cumulative distribution functions corresponding to the Gamma distribution and $\pi(\alpha, \beta)$ is the prior density. Here, interest concerns the unknown parameters α and β which describe the fouling time distributions.

3.1 Prior Distribution

With 30 players of interest (see Chapter 4) and six foul levels $n = 1, \ldots, 6$, this leads to $2(30)(6) = 360$ parameters α and β in (3.1). In hierarchical models, we can effectively reduce the parameterization by borrowing information between parameters. We let p_i denote the position of player i where p_i takes on the values 1 (denoting big), 2 (denoting forward) and 3 (denoting guard). We have seen from the exploratory data analysis that the fouling time distributions depend on both position and foul level. We therefore consider a prior structure where $(\alpha_{in}, \beta_{in})$ arise from a distribution that depends on both the player position p_i and the foul level n.

We implemented the prior structure by imposing independence between the $(\alpha_{in}, \beta_{in})$ pairs and specifying

$$(\alpha_{in}, \beta_{in})' \sim \text{truncated\_Normal}_2((a, b)', \Sigma) \tag{3.2}$$

where $\Sigma = (\sigma_{ij})$. The truncations on the bivariate Normal distributions are imposed so that $\alpha_{in} > 0$ and $\beta_{in} > 0$ according to the definition of the Gamma distribution. We have used some simplifying notation in (3.2) where it is emphasized that the hyperparameters a, b and Σ depend on the combination of the player position p_i and the foul level n.

The hyperparameters (a, b) were informed from the "old" 2013-2014 through 2016-2017 regular season data. At each foul level $n = 1, \ldots, 6$, we first obtained maximum likelihood estimates (mles) $\hat{\alpha}_{in}$ and $\hat{\beta}_{in}$ of the Gamma parameters for all players who accumulated more than five fouls and played at least 1640 minutes in the 2016-2017 season. We then grouped the mles accordingly to the $3(6) = 18$ combinations corresponding to player position and foul level. The hyperparameters a and b were then determined by averaging the values of $\hat{\alpha}_{in}$ and $\hat{\beta}_{in}$ in each group. For the specification of the hyperparameter matrix Σ, we proceeded in the same fashion by calculating the second moments corresponding to $\hat{\alpha}_{in}$ and $\hat{\beta}_{in}$ in each group.

There is one exception to the hyperprior specification described above. We grouped the case $(p_i = 1, n = 5)$ with $(p_i = 1, n = 6)$, we grouped the case $(p_i = 2, n = 5)$ with $(p_i = 2, n = 6)$, and we grouped the case $(p_i = 3, n = 5)$ with $(p_i = 3, n = 6)$. This was carried out because there were fewer fouls at the higher foul levels $n = 5, 6$, and grouping provided more reliable estimation.

We introduced one additional feature in the prior specification based on the discovery from Chapter 2.3. We impose the constraint $\alpha_{i1}/\beta_{i1} \geq \alpha_{i2}/\beta_{i2} \geq \cdots \geq \alpha_{i6}/\beta_{i6}$ to reflect our knowledge that the mean fouling time decreases with increasing n. Further, if we believe that fouling time distributions have an increasing hazard function, then we may introduce the constraint $\alpha_{in} \geq 1.0$. Additional model restrictions may be useful when data used to inform the prior is not abundant.

3.2 Predictive Distributions

In the Bayesian setting, there is a convenient framework for handling predictive inference. Suppose that we are interested in the predictive distribution for the playing time $X_i^{(n)*}$ between the $(n-1)$st foul and the nth foul for player i. The density for the predictive distribution of $X_i^{(n)*}$ is given by

$$f(x) = \int f(x \mid \alpha_i, \beta_i)\ \pi(\alpha, \beta \mid y)\ d\alpha\ d\beta \tag{3.3}$$

where y denotes the historical data used in the determination of the posterior (3.1).

Fortunately, obtaining a sample from the predictive distribution (3.3) is a by-product of Markov chain Monte Carlo (MCMC). In the kth iteration of MCMC, we generate the parameter vector $(\alpha^{(k)}, \beta^{(k)})$. We then generate $x^{(k)} \sim f(x \mid \alpha_i^{(k)}, \beta_i^{(k)})$. Repeating the procedure,

we have a sample $x^{(1)}, x^{(2)}, \ldots, x^{(N)}$ from the predictive distribution. As demonstrated in Chapter 4, the sample allows us to address various questions associated with fouling times.

3.3 Computation

With complex high-dimensional posterior distributions, one typically resorts to sampling-based methods to approximate posterior summaries. In this application, we use MCMC methods to generate variates from the posterior. In particular, we use the Bayesian software package Stan which is relatively simple to use and avoids the need of special purpose MCMC code. In Stan, the user only needs to specify the likelihood, the prior and the data; the determination of appropriate proposal distributions and sampling schemes are done in the background. Stan is open source software (https://mc-stan.org) and can be accessed through RStan (https://mc-stan.org/rstan/) which is the R interface to Stan. For example, if we are able to generate variates $\alpha_{in}^{(1)}, \ldots, \alpha_{in}^{(N)}$ from the posterior (1), then $\hat{\alpha}_{in} = (1/N) \sum_{k=1}^{N} \alpha_{in}^{(k)}$ provides an estimate of the posterior mean of α_{in}.

Chapter 4

Results

Data were taken from the Eight Thirty Four website (Evans and Saini 2019) which consists of enhanced play-by-play data from the 2012-2013 through 2018-2019 NBA regular seasons.

Recall that the 2013-2014 through 2016-2017 NBA regular season data were used to specify the prior distribution. We now consider the posterior density (1) based on data from the 2017-2018 and 2018-2019 NBA regular seasons. We use fouling time data corresponding to the 30 players listed in Table 4.1. The players consisted of 10 bigs, 10 forwards and 10 guards. The players selected were those who had the most minutes of playing time at their respective positions during the 2017-2018 season.

Bigs			Forwards			Guards		
Player	Team	Fouls	Player	Team	Fouls	Player	Team	Fouls
R Drummond	Pistons	249	G Antetokounmpo	Bucks	231	B Beal	Wizards	160
C Capela	Rockets	185	H Barnes	Mavericks	94	C McCollum	Trailblazers	168
D Jordan	Clippers	203	J Ingles	jazz	178	D Lillard	Trailblazers	117
J Nurkic	Trailblazers	247	K Middleton	Bucks	270	D DeRozan	Raptors	151
K Towns	Timberwolves	285	L James	Cavaliers	136	D Mitchell	Jazz	213
M Gasol	Grizzlies	185	P George	Thunder	233	J Holiday	Pelicans	201
M Gortat	Wizards	175	R Covington	Sixers	238	K Walker	Hornets	98
N Jokic	Nuggets	212	T Gibson	Timberwolves	218	L Williams	Clippers	106
S Adams	Thunder	215	T Young	Pacers	175	R Westbrook	Thunder	200
W Cauley-Stein	Kings	185	T Harris	Pistons/Clippers	164	W Barton	Nuggets	168

Table 4.1: For the three positions, we list the players studied from the 2017-2018 season and include the total fouls accumulated.

4.1 Example: Giannis Antetokounmpo - predictive distributions

We illustrate the fouling tendencies of Giannis Antetokounmpo of the Milwaukee Bucks based on his fouling data from the 2017-2018 and 2018-2019 regular seasons. Following Chapter 3.2, we approximated predictive distributions for the fouling times at the foul levels $n = 1, \ldots, 6$. The estimated predictive densities are shown in Figure 4.1. The densities are based on 3,000 draws from the predictive distributions which are estimated by the func-

tion *geom_density_ridges* from the *ggridges* package in R. We observe that the predictive densities have long right-skewed tails indicating that there is possibility of playing a long time without fouling. Like all players, we further observe that Giannis fouls quicker at later foul levels. For example, the mean predictive fouling time for Giannis is 11.6 minutes for his first foul and 10.9 minutes for his third foul.

Referring to Table 4.1, Harrison Barnes is also a forward but he fouls much less frequently than Giannis Antetokounmpo. We observe that the mean predictive fouling time for Barnes is 23.8 minutes for his first foul and 16.8 minutes for his third foul.

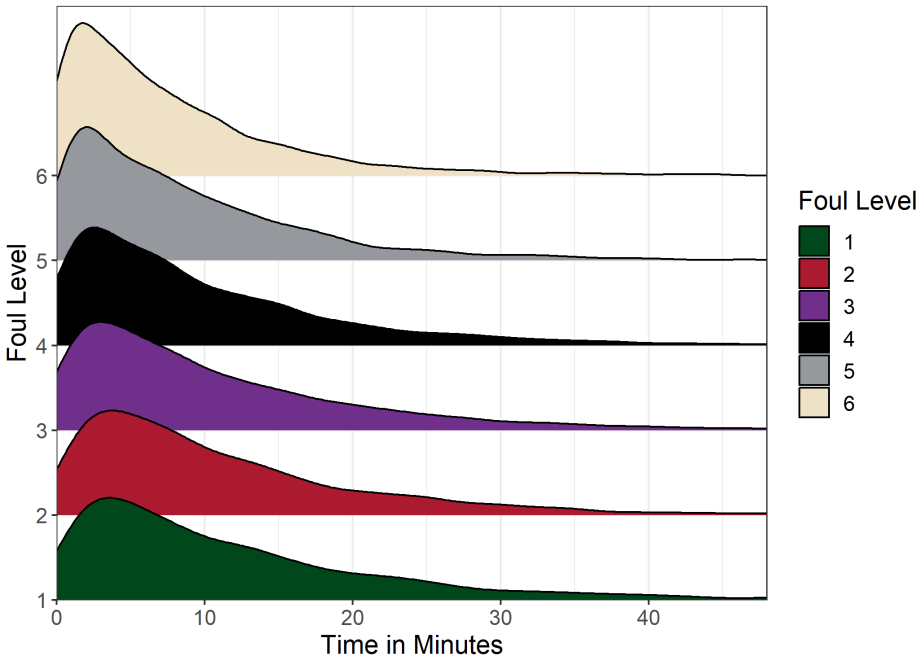

Figure 4.1: Predictive densities of the fouling time distribution for Giannis Antetokounmpo at the foul levels $n = 1, \ldots, 6$ based on data from the 2017-2018 and 2018-2019 regular seasons.

4.2 Example: Lebron James - endgame scenario

Lebron James has been a star NBA player for his entire career. Any coach of Lebron would like to see him playing at the end of a match where the outcome is in the balance. Let's imagine that Lebron has picked up his fifth foul midway through the third quarter where there is 18 minutes left to play. Should Lebron's coach force Lebron to sit or should he continue to play? Based on the MCMC output, Lebron's estimated posterior mean time for the sixth foul is 9.9 minutes (2.2 minutes longer than Giannis). However, the mean fouling time does not provide a complete picture for the problem at hand. We use the MC algorithm and the predictive distribution (3) to generate fouling times $x^{(1)}, x^{(2)}, \ldots, x^{(N)}$

corresponding to Lebron's sixth foul. The 10th percentile of the predictive sample based on $N = 1,000$ is 1.3 minutes. Therefore, if the coach wants to be 90% confident that Lebron is playing at the end of the match, the coach should force Lebron to sit and re-enter the match with 1.3 minutes remaining. This strikes us as an overly conservative strategy, where we suggest that coaches ought to be willing to have Lebron re-enter the match earlier than 1.3 minutes remaining. For reference, the 30th and 50th percentiles for Lebron are 4.2 and 7.5 minutes, respectively.

The same scenario and analysis is considered for another impact player, Karl-Anthony Towns. If his coach wants to be 90% confident that Towns is playing at the end of the match having committed his fifth foul, the coach should force Towns to sit and re-enter the match with 0.8 minutes remaining. The 30th and 50th percentiles for Towns are 2.6 and 4.8 minutes, respectively.

4.3 Example: Damian Lillard - cumulative fouls

We provide a third example which further illustrates the convenience of simulation-based inference using the proposed model. Following the description of the generation of predictive variates in Chapter 3.2, suppose we are interested in the total time T that Damian Lillard can play following his second foul. If x_j is the predicted time between the $(j-1)$st foul and the jth foul, then our interest concerns $T = x_3 + x_4 + x_5 + x_6$.

In Figure 5, we provide the survival curve corresponding to T for Damian Lillard. We observe that Lillard's median time for fouling out exceeds 48 minutes (ie. the length of a match). Therefore, in the case of Damian Lillard, it may be unnecessarily cautious for coaches to follow the $Q + 1$ rule.

4.4 Example: Victor Oladipo - playoff scenario

We provide a fourth example which examines an actual game between the fourth seeded Cleveland Cavaliers and the fifth seeded Indiana Pacer in their series from the 2017-2018 NBA finals. This scenario was described by Falk (2018) and was the motivation for his discussion of the use of the $Q + 1$ rule.

We suggest developing a "coach's cheat sheet". This denotes the probability of fouling out for a given player given a budget of minutes. The probability of fouling out is calculated from the predictive distribution. The survival curve is plotted as a colour gradient to condense the information into a smaller area and to make the information more visually appealing for a coach. A coach can discuss with their staff pre-game to determine their level of risk. We present an example for Victor Oladipo in Figure 4.3. The cheat sheet could have been used to help inform decisions throughout the upcoming scenario.

In Game 2 of the series, Victor Oladipo the star player of the Indiana Pacers picked up his second foul at 1:02 into the first quarter. Following his second foul in the first quarter,

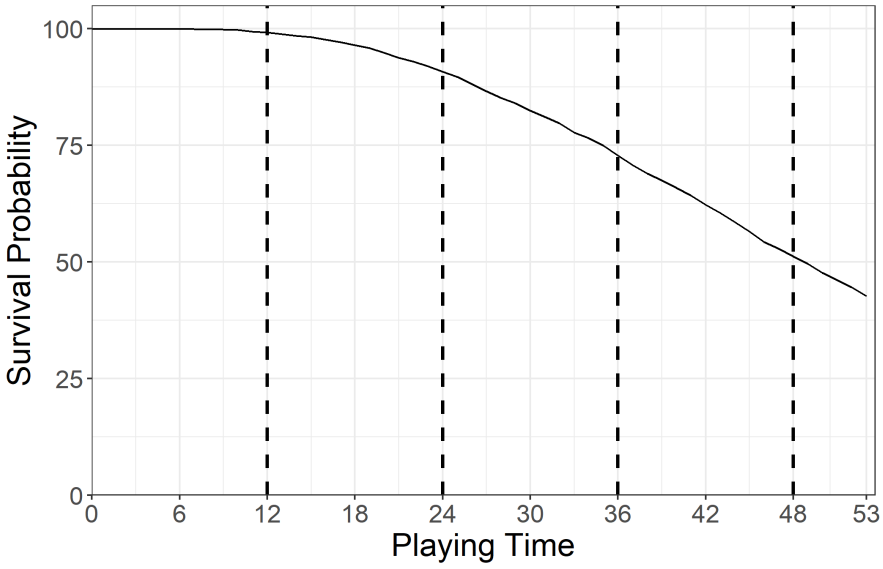

Figure 4.2: Survival curve of the total playing time $T = x_3 + x_4 + x_5 + x_6$ prior to fouling out for Damian Lillard after he has committed his second foul.

Head Coach of the Indiana Pacers, Nate McMillan, removed Oladipo from the game. He was re-inserted into the game at the start of the second quarter with 36 minutes remaining in the game.

We use the predictive distributions of Oladipo's time to foul out following his second foul with data from the 2016-2017 and 2017-2018 seasons in order to avoid using data from the game in question. We predict that Oladipo would have been able to play the 42 minutes (the most minutes he played in a non-overtime game that season) in the game 42% of the time without fouling out. In the "coach's cheat sheet" under "Next Foul: 3" the foul out percentage is roughly 60%, or a survival percentage of 40%. Instead, Oladipo sat on the bench until it was impossible to reach the 42 minute total. When he was reinserted into the game we predicted that he would be able to play the remaining 36 minutes (assuming no substitutions for rest) 53% of the time without fouling out.

When Oladipo re-entered the game the Pacers were losing 33-18. The Pacers cut the lead to 41-35 when Oladipo picked up his third foul, with 5 minutes left in the second quarter (29 minutes left in the game). Oladipo was again removed from the game as he had three fouls in the second quarter and sat until the end of the half. When Oladipo picked up his third foul, we predicted that he'd be able to play 24 minutes (the 5 remaining minutes of the second quarter and the 19 minutes he'd play in the third and fourth quarters) without fouling out 50% of the time. By sitting the 5 minutes in the second quarter Oladipo's probability of surviving without picking up his sixth foul increased to 66% for the 19 minutes that he played in the second half.

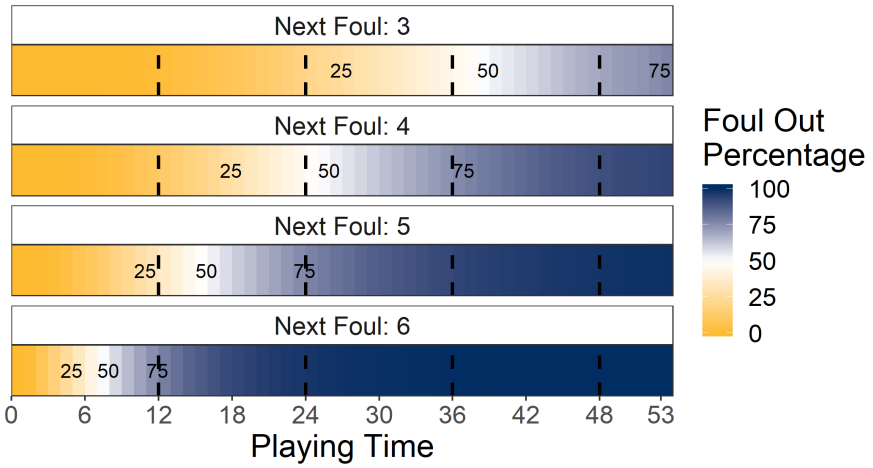

Figure 4.3: Example of a coach's cheat sheet for Victor Oladipo.

Oladipo finished the game with three fouls as he did not pick up another foul for the rest of the game. We wonder if Coach McMillan understood how each minute of playing time affected Oladipo's probability of fouling out? If not, would Coach McMillan have played Oladipo more minutes? That is, would Coach McMillan trade an 11 minutes in the first quarter and 5 minutes in the second quarter for an 11% and 16% increase in probability of survival respectively.

We also note that the distribution for the time to fouling out is right skewed so an appetite for more risk, especially earlier in the game, creates the possibility for a bigger reward.

4.5 Model Validation

It is important to validate the accuracy of our model. In particular we compare the accuracy of our predictive distributions to predictions from three other sources, a baseline survival probability of 0.5, Kaplan-Meier curves, and maximum likelihood estimates for a Gamma distribution.

For assessing the accuracy of the models we consider the integrated time-dependent expected Brier score as suggested by Mogensen, Ishwaran and Gerds (2012) and adjusted for right censoring times as described by Gerds and Schumacher (2006). The Brier score evaluates the accuracy of a predicted survival function for any given time t. For a given time t, it is the squared error between the observed survival status (0,1) and predicted survival probability of an observation. Under right censoring, we adjust the Brier score by weighting the squared error using the inverse probability of censoring weights method. The

time-dependent expected Brier Score using this adjustment is given by

$$\hat{B}(t) = \frac{1}{K}\sum_{i=1}^{30}\sum_{j}\sum_{n}\left(\frac{\left(0-\hat{S}_i^{(n)}(t)\right)^2 \cdot 1_{y_{ij}\leq t,\delta_i=1}}{\hat{G}(y_{ij})} + \frac{\left(1-\hat{S}_i^{(n)}(t)\right)^2 \cdot 1_{y_{ij}>t}}{\hat{G}(t)}\right) \tag{4.1}$$

where $\hat{G}(t) = P[C > t]$ is the estimator of the conditional survival function of the censoring times calculated using the Kaplan-Meier method, K is the total number of observations, $\hat{S}_i^{(n)}(t)$ is the estimated survival function given by the predictive distribution for player i and foul level n.

We calculate $\hat{B}$ for $t = 0, \ldots, 48$ minutes and approximate the integral with respect to time t to obtain the integrated time-dependent expected Brier score. The integrated time-dependent expected Brier scores for each methodology are calculated from out of sample predictions in the 2018-2019 NBA season from models that were fit on data from the 2016-2017 and 2017-2018 NBA seasons and reporter in Table 4.2.

Foul Level	Bayesian	Kaplan-Meier	MLE	0.5 Baseline
1	0.123	0.123	0.123	0.250
2	0.132	0.133	0.133	0.251
3	0.136	0.137	0.138	0.251
4	0.161	0.165	0.167	0.245
5	0.147	0.155	0.159	0.222
6	0.160	0.199	0.150	0.221

Table 4.2: Integrated time-dependent expected Brier Score by foul level for each model.

Our Bayesian model outperforms the other models for all cases except for foul level 6, where the maximum likelihood estimates provide the best out of sample estimates.

Chapter 5

Discussion

This paper introduces parametric models in a Bayesian framework for the analysis of fouling time distributions. The problem is important since NBA players foul differently, and coaches should take this into account when managing playing time for players in foul trouble.

Some of the messages in this paper include (1) that the Gamma distribution provides a flexible and appropriate distribution for fouling times, (2) that mean fouling times decrease across the positional types given by bigs, forwards and guards, and (3) that mean fouling times decrease as more fouls are accumulated. Future work may consider the impact of consecutive playing time versus segmented playing time, and other covariates such as those suggested by Evans (2017a, 2017b).

Additionally, another avenue of future research could analyse the value of each part of the game. In other words, are there parts of the game where a player will have more of an impact on the probability of winning than others? Then one could combine both analyses to determine playing time for a player in foul trouble.

We envision our formalization of player foul rate estimation to be a launching point for future analysis from a statistical point of view and in depth discussions about player rotations from basketball minds. For those coming from a basketball perspective there are a couple of other factors to consider. It is most likely beneficial to remove a player from the game immediately after they are perceived to be in foul trouble for a variety of reasons. Evans (2017b) notes that players who have just been rested are less likely to foul. Pomeroy (2016) mentions that players have self preservation habits with respect to foul trouble and Maymin, Maymin and Shen (2012) found similar results with respect to other aspects of a players performance. Therefore, the perception of foul trouble should not be ignored. We still support the notion of removing the player perceived to be in foul trouble for a short period of time. However, we suggest understanding the player's foul habits and in most cases returning the player to play before the usual guideline suggests. Coaches should also take into account that players will naturally get rest on the bench through regular player rotations and should account for this when budgeting a player's remaining minutes. Finally, our estimation is done with respect to a player's usual of defensive assignments. Therefore,

a coach may be able to change a player's underlying foul rate by changing the player's defensive assignment, or the team's defensive scheme.

It is our hope that the methods presented here may help NBA teams make better substitution decisions. Should teams implement the methods, we suggest that they remove intentional fouls from the dataset. Although intentional fouls are infrequent, they should not be included as they do not characterize individual fouling behaviour.

Finally, while this analysis was done with respect to NBA data we are excited by the possibility of quantifying foul rates for the Women's National Basketball Association (WNBA), Women's and Men's College Basketball and International Basketball competitions.

Bibliography

Beaudoin, D. and Swartz, T.B. (2010). "Strategies for pulling the goalie in hockey", *The American Statistican*, 64, 197-204.

Chu, D. and Swartz, T.B. (2019). "Foul Accumulation in the NBA", *Submitted*

Evans, K. (2017a). "NBA Fouls - I love DeMarcus Cousins", accessed September 20, 2019 at https://causalkathy.com/2017/01/31/i-love-demarcus-cousins/

Evans, K. (2017b). "NBA Fouls - Survival Analysis", accessed November 3, 2019 at https://causalkathy.com/2017/02/21/nba-fouls-survival-analysis/

Evans, K. and Saini U. (2019). "Eight Thirty Four", accessed November 11, 2019 at https://eightthirtyfour.com/data

D'Agostino, R.B. and Stephens, M.A. (1986). *Goodness-of-Fit Techniques*, Marcel Dekker, New York.

Davis, N. and Lopez, M. (2015). "NHL coaches are pulling goalies earlier than ever", *FiveThirtyEight*, accessed September 20, 2019 at https://fivethirtyeight.com/features/nhl-coaches-are-pulling-goalies-earlier-than-ever/

Falk, B. (2018). "The trouble with foul trouble", *Cleaning The Glass*, accessed September 22, 2019 at https://cleaningtheglass.com/the-trouble-with-foul-trouble/

Gerds, T. and Schumacher, M. (2006). "Consistent Estimation of the Expected Brier Score in General Survival Models with Right-Censored Event Times", Biometrical Journal, 48(6), 1029–1040. https://doi.org/10.1002/bimj.200610301

Harter, H.L. and Moore, A.H. (1965). "Maximum-likelihood estimation of Gamma and Weibull populations from complete and censored samples", *Technometrics*, 7, 639-643.

Kalbfleisch, J.D. and Prentice, R.L. (2002). *The Statistical Analysis of Failure Time Data, Second Edition*, John Wiley & Sons, Inc., Hoboken, New Jersey.

Klobuchar, A. (2018). "Basketball foul analysis: A look at the inefficiency in current NBA substitution-after-fouls behavior", *Aidan's Musings*, accessed September 22, 2019 at http://aidanklobuchar.com/home/data-based-writings/fouls/

Lewis, M. (2013). *Moneyball: The Art of Winning an Unfair Game*, WW Norton, New York.

Maymin, P., Maymin, A. and Shen, E. (2012). "How much trouble is early foul trouble? Strategically idling resources in the NBA", *International Journal of Sport Finance*, 7, 324-339.

Mogensen, U. B., Ishwaran, H., and Gerds, T.A. (2012). "Evaluating Random Forests for Survival Analysis using Prediction Error Curves". Journal of statistical software, 50(11), 1–23. doi:10.18637/jss.v050.i11

Partnow, S. (2019). "Partnow November Mailbag, Part 2: The Turduckening (in seven seconds or less)", *The Athletic*, accessed November 29, 2019 at https://theathletic.com/1413858/2019/11/28/partnow-november-mailbag-part-2-the-turduckening-in-seven-seconds-or-less/

Pomeroy, K. (2016). "Foul trouble as defined by players", *kenpom*, accessed November 29, 2019 at https://kenpom.com/blog/foul-trouble-as-defined-by-players/

Rochford, A. (2017). "NBA foul calls and Bayesian item response theory", accessed September 22, 2019 at https://austinrochford.com/posts/2017-04-04-nba-irt.html

Yam, D. and Lopez, M. (2018). "Quantifying the causal effects of conservative fourth down decision making in the National Football League", Available at http://dx.doi.org/10.2139/ssrn.3114242