ABSTRACT

Bayesian Method of Predicting In-Game Win Probability Across Sports

Jason Maddox, Ph.D.

Mentor: Jane Harvill, Ph.D.

In this dissertation, we create different in-game win probability models for several sports using a Bayesian methodology. In the first chapter, we create a college basketball model using score differential and time and compare the model to other models found in literature. In the second chapter, we extend the model from the first chapter into the NBA. In doing so, we also make adjustments to aid in the performance of the model. In the third chapter, we create a college football win probability model, accounting for many more factors than the score differential and time.

Bayesian Method of Predicting In-Game Win Probability Across Sports

by

Jason Maddox, B.S., M.S.

A Dissertation

Approved by the Department of Statistical Science

James Stamey, Ph.D., Chairperson

Submitted to the Graduate Faculty of
Baylor University in Partial Fulfillment of the
Requirements for the Degree
of
Doctor of Philosophy

Approved by the Dissertation Committee

Jane Harvill, Ph.D., Chairperson

David Kahle, Ph.D.

James Stamey, Ph.D.

Philip Young, Ph.D.

Ryan Sides, Ph.D.

Accepted by the Graduate School
December 2022

J. Larry Lyon, Ph.D., Dean

Page bearing signatures is kept on file in the Graduate School.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Thank you Dr. Harvill for her dedication to me and her willingness to branch out into unorthodox fields of research. Thank you also to all pioneers before me that have shaped the field of Sport Analytics into what it is today, making our research possible.

DEDICATION

To my wife, Natasha, who has challenged and encouraged me all along the way. This dissertation may have never come to light without her. I would also like to dedicate this to my parents who have thoroughly supported me throughout my entire journey of life.

CHAPTER ONE

Introduction and Literature Review

1.1 College Basketball Literature Review

Sports analytics is not a new science. Work spans more than 30 years and a large range of difficulty. Since the early 2000s, research in statistical methods for sports analytics has risen dramatically. The review articles of Kubatko et al. (2007), Santos-Fernandez et al. (2019), and Terner and Franks (2021) provide a fairly comprehensive review for sports analytics for a wide variety of sports, including basketball.

Generally speaking, models for predicting the outcome of a sporting event can be classified into two systems: (1) pregame prediction or (2) in-game, or in-play, prediction. Pregame prediction involves determining the outcome of a game before play begins. In contrast, in-game prediction attempts to use the progress during a game to determine win probabilities that vary as a function of in-game variables, for example, elapsed game time, or score difference. The focus of this paper is in-game prediction, and more specifically, during the course of the game estimating the probability the home team wins, or the "in-game win probability."

Estimating in-game win probability has long been a problem of interest. Many different methods for accurately estimating in-game win probability are found in the literature. Cooper et al. (1992) collected and analyzed data from 200 basketball games, 100 baseball games, 100 hockey games, and 100 football games to investigate when, during the course of a game for each of the sports, it is most likely to know the final outcome of the game. With respect to basketball they concluded, "late game leaders in basketball go on to win about four in five times," and that percentage is

"no different in football or hockey." They also found that "home teams in basketball were more than three times as likely as visiting teams to make a fourth-quarter comeback," and that the home winning percentage for basketball (64.1%) was the strongest evidence of a home-team advantage across major American sports.

One of the earlier papers on predicting in-game win probability for baseball is attributed to Lindsey (1963) who used the maximum likelihood estimator to determine the probability the home team wins given the inning the game is in and the home team's lead. He looked into how many times a team had won in prior games when in the same position as the current game was in. That information was then used to determine that team's empirical winning percentage. Up until his work, baseball decisions were based on what would maximize the scoring output for an inning. Lindsey's groundbreaking research instead focused on determining how each decision would affect the probability that a team wins instead of only the change in expected score. A more recent development in estimating in-game win probability is Lock and Nettleton (2014) who use random forests that combine pre-play variables to estimate win probability before each play of an NFL game. More detail the methodology by Lock and Nettleton can be found in Section 1.3.

In-game basketball analytics began to surface when Westfall (1990) developed a graphical summary of the scoring activity for a basketball game that is a real-time plot of the score difference versus the elapsed time. The features of the graph provided easy access to largest leads, lead changes, come-from-behind activity, and other interesting game features. Over time, models for forecasting in-game win probability have become more complex. Some are built on expert predictions, some on betting paradigms, and others on within-game metrics.

Shirley (2007) modeled a basketball game using a Markov model with three states and used that model for estimating in-game win probability. He defines the three states of the Markov model by considering which team has possession (having

two values), how that team gained the possession (with five values), and the number of points scored on the previous possession (four values), with a total of 40 states. Shirley's goals were to provide a detailed "microsimulation" of a basketball game. Rows of the transition matrix were modeled using multinomial logit models. He then incorporates effects for the transition probabilities, resulting in a baseline transition matrix, and then a unique transition matrix for every match-up between two teams.

Štrumbelj and Vračar (2012) improved upon the model of Shirley by taking into consideration the strengths of the two teams and estimated transition probabilities using performance statistics. They evaluated this approach along with logit regression, a latent strength rating method, and bookmaker odds. They found that the Markov model approach is appropriate for modeling a basketball game and produces forecasts of a quality comparable to that of other statistical approaches, while giving more insight into basketball. Vračar et al. (2016) extended the state description to capture other facets beyond in-game states so that the transition probabilities become conditional on a broader game context. Apart from the in-game event label, the extended state description also includes game time, the points difference, and the opposing teams' characteristics. They argue that by doing so, the model's transition probabilities become conditional on a broader game context (and not solely on the current in-game event), which brings several advantages: it provides a means to infer the teams' specific behavior in relation to their characteristics, and to mitigate the intrinsic non-homogeneity of the progression of a basketball game (which is especially evident near the end of the game).

Bashuk (2012) proposed using cumulative win probabilities over the duration of a game to measure team performance. Using five years of game play, he generated a Win Probability Index for NCAA basketball. He created the index using the maximum likelihood estimate (MLE) of $\frac{\text{home wins}}{\text{total games}}$ for each combination of minutes

left and score differential in the data. Using that index he created an open system to measure the impact, in terms of win probability added, of each play.

More recently, Benz (2019) developed a logistic regression approach. In this model, the coefficients of the covariates are allowed to change a function of time, i.e., the effects of the coefficients are dynamic in nature. More on the methodology of Benz can be found in Section 2.3.2. Chen and Fan (2018) developed a method for modeling point differences using a functional data analysis (FDA) approach. They argue that there are two major advantages of modeling the latent score difference intensity process using FDA. First, it allows for arbitrary dependent structure among score change increments. This removes potential model mis-specifications and accommodates momentum which is often observed in sports games. Second, further statistical inferences using FDA estimates will not suffer from inconsistency due to the issue of having a continuous model yet discretely sampled data.

Shi and Song (2019) develop a discrete-time, finite-state Markov model for the progress of basketball scores and use it to conditionally predict the probability the home team wins or loses by a certain amount. They find that an empirical study shows that the proposed model performs well, and more profoundly it can have positive return when they bet with the market. Song and Shi (2020) present an in-play prediction model based on the gamma process. The model is team-specific; it takes account of the relative strengths of the two teams playing in a match. They apply a Bayesian dynamic forecasting procedure that can be used to predict the final score and total points. Finally, Song et al. (2020) modify the gamma process by employing betting lines, letting the expectation of the final points total equal the pregame betting line. They find their model can produce a positive return on the over-under betting market, and their model has an application in monitoring the betting market.

While the techniques outlined above have various pros and cons, the methods for estimating the in-game win probability proposed here are aimed at improving the approaches of Stern (1994), Deshpande and Jensen (2016), and Benz (2019). Stern modeled the difference between the home and visiting teams' scores as a Brownian motion process with drift equal to the points in favor of the home team. This model, which is equivalent to a probit regression model, results in a relationship between the home team's lead and the probability of victory for the home team. His approach is one of the earliest to provide a mechanism for allowing time to be continuous, and to use that continuity in modeling the probabilities. Deshpande and Jensen extended the work of Stern by applying a Bayesian framework to the probit model. Benz extended Stern by allowing for multiple covariates with dynamic coefficients.

1.2 NBA Literature Review

Sports analytics has become a well-established area of research. Work spans more than 30 years and a large range of difficulty. Since the early 2000s, research in statistical methods for sports analytics has risen dramatically. The review articles of Kubatko et al. (2007), Santos-Fernandez et al. (2019), and Terner and Franks (2021) provide a fairly comprehensive review for sports analytics for a wide variety of sports, including basketball. One problem of interest is predicting the probability that the home team wins during the course of the game, or predicting "in-game win probability."

Speaking broadly, models for predicting the outcome of a sporting event can be classified into two systems: (1) pregame prediction or (2) in-game, or in-play, prediction. Pregame prediction involves determining the outcome of a game before play begins. Once play begins, the process of predicting the outcome ends. In contrast, in-game prediction attempts to use the progress during a game to determine

win probabilities that vary as a function of in-game variables, for example, elapsed game time or score difference. The focus of this paper is in-game prediction.

For a variety of sports, there are many different methods for accurately estimating in-game win probability found in the literature. For basketball, one of the first attempts to estimate in-game basketball analytics is Westfall (1990) who developed a graphical summary of the scoring activity for a basketball game that is a real-time plot of the score difference versus the elapsed time. The features of the graph provided easy access to largest leads, lead changes, come-from-behind activity, and other interesting game features. As computing technology and algorithms have become more sophisticated, models for forecasting in-game win probability have become more complex. Some are built on expert predictions, some on betting paradigms, and others on within-game metrics. Shirley (2007) modeled a basketball game using a Markov model with three states and used that model for estimating in-game win probability. Štrumbelj and Vračar (2012) improved upon that by taking into consideration the strengths of the two teams and estimated transition probabilities using performance statistics. Vračar et al. (2016) extended the state description to capture other facets beyond in-game states so that the transition probabilities become conditional on a broader game context. Bashuk (2012) proposed using cumulative win probabilities over the duration of a game to measure team performance. Using five years of game play, he generated a win probability index for NCAA basketball. Using the index, he created an open system to measure the impact, in terms of win probability added, of each play. More recently, Benz (2019) developed a logistic regression approach where the coefficients of the covariates are allowed to change a function of time, i.e., the effects of the coefficients are dynamic in nature. Chen and Fan (2018) developed a method for modeling point differences using a functional data approach. Shi and Song (2019) develop a discrete-time, finite-state Markov model for the progress of basketball scores, and

use it to conditionally predict the probability the home team wins or loses by a certain amount. Song and Shi (2020) present an in-play prediction model based on the gamma process. They apply a Bayesian dynamic forecasting procedure that can be used to predict the final score and total points. Song et al. (2020) modify the gamma process by employing betting lines, letting the expectation of the final points total equal the pregame betting line. Maddox et al. (2022a) develop three Bayesian approaches with dynamic priors. The adjusted model with a dynamic prior is, overall, the better of their three proposed methods. In this paper, we adopt the approach of Maddox et al. for predicting in-game win probabilities for games in the National Basketball Association (NBA). Additional considerations are made to their methodology upon the extension to the NBA, such as an refinement of the prior, improvement of the binning method and a more sophisticated adjustment from pregame information into the model for estimating in-game win probabilities. In addition, the win probabilities that ESPN publishes on their website have been collected to compare to the methodology proposed by Maddox et al..

1.3 CFB Literature Review

Sports analytics has become a well-established area of research. Work spans more than 30 years and a large range of difficulty. Since the early 2000s, research in statistical methods for sports analytics has risen dramatically. The review articles of Kubatko et al. (2007), Santos-Fernandez et al. (2019), and Terner and Franks (2021) provide a fairly comprehensive review for sports analytics for a wide variety of sports, including football. One problem of interest is predicting the probability that the home team wins during the course of the game, or predicting "in-game win probability."

Speaking broadly, models for predicting the outcome of a sporting event can be classified into two systems: (1) pregame prediction or (2) in-game, or in-play,

prediction. Pregame prediction involves determining the outcome of a game before play begins. Once play begins, the process of predicting the outcome ends. In contrast, in-game prediction attempts to use the progress during a game to determine win probabilities that vary as a function of in-game variables, for example, elapsed game time or score difference. The focus of this paper is in-game prediction for college football.

One of the earlier papers on predicting in-game win probability for baseball is attributed to Lindsey (1963) who used the maximum likelihood estimator to determine the probability the home team wins given the inning the game is in and the home team's lead. He looked into how many times a team had won in prior games when in the same position as the current game was in. That information was then used to determine that team's empirical winning percentage. Up until his work, baseball decisions were based on what would maximize the scoring output for an inning. Lindsey's groundbreaking research instead focused on determining how each decision would affect the probability that a team wins instead of only the change in expected score. A more recent development in estimating in-game win probability was approached by Benz (2019) who modeled college basketball in-game win probability by using a series of logistic regressions at different times throughout games on score differential and pregame win probability. He then smooths the multiple logistic regression models into a single smooth function. Maddox et al. (2022b) propose a Bayesian model for the National Basketball Association (NBA) based on time and score differential as the two predictors. Much of their methodology can be extended to college football, but different predictors must be considered. Score and time are not as solely informative about win probability in football as in basketball because many more factors play a large role, such as field position or down and distance. This paper instead uses separate models for expected possessions remaining

and expected score differential following the current and subsequent possessions as predictors themselves for the win probability model.

Within the game of football, Lock and Nettleton (2014) use random forests that combine pre-play variables of everything contributing to the state of the game to estimate win probability before each play of a National Football League (NFL) game. The model just builds a random forest on all the variables they have access to. For more detail on how random forests are built, see Section 4.4.1. Pro Football Reference (2012) create a quasi "black box" model, where they go into some detail about creating their win probability model using expected points along with pregame win probability and the known standard deviation of the end of the game score differential. However, many of the details used for their model are not mentioned.

Later, Ruscio and Brady (2021) compare the performance of the random forest model by Lock and Nettleton and the model put forth by Pro Football Reference when applied to the NFL. Their findings were that there was no discernable difference between the two models. They then modified the way the Pro Football Reference model uses game time to improve accuracy and to handle plays in overtime. Their modified model performs slightly better than the random forest model by Lock and Nettleton for plays throughout regulation and in overtime. Ruscio and Brady were able to obtain the Pro Football Reference model for their paper to reproduce the results. Pro Football Reference were also reached out to for this paper, but they declined to provide more detail on their model. Therefore, the Pro Football Reference model is unable to be reproduced and compared in this paper. However, as Ruscio and Brady found the Lock and Nettleton and Pro Football Reference performed similarly, it may be assumed that is also the case for college football.

CHAPTER TWO

Bayesian Estimation of In-Game Home Team Win Probability for Division-I College Basketball

2.1 Introduction

Sports analytics is not a new science. Work spans more than 30 years and a large range of difficulty. Since the early 2000s, research in statistical methods for sports analytics has risen dramatically. The review articles of Kubatko et al. (2007), Santos-Fernandez et al. (2019), and Terner and Franks (2021) provide a fairly comprehensive review for sports analytics for a wide variety of sports, including basketball.

Generally speaking, models for predicting the outcome of a sporting event can be classified into two systems: (1) pregame prediction or (2) in-game, or in-play, prediction. Pregame prediction involves determining the outcome of a game before play begins. In contrast, in-game prediction attempts to use the progress during a game to determine win probabilities that vary as a function of in-game variables, for example, elapsed game time, or score difference. The focus of this paper is in-game prediction, and more specifically, during the course of the game estimating the probability the home team wins, or the "in-game win probability."

Estimating in-game win probability has long been a problem of interest. Many different methods for accurately estimating in-game win probability are found in the literature. Cooper et al. (1992) collected and analyzed data from 200 basketball games, 100 baseball games, 100 hockey games, and 100 football games to investigate when, during the course of a game for each of the sports, it is most likely to know the final outcome of the game. With respect to basketball they concluded, "late game leaders in basketball go on to win about four in five times," and that percentage is

"no different in football or hockey." They also found that "home teams in basketball were more than three times as likely as visiting teams to make a fourth-quarter comeback," and that the home winning percentage for basketball (64.1%) was the strongest evidence of a home-team advantage across major American sports.

One of the earlier papers on predicting in-game win probability for baseball is attributed to Lindsey (1963) who used the maximum likelihood estimator to determine the probability the home team wins given the inning the game is in and the home team's lead. He looked into how many times a team had won in prior games when in the same position as the current game was in. That information was then used to determine that team's empirical winning percentage. Up until his work, baseball decisions were based on what would maximize the scoring output for an inning. Lindsey's groundbreaking research instead focused on determining how each decision would affect the probability that a team wins instead of only the change in expected score. A more recent development in estimating in-game win probability is Lock and Nettleton (2014) who use random forests that combine pre-play variables to estimate win probability before each play of an NFL game. Additionally, Ryall (2011) used play-by-play data with pregame Elo rankings to develop a model for Australian Rules football. The concept of using pregame power rankings is one that will be adopted here.

In-game basketball analytics began to surface when Westfall (1990) developed a graphical summary of the scoring activity for a basketball game that is a real-time plot of the score difference versus the elapsed time. The features of the graph provided easy access to largest leads, lead changes, come-from-behind activity, and other interesting game features. Over time, models for forecasting in-game win probability have become more complex. Some are built on expert predictions, some on betting paradigms, and others on within-game metrics. Shirley (2007) modeled a basketball game using a Markov model with three states and used that model for

estimating in-game win probability. Štrumbelj and Vračar (2012) improved upon the model of Shirley by taking into consideration the strengths of the two teams and estimated transition probabilities using performance statistics. Vračar et al. (2016) extended the state description to capture other facets beyond in-game states so that the transition probabilities become conditional on a broader game context. Bashuk (2012) proposed using cumulative win probabilities over the duration of a game to measure team performance. Using five years of game play, he generated a Win Probability Index for NCAA basketball. Using that he created an open system to measure the impact, in terms of win probability added, of each play.

More recently, Benz (2019) developed a logistic regression approach. In this model, the coefficients of the covariates are allowed to change a function of time, i.e., the effects of the coefficients are dynamic in nature. Chen and Fan (2018) developed a method for modeling point differences using a functional data approach. Shi and Song (2019) develop a discrete-time, finite-state Markov model for the progress of basketball scores, and use it to conditionally predict the probability the home team wins or loses by a certain amount. Song and Shi (2020) present an in-play prediction model based on the gamma process. They apply a Bayesian dynamic forecasting procedure that can be used to predict the final score and total points. Finally, Song et al. (2020) modify the gamma process by employing betting lines, letting the expectation of the final points total equal the pregame betting line.

While the techniques outlined above have various pros and cons, the methods for estimating the in-game win probability proposed here are aimed at improving the approaches of Stern (1994), Deshpande and Jensen (2016), and Benz (2019). Stern modeled the difference between the home and visiting teams' scores as a Brownian motion process with drift equal to the points in favor of the home team. This model, which is equivalent to a probit regression model, results in a relationship between the home team's lead and the probability of victory for the home team. His approach

is one of the earliest to provide a mechanism for allowing time to be continuous, and to use that continuity in modeling the probabilities. Deshpande and Jensen extended the work of Stern by applying a Bayesian framework to the probit model. Benz extended Stern by allowing for multiple covariates with dynamic coefficients.

The remainder of the chapter is organized as follows. Section 2.2 contains a more thorough overview of Stern (1994) and Deshpande and Jensen (2016), two methods for in-game probability designed specifically for NBA games, and Benz (2019), which was designed for NCAA games. In Sections 2.2.1 and 2.2.2, a modified enhanced Bayesian approach is proposed that not only improves in-game predictions compared to existing methods, but also is suitable for application to NCAA basketball. Section 2.3 presents a description of the data used in the study. Following that description is the result of estimating the in-game home team win probability to over 30,000 NCAA basketball games, and then using the estimated model to predict the outcome of over 10,000 NCAA games. To illustrate utility in Section 2.4, the models are applied to the 2016 Division 1 NCAA Tournament Championship game between the University of North Carolina and Villanova University. Finally, Section 2.5 contains a summary and concluding remarks.

2.2 Estimating In-Game Win Probability

For a specific game, consider the random process that is the home team's lead at time $t = 0, 1, \ldots, 2399$, where t is the game time elapsed in seconds. At a specific time t and for a specific home team lead ℓ, let $p_{t,\ell}$ denote the in-game probability that the home team will win the game at the end of regulation. At the beginning of the game, $t = \ell = 0$, the estimator of $p_{0,0}$ is dependent on the method used for estimating in-game win probability.

When considering multiple games $i = 1, 2, \ldots, M$, let $Y_i = 1$ if the home team wins game i and 0 otherwise. Consider $p_{t,\ell}$ as a continuous function of t and ℓ.

A classic approach to estimating $p_{t,\ell}$ in any (t, ℓ) cell is the maximum likelihood estimator. Specifically, define $N_{t,\ell}$ as the number of games in which the home team leads by ℓ points after t seconds, and define $n_{t,\ell} = \sum_{i=1}^{N_{t,\ell}} Y_i$, the number of games in the (t, ℓ) cell that the home team wins in regulation. Then on each (t, ℓ) cell, the random variable $n_{t,\ell}$ has a binomial$(N_{t,\ell}, p_{t,\ell})$ distribution. Within each cell, the maximum likelihood estimator of $p_{t,\ell}$ is $\bar{p}_{t,\ell} = n_{t,\ell}/N_{t,\ell}$. At $(0, 0)$, $\bar{p}_{0,0} = n_{0,0}/N_{0,0}$.

Let X_t represent the home team lead after t seconds. Another approach to estimating $p_{t,\ell}$ is found in Stern (1994). Stern estimates in-game home team win probability via a Brownian motion process with drift μ points per second home team lead and finite variance σ^2; that is,

$$\tilde{p}_{t^*,\ell} = P(X_1 > 0 \,|\, X_{t^*} = \ell) = \Phi\left(\frac{\ell + (1 - t^*)\mu}{\sqrt{(1 - t^*)\sigma^2}}\right), \tag{2.1}$$

where $t^* \in [0, 1)$ represents re-scaled time t and $\Phi(\cdot)$ denotes the standard normal cumulative distribution function. Although X_t is discrete, the model in equation (2.1) treats X_t as a continuous random variable. Stern suggests a continuity correction factor be applied ℓ, although he also noted the continuity correction factor results in little improvement in the model's performance. He provides empirical evidence that the Brownian motion model provides a good fit to the score differences when applied across multiple games. Stern noted the model (2.1) can be interpreted as a probit regression model relating the game outcome to the transformed variables $\ell/\sqrt{1 - t^*}$ and $\sqrt{1 - t^*}$, and with coefficients σ^{-1} and μ/σ. At $t^* = 0$ and $\ell = 0$, $\tilde{p}_{0,0} = \Phi(\mu/\sigma)$, signifying that μ/σ indicates the magnitude of the home field advantage. Specifying the home team advantage can be accomplished several ways. Stern (1994) suggests letting μ be the home field advantage in the particular sport, and for the NBA, $\mu = 5$ or 6 points. They also note that the home team wins in approximately 55% to 65% of games, and so σ can be chosen so that values of μ/σ are in the 0.12 to 0.39 range.

Taking the approach as outlined in Deshpande and Jensen (2016), the game is partitioned into cells based on time (in seconds) and point differential. Some (t, ℓ) cells may have small values of $N_{t,\ell}$, which have the potential to result in estimators of $p_{t,\ell}$ with very large standard errors. To address this issue, windows can be defined and centered on (t, ℓ) in such a way that the in-game win probability remains relatively constant across the window. In basketball, since no offensive possession can result in more than four points the window with respect to ℓ can be reasonably defined as $[\ell - 2, \ell + 2]$. Moreover, since most offensive possessions last at least six seconds the width of the time window is taken to be six. The same notation will be adopted for any $[t - 3, t + 3] \times [\ell - 2, \ell + 2]$ window; that is, $N_{t,\ell}$ is the number of games in the window in which the home team has led by any value in $[\ell - 2, \ell + 2]$ points after any time in $[t - 3, t + 3]$ seconds and $n_{t,\ell} = \sum_{i=1}^{N_{t,\ell}} Y_i$, distributed as a binomial$(N_{t,\ell}, p_{t,\ell})$ random variable.

Benz (2019) presents a logistic regression model for estimating $p_{t,\ell}$. The model improves upon his win probability model built into his `ncaahoopR` package. For a specific game, Benz considers time intervals of the form $(t' - \Delta, t']$, where $t' = 2400 - t$ is the time remaining in the game. Pregame spread and score differential are included as covariates. Each covariate has a coefficient that is allowed to change as function of time remaining. The time intervals overlap each other by 90%. The value of Δ is adjusted as time remaining approaches zero. Using a training set of 10,949 games from the 2016-2017 and 2017-2018 seasons, Benz finds empirical values for the model coefficients. He illustrates that, as the amount of time remaining in a game decreases, the importance of of pregame spread decreases in a nonlinear fashion, and the importance of score differential increases, also in a nonlinear fashion. An investigation of the scale of the coefficients reveals that score differential becomes the more important predictor after the half.

Bayesian methods provide another approach to estimating $p_{t,\ell}$. It is well-known that the beta family is a conjugate prior for estimating $p_{t,\ell}$. Let $\alpha_{t,\ell} > 0$ and $\beta_{t,\ell} > 0$ be the shape parameters of the beta prior on $p_{t,\ell}$. Then for each window within the (t, ℓ) plane, the Bayes estimator of $p_{t,\ell}$ is the mean of the posterior; specifically

$$\hat{p}_{t,\ell} = \frac{n_{t,\ell} + \alpha_{t,\ell}}{N_{t,\ell} + \alpha_{t,\ell} + \beta_{t,\ell}}. \tag{2.2}$$

For all $t = 0, 1, \ldots, M$, on all windows centered at (t, ℓ), Deshpande and Jensen (2016) propose the beta prior

$$p_{t,\ell} \sim \begin{cases} \text{beta}(0, 10), & \text{for } \ell < -20, \\ \text{beta}(5, 5), & \text{for } -20 \leq \ell \leq 20, \\ \text{beta}(10, 0), & \text{for } \ell > 20. \end{cases} \tag{2.3}$$

The choice of prior depends only on the home team lead ℓ, and does not take into account the time remaining in the game t.

2.2.1 Dynamic Prior for In-Game Home Team Win Probability

The scale parameters in (2.3) rely only on the home team lead ℓ, not taking into account the interaction between ℓ and time remaining. Also notable is that for $|\ell| > 20$, which occurs 8.5% of the time, the prior parameters yields an improper prior. In these cases, the prior may overwhelm the information in the data. To model the interaction between the home team lead and the time elapsed, we propose choosing the scale parameters for the beta prior dynamically, as illustrated in Figure 2.1 and specified in Table 2.1. The choices of the time scales and prior parameters were determined based upon the lead author's six years of experience working as an analyst for a major Division 1 NCAA Basketball men's team. From a basketball coaching perspective, there are not many times in the first half of a game that strategies or game plans change, so the prior for the first half of a game remained relatively non-informative and constant across time. During the second half, coaches often have a goal of winning the first ten minutes, minutes 10–15, 15–19, and the

last minute, which is where the changes in the prior along time take place. At each time point, the change in prior based on score differential were shrunk down closer to a score differential of zero because small differences in score differential indicate more as the game moves further on.

Table 2.1: Specification of dynamic beta prior.

Color	Prior
Red	beta(19,1)
Orange	beta(9,1)
Yellow	beta(4,1)
White	beta(1,1)
Green	beta(1,4)
Light blue	beta(1,9)
Blue	beta(1,19)

The time intervals illustrated in Figure 2.1 are: $[0, 1200)$, $[1200, 1800)$, $[1800, 2100)$, $[2100, 2340)$, and $[2340, 2400]$. Observed home-team differential intervals are time dependent. For the small number of games when the $|\ell| > 50$, the beta(19, 1) or beta(1, 19) prior is applied accordingly. In the rare instance that a specific (t, ℓ) window had no observed data, for a positive differential, the largest posterior probability of a home team win is used; for a negative differential, the smallest posterior probability of a home team win is used. This choice of posterior estimates is justified since windows with empty cells are those windows with large score differentials relative to the elapsed time. Moreover, that the combination hasn't occurred in over 30,000 games. Therefore, the choice of the largest or small-

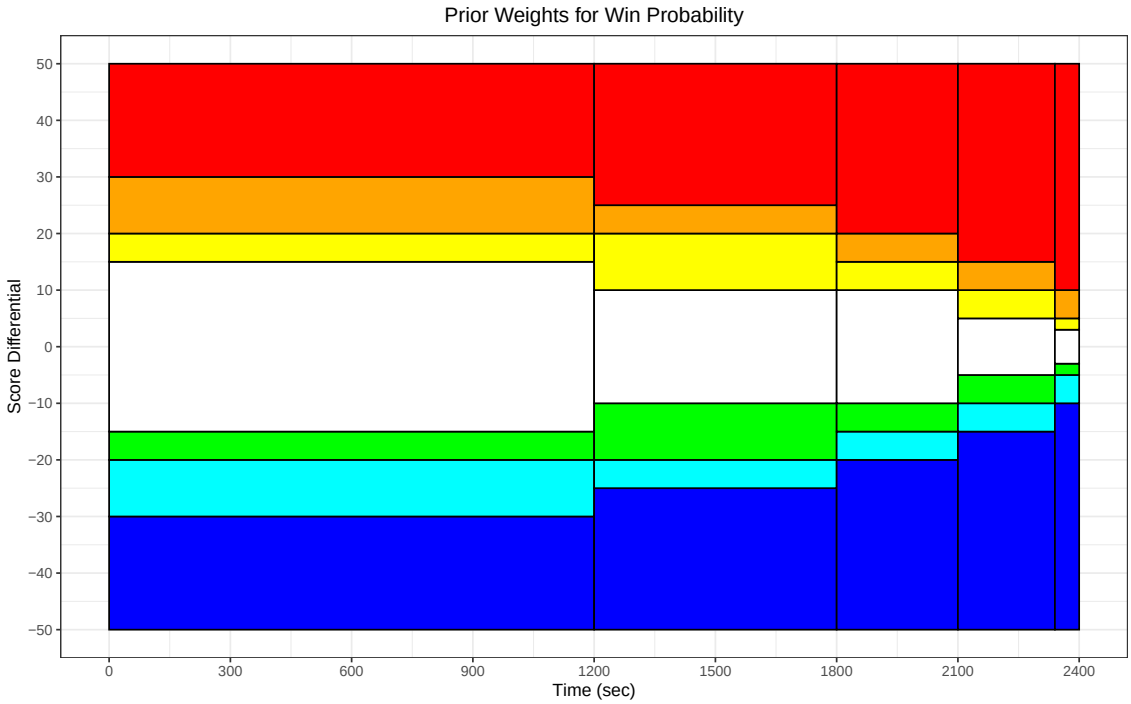

Figure 2.1: Illustration of dynamic beta prior.

est estimate is reasonable. Using the dynamic prior, for any (t, ℓ) in the plane, the dynamic Bayesian estimator of $p_{t,\ell}$ is given as in equation (2.2).

The newly proposed choice of prior is proper for all (t, ℓ) combinations and is much less likely to overwhelm the information in the data. More importantly, the dynamic prior models the in-game home team win probability as a combination of (t, ℓ). As with the prior in (2.3), the size of the score differential $|\ell|$ has an affect on the estimated in-game win probability. However, the newly proposed dynamic prior also models the estimated in-game win probability as a function of the game time elapsed. In particular, early in the game (for small t), even for a relatively large score differential, the dynamic prior allows for a larger probability of a comeback than late in the game, for the same differential.

2.2.2 Adjustment of Bayesian Estimator

For any game, the skill of the two teams playing has a definitive impact on how likely it is a team will win. If two teams are evenly matched, then accurately predicting which team will win is more difficult than if one team is much more skilled than the other. Accordingly, team skill should also be taken into account when estimating the in-game win probability. None of the previously described estimators incorporate such a measure. In the following, we propose an adjusted Bayesian estimator for in-game win probability, where the adjustment incorporates the skills of the teams in the game.

To determine the adjustment, or the pregame point spread, the team ratings and home team advantage are used. The point differential is commonly modeled using a normal distribution with a standard deviation of 10 points; see for example, Adams (2019). The mean for each game is computed using the difference in the two teams' ratings with a 3.5 point advantage added for the home team. The value 3.5 commonly used as an adjustment that shows the worth of the home court in evenly matched games. It differs from the parameter μ in (2.1), which measures home court advantage in conference and non-conference games. Since in non-conference games the home team tends to be better, the value of μ will be greater than the 3.5 point advantage. Using the normal quantile function, the pregame point spread can be converted to a pregame winning probability, $\hat{p}_p$ say, for each team. For the rare case when a non Division 1 team appeared in the database, that team is given a rating slightly lower than the lowest ranked Division 1 team. While this may not be accurate for each of these occurrences, no power rankings on Division 2 and lower teams are available. During the game at each time t, the probability the home team wins $\hat{p}^*_{t,\ell}$ is the weighted average of the pregame win probability and the current unadjusted in-game probability, $\hat{p}_{t,\ell}$ found using the dynamic Bayesian model described

in Section 2.2.1; that is,

$$\hat{p}^*_{t,\ell} = \left(\frac{2400 - t}{2400}\right)\hat{p}_p + \left(\frac{t}{2400}\right)\hat{p}_{t,\ell} \tag{2.4}$$

The adjustment has the following effect. In a tight contest, the in-game win probability is shifted such that the higher ranked team is predicted to be more likely to win. When the higher ranked team is winning, the probability the higher ranked time will win is larger than a lower ranked team that is winning at the same time by the same margin. The choice of a linear weighting function for the influence of the pregame win probability $\hat{p}_p$ on $\hat{p}^*_{t,\ell}$ was influenced by its interpretability, simplicity, and effectiveness of a linear model. Further work on the effectiveness of the improvements that might be obtained by more complicated weighting models.

2.3 Comparison of Methods for Estimation and Prediction

In the following, the performances of the MLE (baseline), the probit model estimator of Stern (1994), the Bayesian estimator with prior given in (2.3) from Deshpande and Jensen (2016), and the proposed Bayesian estimator with dynamic prior in Table 2.1 and Figure 2.1 are compared from both the modeling and prediction perspectives. The games that are used in conducting the study are from the 2012-2013 seasons up until the 2019-2020 season, which was cut short due to COVID. In Subsection 2.3.1, the data is more thoroughly described, and the data collection process and its challenges are discussed. In Section 2.3.2 the models are used to compute estimates of in-game home team win probability using 30,789 games beginning with the 2012-2013 season and ending with the 2017-2018 season. In Section 2.3.3, the model estimates are used to predict the outcome of the 10,853 games from the 2018-2019 season and the 2019-2020 season. The performances of the models are compared through Brier Score and misclassification rates.

2.3.1 NCAA Data, Collection and Challenges

Play-by-play data from ESPN was scraped using *R* and the package `rvest`. Since play-by-play data is not readily available on ESPN's site for college basketball games prior to 2012, the data collected begins with the 2012-2013 season and continues through the completion of the 2019-2020 season.

Due to inconsistencies in data collection and formatting, the data used does not contain play-by-play data for every game for all teams. From each conference, a randomly selected team was chosen for inclusion in Tables 2.2 and 2.3, which shows the percentage of all games played for each season in the data. In general, smaller schools have a lower percentage of games with available data than larger schools. However, the lack of information did not appear systematic, and is not seen as problematic, since the data does include play-by-play information for many other similar teams.

Table 2.2: Percentage of games played for which data was collected by ESPN for selected programs from each conference (Part I).

Team	Conference	12-13	13-14	14-15	15-16	16-17
Cent Arkansas	ASUN	76.7%	93.1%	79.3%	100.0%	100.0%
UMass Lowell	America East	NA	89.3%	89.7%	96.6%	93.5%
UCF	American	90.0%	93.5%	93.3%	96.7%	97.2%
George Mason	Atlantic 10	78.9%	87.1%	87.1%	100.0%	88.2%
NC State	ACC	17.1%	94.4%	97.2%	97.0%	100.0%
Texas Tech	Big 12	19.4%	90.6%	96.9%	93.8%	96.9%
Xavier	Big East	67.7%	91.2%	83.8%	100.0%	89.5%
N Arizona	Big Sky	87.5%	90.6%	92.1%	86.7%	100.0%
Gardner-Webb	Big South	85.3%	90.9%	94.3%	87.9%	93.9%
Ohio State	Big Ten	10.8%	85.7%	85.7%	100.0%	100.0%

Team	Conference	12-13	13-14	14-15	15-16	16-17
UC Irvine	Big West	75.7%	91.4%	85.3%	86.8%	91.7%
Delaware	Colonial	60.6%	94.3%	90.0%	86.7%	100.0%
UNC Charlotte	Conference USA	75.0%	90.3%	90.6%	93.9%	96.7%
Oakland	Horizon	81.8%	87.9%	90.9%	94.3%	97.1%
Harvard	Ivy	69.0%	90.6%	80.0%	90.0%	96.4%
Fairfield	MAAC	74.3%	87.5%	96.8%	97.0%	96.8%
Kent State	Mid-American	80.0%	78.1%	91.4%	93.8%	91.7%
Morgan State	MEAC	62.5%	41.9%	58.1%	41.9%	63.3%
Drake	Missouri Valley	78.1%	90.3%	96.8%	96.8%	100.0%
San Diego State	Mountain West	44.1%	91.7%	94.4%	92.1%	90.9%
St Francis (PA)	Northeast	89.7%	80.6%	100.0%	96.7%	97.1%
SE Missouri St	Ohio Valley	63.6%	53.1%	96.7%	93.1%	97.0%
Colorado	Pac-12	36.4%	94.3%	90.9%	94.1%	94.1%
Bucknell	Patriot League	79.4%	90.0%	88.2%	93.5%	100.0%
Missouri	SEC	32.4%	94.3%	96.9%	100.0%	93.8%
Furman	Southern	90.3%	83.3%	93.9%	94.3%	91.4%
McNeese State	Southland	54.8%	87.1%	87.1%	93.1%	100.0%
Jackson State	SWAC	41.4%	32.3%	43.8%	41.7%	34.4%
Omaha	Summit League	80.6%	46.9%	69.0%	90.6%	96.9%
South Alabama	Sun Belt	83.3%	90.3%	97.0%	87.9%	96.9%
BYU	West Coast	58.3%	94.3%	94.3%	97.3%	94.1%
Chicago State	WAC	75.8%	90.6%	62.5%	87.5%	100.0%

Table 2.3: Percentage of games played for which data was collected by ESPN for selected programs from each conference (Part II).

Team	Conference	17-18	18-19	19-20	Total
Cent Arkansas	ASUN	91.4%	93.9%	90.3%	90.6%
UMass Lowell	America East	93.3%	96.9%	90.6%	92.8%
UCF	American	96.9%	87.9%	96.7%	94.0%
George Mason	Atlantic 10	97.0%	93.9%	96.9%	91.1%
NC State	ACC	97.0%	100.0%	100.0%	87.8%
Texas Tech	Big 12	97.3%	97.4%	93.5%	85.7%
Xavier	Big East	97.1%	97.1%	96.9%	90.4%
N Arizona	Big Sky	93.8%	96.8%	86.7%	91.8%
Gardner-Webb	Big South	96.9%	85.7%	90.9%	90.7%
Ohio State	Big Ten	97.1%	97.1%	100.0%	84.6%
UC Irvine	Big West	91.4%	89.2%	100.0%	88.9%
Delaware	Colonial	100.0%	90.9%	90.9%	89.2%
UNC Charlotte	Conference USA	93.1%	93.1%	100.0%	91.6%
Oakland	Horizon	97.0%	100.0%	84.8%	91.7%
Harvard	Ivy	100.0%	93.5%	89.7%	88.7%
Fairfield	MAAC	93.9%	93.5%	93.8%	91.7%
Kent State	Mid-American	97.1%	100.0%	96.9%	91.1%
Morgan State	MEAC	81.3%	83.3%	71.0%	62.9%
Drake	Missouri Valley	97.1%	97.1%	97.1%	94.1%
San Diego State	Mountain West	93.9%	94.1%	100.0%	87.7%
St Francis (PA)	Northeast	96.8%	90.9%	84.4%	92.0%
SE Missouri St	Ohio Valley	93.5%	87.1%	100.0%	85.5%
Colorado	Pac-12	93.8%	97.2%	93.8%	86.8%

Team	Conference	17-18	18-19	19-20	Total
Bucknell	Patriot League	97.1%	97.0%	97.1%	92.8%
Missouri	SEC	97.0%	96.9%	100.0%	88.9%
Furman	Southern	93.9%	90.9%	100.0%	92.3%
McNeese State	Southland	96.4%	93.5%	90.6%	87.8%
Jackson State	SWAC	50.0%	40.6%	90.6%	46.8%
Omaha	Summit League	93.5%	90.6%	93.8%	82.7%
South Alabama	Sun Belt	96.9%	97.1%	96.8%	93.3%
BYU	West Coast	100.0%	90.6%	100.0%	91.1%
Chicago State	WAC	96.9%	84.4%	89.7%	85.9%

Part of the play-by-play data includes the arena where the game was played. Teams are assigned multiple home arenas in situations where that is appropriate due to design, i.e., Villanova, or because of a one off, for example, TCU playing their home games in a different arena for the 2014-2015 season due to renovations. While ESPN is not globally consistent in the way information is stored, this effect on home-arena determination is rare. It should also be noted that occasionally teams play in what could be considered a home game with respect to fan attendance and location. An example of this is when the University of Kansas plays in Kansas City. Despite these issues, the database of games is large enough that the overall accuracy of the data used in the analysis is unaffected.

Finally, to capture each team's power rating prior to each game, data was scraped from the website `teamrankings.com`.[1] This site was chosen over other systems that may outperform `teamrankings.com` because `teamrankings.com` has historical daily power ratings available and the other sites do not. An alternative to using these ratings might be the betting spread from Las Vegas betting lines.

2.3.2 Estimating Home Team In-Game Win Probability

Estimates of in-game home team win probability from the MLE, probit model, Bayes model using the prior in (2.3), and the dynamic Bayes model are given in Figures 2.2 through 2.5. For each $t = 0, 1, \ldots, 2399$ and $|\ell| = 0, 1, \ldots, 104$ cell, the estimates are illustrated by letting the color of any (t, ℓ) represent the estimated value of $p_{t,\ell}$. The values $|\ell| = 104$ are taken from a game played on Dec. 30, 2013, when Southern University beat Champion Baptist College by a score of 116-12, the largest margin of victory for any game included in the data described in Section 2.3.1. Historically, the largest margin of defeat in college men's basketball occurred on Jan. 12, 1992 when Troy State University, the home team, beat DeVry University of Atlanta by a score of 258-141. For any cell, blue represents an estimated in-game home team win probability of approximately zero, white an estimated probability of approximately 0.5, and red an estimated probability of approximately one. Examination of these four figures reveals some common features. In particular, when the score differential is large, in almost all (t, ℓ) cells, all models result in approximately equal estimated probabilities. Additionally, by the end of the game the estimated win probability goes to either zero or one for all four models.

Estimated probabilities using the traditional MLE and probit models are seen in Figures 2.2 and 2.3, respectively. The probit model of Stern (1994) was fit to the data, and values for μ and σ were estimated via maximum likelihood. The estimates

[1] `teamrankings.com` lacks data for December 3, 2012. For games played on that date, the team ratings from the previous day were used.

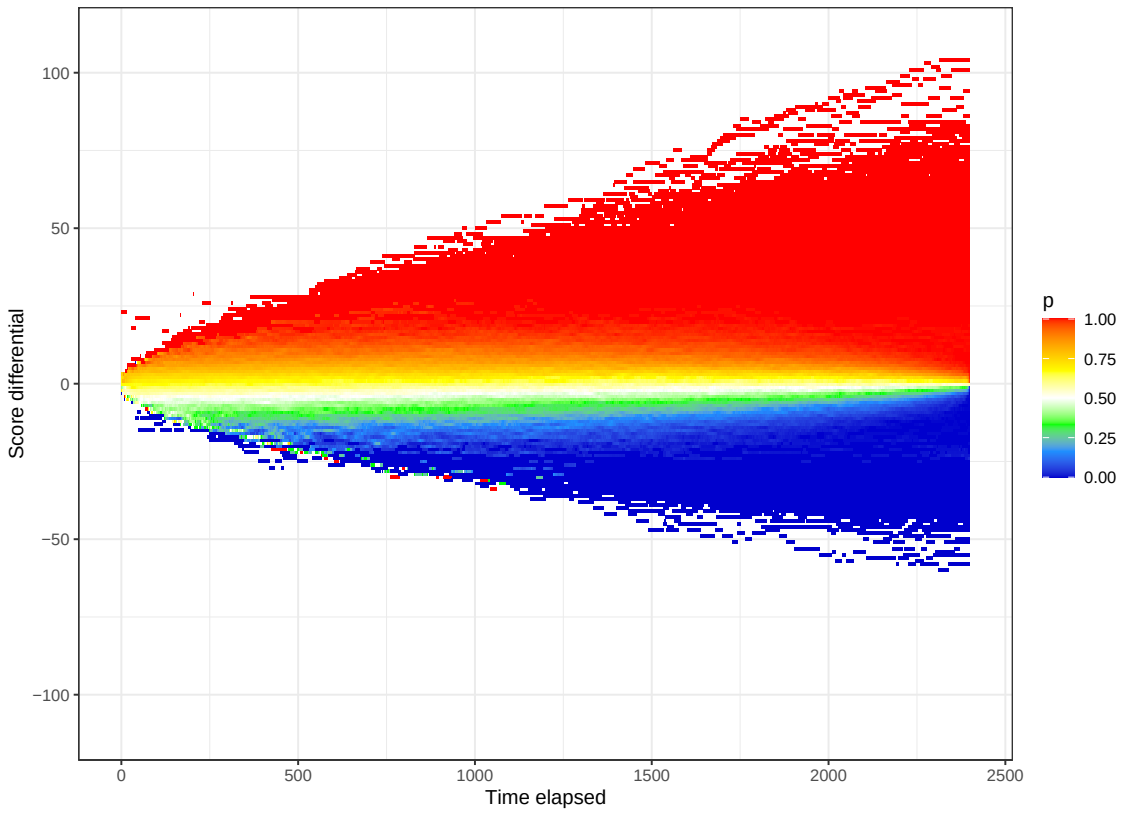

Figure 2.2: Maximum likelihood estimates of in-game home team win probability.

were $\hat{\mu} = 5.88$ and $\hat{\sigma} = 14.26$, giving $\hat{\mu}/\hat{\sigma} = 0.41$. While the estimate for μ does fall within the suggested five or six points, the ratio of the mean to the standard deviation is slightly larger than the upper bound 0.39. This should not be of concern, since in NCAA Division 1 basketball, the home team wins approximately 67% of the time, compared to only 55% - 65% for the NBA. The MLE and probit models are fitted using information in a cell, as opposed to a window, and so are more cells with missing data than the Bayesian estimates in Figures 2.4 and 2.5. Additionally, the MLE are less smooth than the other three. In Figure 2.2, up until $t = 1100$, a few windows on the lower bound of the estimates are red, indicating that even though the home team is down early in the game by a very large margin, the probability the home teams wins is approximately one. These cells illustrate one of the problems in

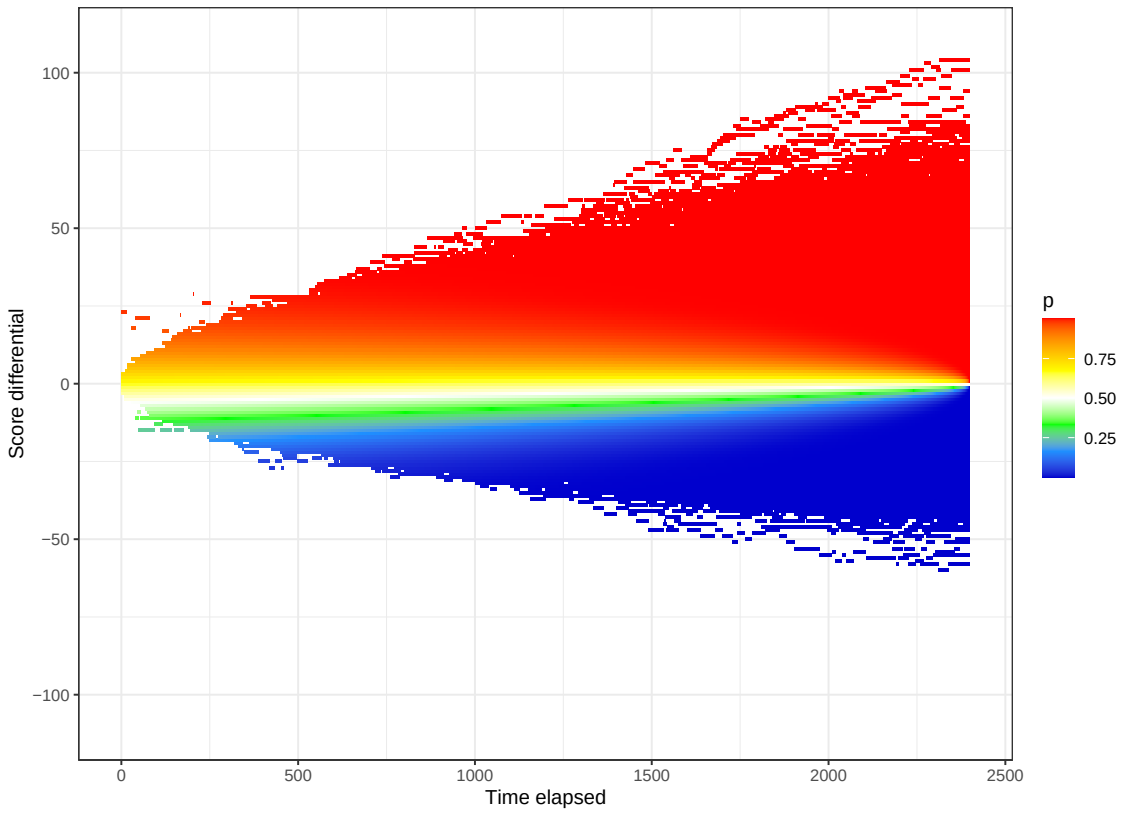

Figure 2.3: Probit model estimates of in-game home team win probability.

using the MLE. The estimates produced in these cells are based on a single game. In February of 2018, Drexel University came back from a 34-point deficit to win over the University of Delaware by a score of 85-83. This comeback is the "largest come-from-behind win in the history of Division I basketball" (ESPN/Associated Press, 2018). Delaware led by a score of 53-19 with 2:36 remaining in the first half. For a many of the cells associated in the first half of this game, both $N_{t,\ell}$ and $n_{t,\ell}$ are one, resulting in a $\bar{p}_{t,\ell} = 1$. For this specific game, on these specific cells, the MLE performs perfectly on this training data. However, as evidenced by other games where the home team fell behind early – just not as badly as Drexel – the home team lost with high relative frequency. In short, cells with low values of $N_{t,\ell}$ result in unreliable estimates because the estimators have large standard errors. Rare events are not as problematic when using the probit model since the drift μ

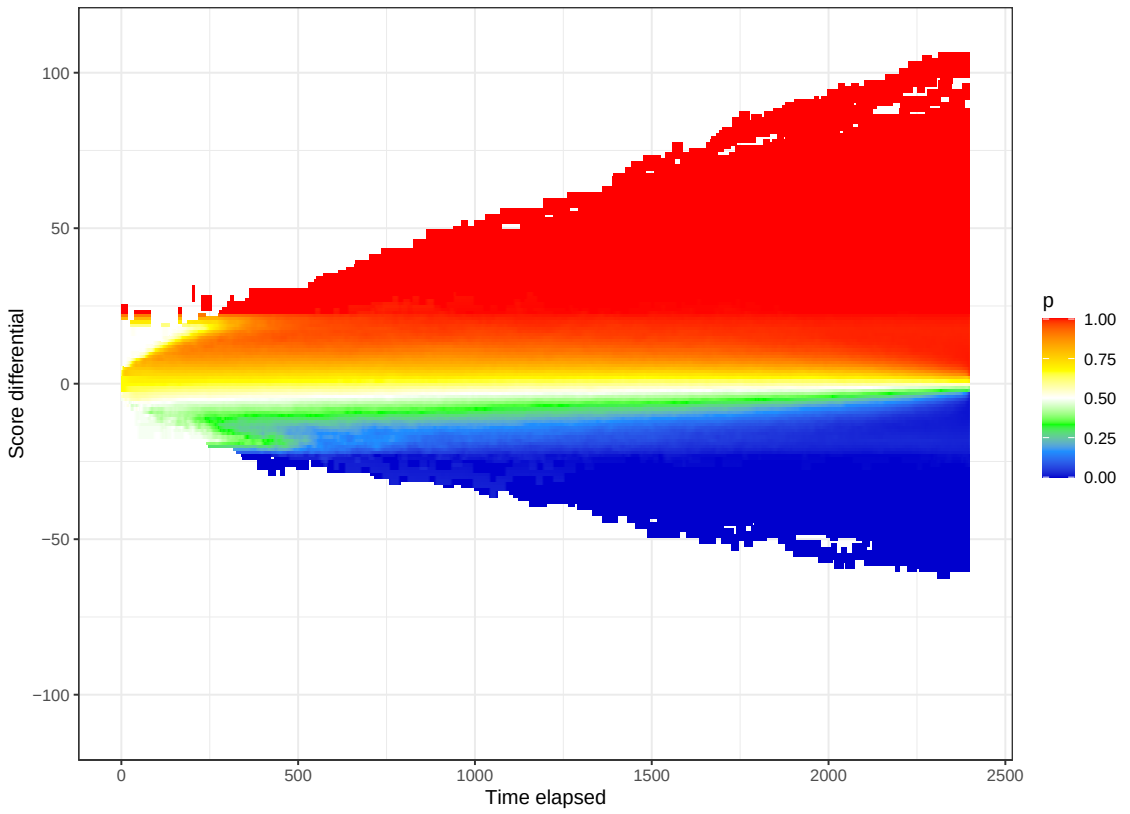

Figure 2.4: Bayesian estimates of in-game home team win probability.

and variance σ are included in the estimator for each cell. The Bayesian estimates perform better estimating probabilities for rare events for two reasons. The first is seen in a type of interpretation of the scale parameters in the choice of prior. The first scale parameter can be interpreted as the number of "pseudo-wins" in a cell, and the second scale parameter as the number of "pseudo-losses." In this way, the scale parameter choices can be seen as increasing the number of games in a specific window in a way that is intuitive for that window. Specifically if the home team is ahead, then first scale parameter being large effectively acts to increase the number of wins in that cell. On the other hand, if the home team is behind, the second scale parameter is large, and thus acting to increase the number of losses. The second reason rare events do not effect the Bayesian estimates as drastically as the MLE or as the probit model is due to the Bayesian estimates being computed over

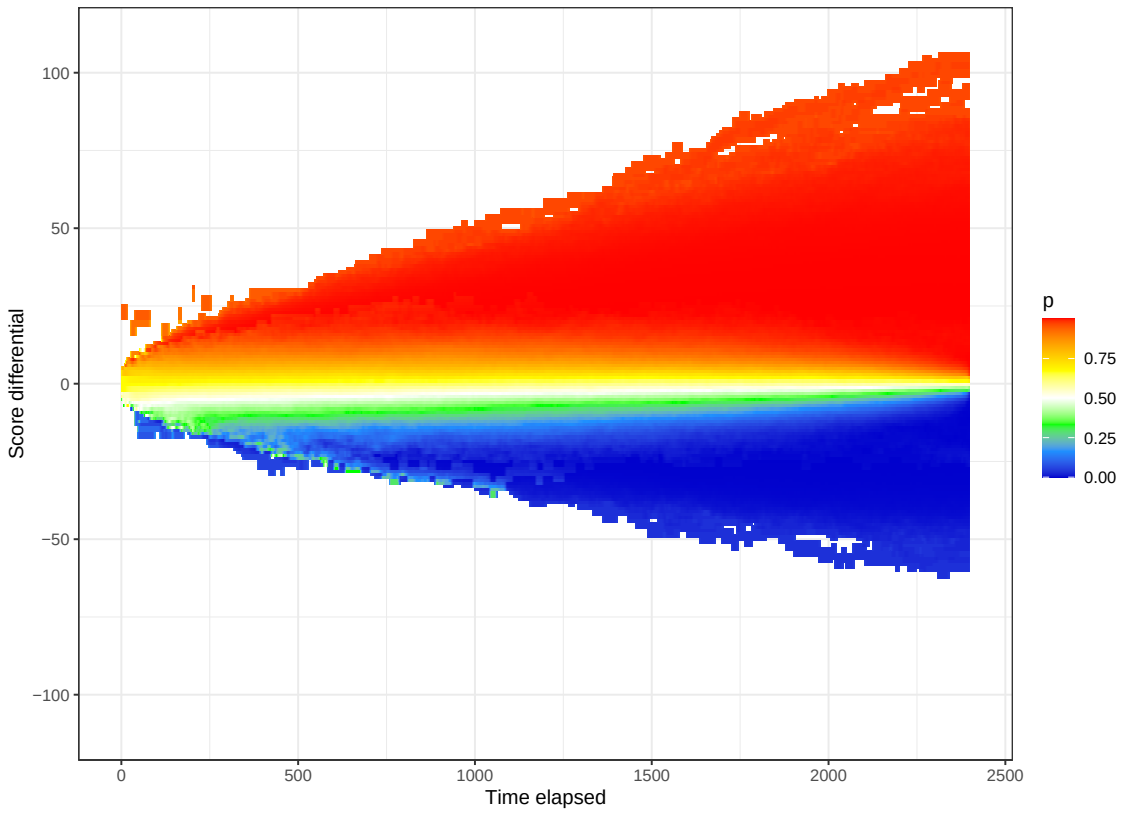

Figure 2.5: Dynamic Bayesian estimates of in-game home team win probability.

a small window, as opposed to on a cell. Estimates from the Bayes models, shown in Figures 2.4 and 2.5, are slower to increase the win probability early in the game compared to the MLE and probit model estimates. As previously mentioned, because the Bayes estimates are computed using information in a small window around the cell, their overall appearance is smoother. The most prominent difference in the two Bayesian estimates is for score differentials between -20 and 20 points. In that range, the Bayes estimates using the prior in (2.3) tends to result in moderate probabilities (0.4 to 0.6) for a early in the game, and for longer period of elapsed time than the Bayes estimates with the dynamic prior.

Figures 2.6 through 2.8 illustrate the estimated in-game home team win probabilities using the adjusted Bayes estimator with dynamic prior. Because the es-

timated probability is affected by the pregame home team win probability, three values of $\hat{p}_p$ were selected to illustrate the performance of the newly proposed estimator. Of the 30,789 games in the data, approximately 67% were won by the home team, motivating the choice of $\hat{p}_p = 0.67$. The other two choices of pregame home team win probability were found by adding and subtracting 0.3 from 0.67. For all of the 30,789 games, the adjusted Bayes estimator of equation (2.4) was applied. For a pregame win probability of $\hat{p}_p = 0.67$, the adjusted in-game home team win probabilities are see in Figure 2.6. Figures 2.7 are the estimates for games with a pregame win probability of 0.97, and Figure 2.8 for those with $\hat{p}_p = 0.37$. A visual comparison of the graphs in Figures 2.2 through 2.4 to Figures 2.6 through 2.8 leads to the conclusion that the unadjusted estimates perform very differently than the adjusted estimates. In particular, the adjusted estimates tend to change more slowly than the adjusted estimates. This is especially true in the first half of the game. The implications of this slower change is that the adjusted estimates are less likely to produce a high win probability early in the game for large leads. Later in the game, the adjusted estimates change more rapidly. An animated Shiny app showing the evolution of the adjusted dynamic Bayes estimates as a function of $\hat{p}_p$ can be found online at `jasontmaddox.shinyapps.io\Win_Probability`.

2.3.3 Assessing the Models Predictive Performance

The six models fitted in Section 2.3.2 were used to predict the outcome of the 10,853 games in 2018-2019 season and the 2019-2020 season. Two of the more common model evaluation methods used in classification of binary prediction models are Brier Score and misclassification rates. The predictive performances of the MLE, probit model, Bayes with prior in (2.3), Bayes with dynamic prior, and adjusted Bayes model with dynamic prior are compared below using the two measures.

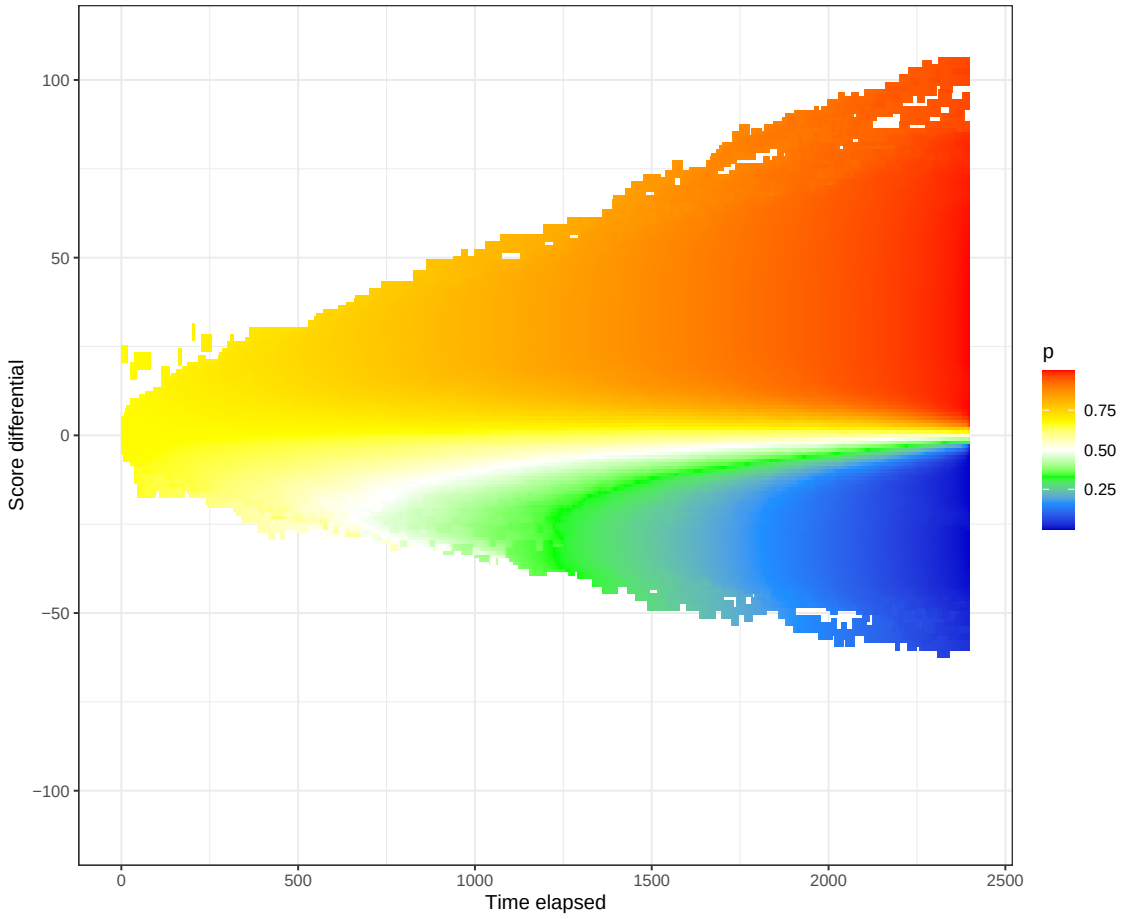

Figure 2.6: Adjusted dynamic Bayesian estimates for a pregame win probability of .67.

Brier Score is analogous to the sum of the squared errors in linear regression and has the advantage of maintaining the continuous information of all estimations. The square of the difference between the estimated probability is compared to the observed binary outcome. In the context of in-game home team win probabilities, this observed binary outcome is denoted y_i and is the observed value of Y_i as defined in Section 2.2. To interpret Brier Score, if $Y_i = 1$ for all i, and the predicted probability is also one for every i, then Brier Score will be zero, indicating perfect prediction. On the other hand, if for all $i, Y_i = 0$, and the estimated probability is one, then Brier Score will be one, the worst possible Brier Score.

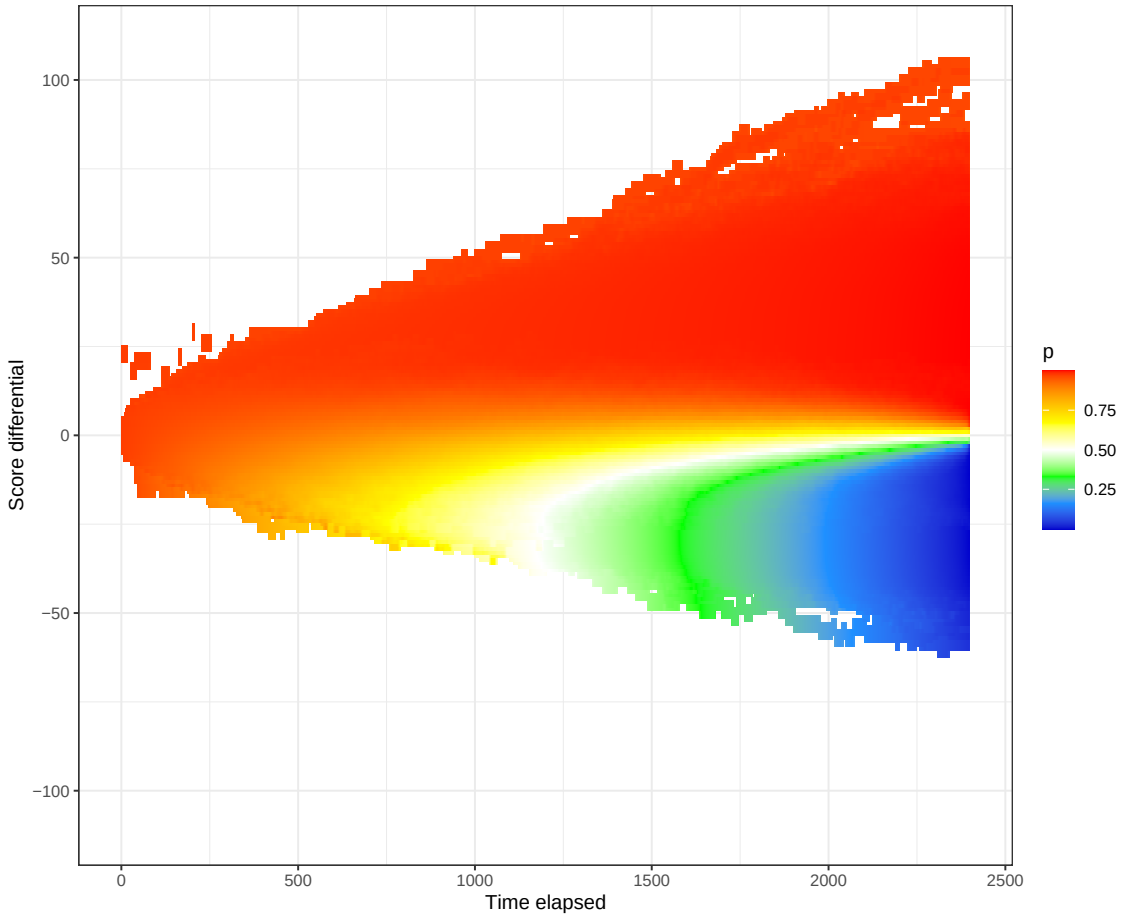

Figure 2.7: Adjusted dynamic Bayesian estimates for a pregame win probability of .97.

To compute Brier Score in this context, let $\tilde{p}_{t,\ell}$ represent the estimated in-game home team win probability for any one of the six methods. For each (t, ℓ) cell in the test data set consisting of the 10,853 games, let $N^*_{t,\ell}$ represent the number of games in the cell in which the home team led by ℓ points at time t; that is, $N^*_{t,\ell}$ is the number of games observed in that cell. Then for non-missing estimated probabilities, Brier Score is

$$B = \frac{1}{Q} \sum_{t=0}^{2399} \sum_{\ell=-104}^{104} \sum_{j=1}^{N^*_{t,l}} \left(\tilde{p}_{t,\ell} - y_j\right)^2,$$

where Q is the sum of $N^*_{t,\ell}$ in cells without missing $\tilde{p}_{t,\ell}$. Values of Q for the models are given in Table 2.4.

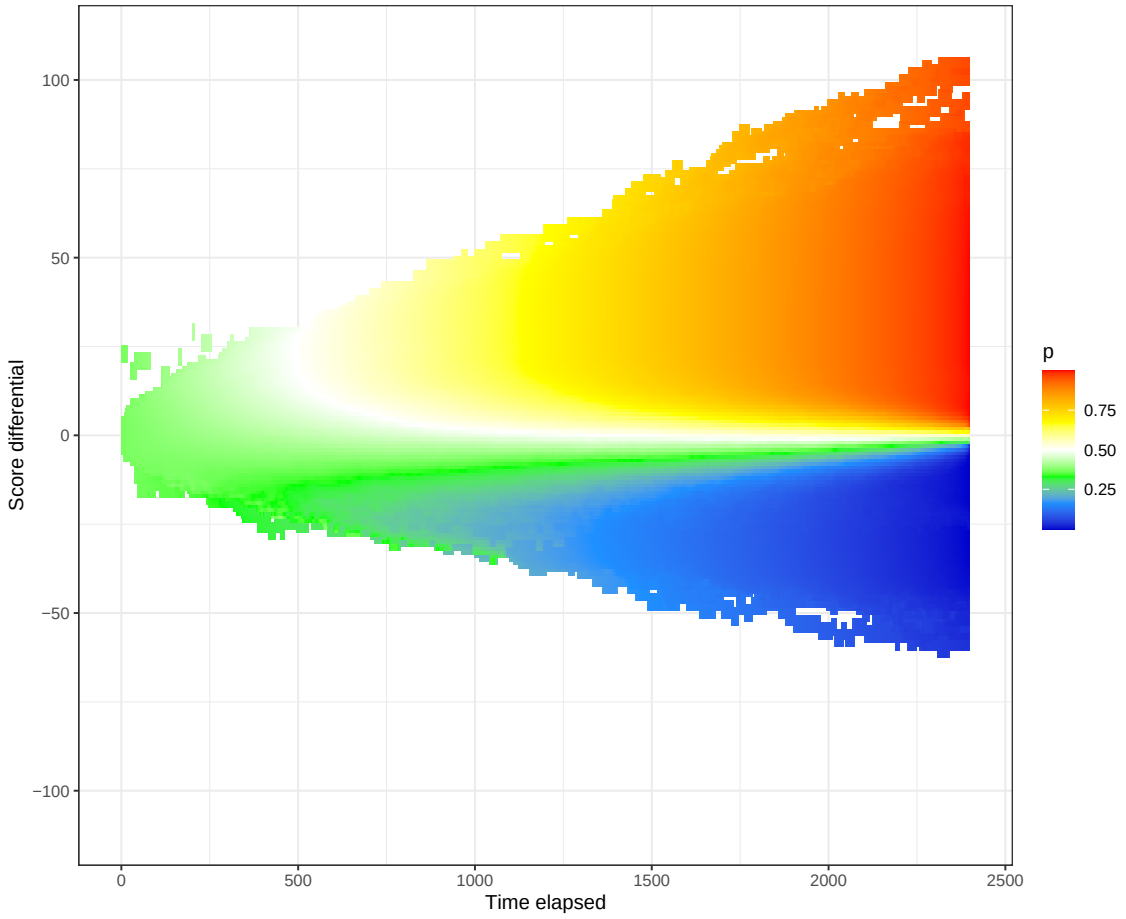

Figure 2.8: Adjusted dynamic Bayesian estimates for a pregame win probability of .37.

Table 2.4: Number of points in computing Brier Score and the misclassification rates for six estimation methods.

	Model				
Season	MLE	Probit	Bayes	(Adj) Dyn Bayes	Benz
2018-2019	13,118,309	13,118,309	13,122,430	13,123,200	3,553,584
2019-2020	12,916,158	12,916,158	12,922,503	12,924,000	3,439,214

Misclassification rates transform the estimated probabilities into binary indicators. Specifically, a false positive occurs if the estimated in-game home team win probability is greater than 0.5 and the home team loses ($Y_i = 0$). A false negative occurs when the estimated in-game home team win probability is less than 0.5 and the home team wins ($Y_i = 1$). According to the same summing operations for calculating Brier Score, the number of false positives FP and false negatives FN are counted, and then the two values are added. Using this, the misclassification rate is defined by

$$MR = \frac{FP + FN}{Q}.$$

Because Brier Score maintains the values of the estimated probabilities, it conveys more information about the models ability to predict than the misclassification rate. However, due to the sum of squares in Brier Score, misclassification rate is more easily interpreted. Both of the evaluation metrics for each of all models are provided in Table 2.5.

Table 2.5: Evaluation of predictive performances the 2018-2019 and 2019-2020 seasons.

	2018-2019 Season		2019-2020 Season	
Model	Brier Score	Misclass Rate	Brier Score	Misclass Rate
MLE	0.1453	0.2183	0.1397	0.2084
Probit	0.1452	0.2180	0.1398	0.2086
Bayes	0.1451	0.2182	0.1396	0.2081
Dyn Bayes	0.1450	0.2182	0.1396	0.2081
Adj Dyn Bayes	0.1284	0.1870	0.1261	0.1827
Benz	0.1247	0.1799	0.1187	0.1688

With respect to these metrics, the MLE, probit model, Bayes model, and Bayes model with dynamic prior appear to perform similarly. However, with values of Q exceeding 26,000,000, these slight differences may be indicative of a true difference in performance. In contrast, the adjusted Bayesian model with dynamic prior clearly out performs the other methods, having both lower Brier Score, and misclassification rate. Benz's model typically outperforms the best of the Bayesian approaches. The faster decrease for the contribution of pregame spread and faster increase for the contribution of point differential may account for the improved performance. Additionally, Benz's model only updates when a play-by-play event occurs as opposed to every second (see Table 2.4) which may explain the differences between the models' performances.

Tables 2.6 and 2.7 show Brier Score and the misclassification rates, respectively for the methods in this paper at the half (20 minutes remaining), with 10, 2, and 1 minute remaining. All methods improved substantially as the game progresses. The two best models are the adjusted dynamic Bayes and Benz's model. There is very little difference in the performance of the other models.

Table 2.6: In-game evaluation of Brier Score for the 2018-2019 seasons.

	2018-2019 Season				2019-2020 Season			
	Time remaining (in minutes)							
	20	10	2	1	20	10	2	1
Probit	0.149	0.109	0.060	0.050	0.141	0.102	0.061	0.051
Bayes	0.149	0.108	0.060	0.050	0.141	0.102	0.060	0.050
Dyn Bayes	0.149	0.108	0.060	0.050	0.141	0.102	0.060	0.050
Adj dyn Bayes	0.139	0.106	0.060	0.050	0.134	0.101	0.060	0.050
Benz	0.130	0.101	0.056	0.039	0.135	0.095	0.048	0.039

Table 2.7: In-game evaluation of missclassification rate for the 2018-2019 seasons.

	2018-2019 Season				2019-2020 Season			
	Time remaining (in minutes)							
	20	10	2	1	20	10	2	1
Probit	0.224	0.156	0.086	0.070	0.206	0.147	0.082	0.070
Bayes	0.223	0.156	0.085	0.070	0.203	0.147	0.082	0.070
Dyn Bayes	0.223	0.156	0.085	0.070	0.203	0.147	0.082	0.070
Adj dyn Bayes	0.197	0.149	0.085	0.070	0.186	0.140	0.082	0.070
Benz	0.192	0.140	0.070	0.051	0.198	0.131	0.063	0.051

2.4 Application of Model to a Specific Game

To illustrate utility for a single game, the six estimation techniques were applied to the 2016 NCAA Division 1 Championship Game between the University of North Carolina (UNC) and Villanova University. The traces of the estimated in-game home team win probabilities for all six methods are seen in Figure 2.9. The game was played at NRG Stadium in Houston, Texas, a neutral site. UNC was listed as the home team, as one of the four #1 seeds in the Tournament that year. Villanova was a #2 seed, but was favored slightly over UNC by `teamrankings.com`. The estimated pregame win probability for UNC to win the game was $\hat{p}_p = 0.49$. The first three points of the game were scored one minute and five seconds into the game by UNC's Joel Berry II. Nineteen seconds later, Kris Jenkins of Villanova made a lay-up, and the score was 2-3. The game stayed close, and at the end of the first half (1,200 seconds into the game), UNC led Villanova 39-34. An investigation of the

six estimators shows that the MLE, probit model, Bayes, and unadjusted dynamic Bayes all quickly react to the early scores, bumping up the UNC win probability to as high as 0.75. Benz's model also increases the UNC win probability, but not as much as the other four. The adjusted dynamic Bayes changes very little. At the half, when UNC's lead was five points, the four unadjusted methods predict UNC will win with a probability of around 0.85. Benz's model returns a probability of 0.75. On the other hand, the adjusted Bayesian estimate does not react as drastically as the other four in the early part of the game, because the pregame win probability is influencing the returned values.

Five minutes and 52 seconds (1,552 seconds) into the second half, Villanova had tied the game. At that time, the estimated probability of a UNC win drops to around 0.5 for all six methods, and Benz's method returns a value of slightly less than 0.5. As the game continued, and the clock wound down, the six estimates increase and decrease according to the score differential. In the last few minutes of the game, the change in the six estimates is more unified in response to the change in score differential than in the first half of the game. At around time 2,370 seconds elapsed, Villanova had an eight point lead, and all models give UNC a very small chance of winning. UNC pulled within one point with one minute remaining. With six seconds remaining in the game, Marcus Page of UNC hit a 3-point jump shot to tie the game at 74-74, and all six estimators jump to a home team win probability of greater than 0.5. With less than one second remaining, Kris Jenkins shot a 3-pointer which went through the net after time expired. Villanova won the championship by a score of 77-74, and all six estimators drop to zero, showing how Villanova taking the lead with no time remaining drastically affects the home team's (UNC) probability of winning.

Figure 2.9: Win probability graph for 2016 NCAA Championship Game.

2.5 Conclusion

For NCAA basketball games, two new methods are proposed for estimating or predicting in-game home team win probabilities. The first newly proposed method is a Bayesian estimator with a prior distribution that changes as a function of lead differential and time elapsed, which was called the Bayesian estimator with a dynamic prior. The second method adds to the original estimate a time-weighted adjustment based on pregame win probability computed from daily ratings. In this paper, the adjustment was applied only to the Bayesian estimate with dynamic prior. It is reasonable to conclude the adjustment would improve the performance of the other estimators, just as it did the dynamic Bayesian estimator. A comparison of the methods for the purpose of estimation shows that the two proposed estimates outperforms the estimates from the three standard methods, and is competitive with the logistic model of Benz. For prediction, the adjusted dynamic Bayesian method

out performs the other, based on a comparison of both Brier Score and misclassification rates. There are a number of additional problems to be investigated. First note the methodology can be easily restructured based on the total time of the game to apply to other basketball leagues, most notably the National Basketball Association (NBA). Also of interest is the effect on the estimators for different window choices. Another consideration is that the adjustment resulted in a substantial improvement on the prediction of the adjusted dynamic Bayesian estimator. The adjusted dynamic Bayes model performed almost as well as Benz's model. Another area of investigation that remains is to determine a function of time that gives pregame win probability a more quickly decreasing role than the linear function considered here. It is also likely that other pregame metrics, or some combination of pregame metrics, in place of the pregame win probability derived from the power rankings (found on `teamrankings.com`) might improve the outcome. These other metrics include different power rankings, ELO rating, score-differential, and other team statistics; any of these can be used singularly, as in this paper, or combined. Finally, the development of these types of models for other sports also presents unique challenges that are worth investigating.

CHAPTER THREE

Bayesian Estimation of In-Game Home Team Win Probability for National Basketball Association Games

3.1 Introduction

Sports analytics has become a well-established area of research. Work spans more than 30 years and a large range of difficulty. Since the early 2000s, research in statistical methods for sports analytics has risen dramatically. The review articles of Kubatko et al. (2007), Santos-Fernandez et al. (2019), and Terner and Franks (2021) provide a fairly comprehensive review for sports analytics for a wide variety of sports, including basketball. One problem of interest is predicting the probability that the home team wins during the course of the game, or predicting "in-game win probability."

Speaking broadly, models for predicting the outcome of a sporting event can be classified into two systems: (1) pregame prediction or (2) in-game, or in-play, prediction. Pregame prediction involves determining the outcome of a game before play begins. Once play begins, the process of predicting the outcome ends. In contrast, in-game prediction attempts to use the progress during a game to determine win probabilities that vary as a function of in-game variables, for example, elapsed game time or score difference. The focus of this paper is in-game prediction.

For a variety of sports, there are many different methods for accurately estimating in-game win probability found in the literature. For basketball, one of the first attempts to estimate in-game basketball analytics is Westfall (1990) who developed a graphical summary of the scoring activity for a basketball game that is a real-time plot of the score difference versus the elapsed time. The features of the graph provided easy access to largest leads, lead changes, come-from-behind activ-

ity, and other interesting game features. As computing technology and algorithms have become more sophisticated, models for forecasting in-game win probability have become more complex. Some are built on expert predictions, some on betting paradigms, and others on within-game metrics. Shirley (2007) modeled a basketball game using a Markov model with three states and used that model for estimating in-game win probability. Štrumbelj and Vračar (2012) improved upon that by taking into consideration the strengths of the two teams and estimated transition probabilities using performance statistics. Vračar et al. (2016) extended the state description to capture other facets beyond in-game states so that the transition probabilities become conditional on a broader game context. Bashuk (2012) proposed using cumulative win probabilities over the duration of a game to measure team performance. Using five years of game play, he generated a win probability index for NCAA basketball. Using the index, he created an open system to measure the impact, in terms of win probability added, of each play. More recently, Benz (2019) developed a logistic regression approach where the coefficients of the covariates are allowed to change a function of time, i.e., the effects of the coefficients are dynamic in nature. Chen and Fan (2018) developed a method for modeling point differences using a functional data approach. Shi and Song (2019) develop a discrete-time, finite-state Markov model for the progress of basketball scores, and use it to conditionally predict the probability the home team wins or loses by a certain amount. Song and Shi (2020) present an in-play prediction model based on the gamma process. They apply a Bayesian dynamic forecasting procedure that can be used to predict the final score and total points. Song et al. (2020) modify the gamma process by employing betting lines, letting the expectation of the final points total equal the pregame betting line. Maddox et al. (2022a) develop three Bayesian approaches with dynamic priors. The adjusted model with a dynamic prior is, overall, the better of their three proposed methods. In this paper, we adopt the

approach of Maddox et al. for predicting in-game win probabilities for games in the National Basketball Association (NBA). Additional considerations are made to their methodology upon the extension to the NBA, such as an refinement of the prior, improvement of the binning method and a more sophisticated adjustment from pregame information into the model for estimating in-game win probabilities. In addition, the win probabilities that ESPN publishes on their website have been collected to compare to the methodology proposed by Maddox et al..

The remainder of the paper is organized as follows. Section 3.2 provides a brief description of the data and the data collection process. Section 3.3 contains an more thorough explanation of the dynamic Bayesian estimator in Maddox et al. (2022a) and describes the adjustments necessary to implement their approach to games in the NBA. Section 3.4 presents the adjusted dynamic Bayesian method for estimating in-game home team win probabilities, along with a comparison of performances of the dynamic Bayesian and adjusted dynamic Bayesian estimators and the ESPN counterpart. To illustrate utility, Section 3.5 applies the Bayesian approaches to a specific NBA game, and compares that to the ESPN counterpart. The paper concludes with a summary in Section 3.6.

3.2 Data Collection

The primary goal of the models that developed within is to find a practical approach for effectively predicting regular season in-game home team win probability for a single NBA game or a collection of NBA games. The data collected for investigating the proposed models performances were taken from ESPN. Specifically, play-by-play data from ESPN was scraped using *R* (R Core Team, 2021) and the package `rvest` (Wickham, 2021). The data was collected starting with the beginning of the 2012-13 NBA season through the 2019-20 season, when play was halted due to the COVID-19 pandemic. Due to the unprecedented and unpredictable nature

of playing the end of the 2019-20 season inside of a "bubble" with all games on a neutral court, along with attendance limitations and inconsistent schedules for the 2020-21 season, the end of the 2019-20 season and the entire 2020-21 season are omitted. The postseason is also not included, since it is conceivable that postseason games have different behavior than regular season games.

Along with collecting the play-by-play data from ESPN, ESPN has their own win probability model. Their model for estimating in-game win probability is not accessible, and may be considered a "black box" model. The predicted win probabilities for the 2018-19 and 2019-20 seasons are available on ESPN's API that can be accessed through the ESPN Developer Center.[1] There is a general form of the URL for NBA game back-end data providing access to the win probabilities. For each game, the game id changes in the URL. The game ids are scraped for all games in each season, then input into this URL to scrape the win probabilities throughout the game.

3.3 Dynamic Bayesian Estimator

For a specific game, consider the random process that is the home team's lead at time $t = 0, 1, \ldots, 2879$, where t is the game time elapsed in seconds. At a specific time t and for a specific home team lead ℓ, let $p_{t,\ell}$ denote the in-game probability that the home team will win the game at the end of regulation. When considering multiple games $i = 1, 2, \ldots, M$, let $Y_i = 1$ if the home team wins game i and 0 otherwise. Consider $p_{t,\ell}$ as a continuous function of t and ℓ. Maddox et al. (2022a) establish an estimator of $p_{t,\ell}$ using a Bayesian approach. For each cell, they combine the data with a beta$(\alpha_{t,\ell}, \beta_{t,\ell})$ prior where $\alpha_{t,\ell}$ and $\beta_{t,\ell}$ are chosen based on t and ℓ. Specifically, the game is partitioned into cells based on time (in seconds) and point differential. Some (t, ℓ) cells may have small values of $N_{t,\ell}$ = number of games in that

[1] The description of the ESPN Developer Center may be found at www.espn.com/apis/devcenter/overview.html.

cell, which have the potential to result in estimators of $p_{t,\ell}$ with very large standard errors. To address this issue, windows can be defined and centered on (t, ℓ) in such a way that the in-game win probability remains relatively constant across the window. In basketball, since no offensive possession can result in more than four points the window with respect to ℓ can be reasonably defined as $[\ell - 2, \ell + 2]$. Moreover, since most offensive possessions last at least six seconds the width of the time window is taken to be six. The same notation will be adopted for any $[t-3, t+3] \times [\ell - 2, \ell + 2]$ window; that is, $N_{t,\ell}$ is the number of games in the window in which the home team has led by any value in $[\ell - 2, \ell + 2]$ points after any time in $[t - 3, t + 3]$ seconds and $n_{t,\ell} = \sum_{i=1}^{N_{t,\ell}} Y_i$, distributed as a binomial$(N_{t,\ell}, p_{t,\ell})$ random variable. Based on the binomial distribution, a simple estimator for in-game home team win probability for for each (t, ℓ) combination is the maximum likelihood estimator

$$\bar{p}_{t,\ell} = \frac{n_{t,\ell}}{N_{t,\ell}}. \tag{3.1}$$

In Section 3.3.2, limitations of this binning method are explained and addressed.

Then Maddox et al. apply a Bayesian methods approach to estimate $p_{t,\ell}$. Since the beta family of distributions is a conjugate prior for the binomial distribution, the beta-binomial connection is used to estimate $p_{t,\ell}$. Let $\alpha_{t,\ell} > 0$ and $\beta_{t,\ell} > 0$ be the shape parameters of the beta prior on $p_{t,\ell}$. Then for each window within the (t, ℓ) plane, the Bayes estimator of $p_{t,\ell}$ is the mean of the posterior beta distribution, specifically

$$\hat{p}_{t,\ell} = \frac{n_{t,\ell} + \alpha_{t,\ell}}{N_{t,\ell} + \alpha_{t,\ell} + \beta_{t,\ell}}. \tag{3.2}$$

In Maddox et al., the choice of $\alpha_{t,\ell}$ and $\beta_{t,\ell}$ are dependent both on the time remaining in the game and the score differential, leading to a "dynamic prior." For the NBA dynamic Bayesian estimator, the choice of parameters for the beta prior is described in Section 3.3.1.

3.3.1 Dynamic Prior for NBA Games

For the NBA, a more precise prior structure for the structure of the prior distribution is proposed compared to Maddox et al. (2022a). A sample of 14 NBA experts, including NBA front office associates were polled. For each combination of elapsed time and home team lead in Table 2.1, the experts provided their estimation of the probability of a team winning, regardless of home team. The sample mean $\tilde{p}_{t,\ell}$ and sample variance $s^2_{t,\ell}$ of the probabilities were computed. The two scale parameters were estimated via a method-of-moments type approach. The system of equations

$$\begin{aligned}\tilde{p}_{t,\ell} &= \frac{\alpha_{t,\ell}}{\alpha_{t,\ell}+\beta_{t,\ell}},\\ s^2_{t,\ell} &= \frac{\alpha_{t,\ell}\beta_{t,\ell}}{\left(\alpha_{t,\ell}+\beta_{t,\ell}\right)^2\left(\alpha_{t,\ell}+\beta_{t,\ell}+1\right)},\end{aligned}$$

is solved for $\alpha_{t,\ell}$ and $\beta_{t,\ell}$, yielding

$$\begin{aligned}\alpha_{t,\ell} &= -\frac{\tilde{p}_{t,\ell}\left(\tilde{p}^2_{t,\ell}-\tilde{p}_{t,\ell}+s^2_{t,\ell}\right)}{s^2_{t,\ell}}\\ \beta_{t,\ell} &= \frac{\left(\tilde{p}_{t,\ell}-1\right)\left(\tilde{p}^2_{t,\ell}-\tilde{p}_{t,\ell}+s^2_{t,\ell}\right)}{s^2_{t,\ell}},\end{aligned}$$

as long as $(\tilde{p}_{t,\ell}-1)\tilde{p}_{t,\ell}(\tilde{p}^2_{t,\ell}-\tilde{p}_{t,\ell}+s^2_{t,\ell}) \neq 0$.

In Table 2.1, the score differential is presented when the home team has the lead. The greater the lead, the more likely the home team will win, and this is modeled with a left-skewed prior. On the other hand, if the visiting team has the lead, then the roles of $\alpha_{t,\ell}$ and $\beta_{t,\ell}$ are reversed, and the prior becomes right-skewed. The prior densities in Figure 3.1 illustrate this principle. The curves are beta density priors plotted by home-team win probability. For example, the fifth row in Table 2.1 for $t = 360, 361, \ldots, 719$ seconds and the home team has a lead of $\ell = 10, 11, \ldots, 19$ points has a beta(19, 7) prior seen in the figure as the blue density curve. On the other hand, if the away team as the same lead, then the prior is a beta(7, 19), represented by the red curve. At any time, if the game is tied ($\ell = 0$), or is

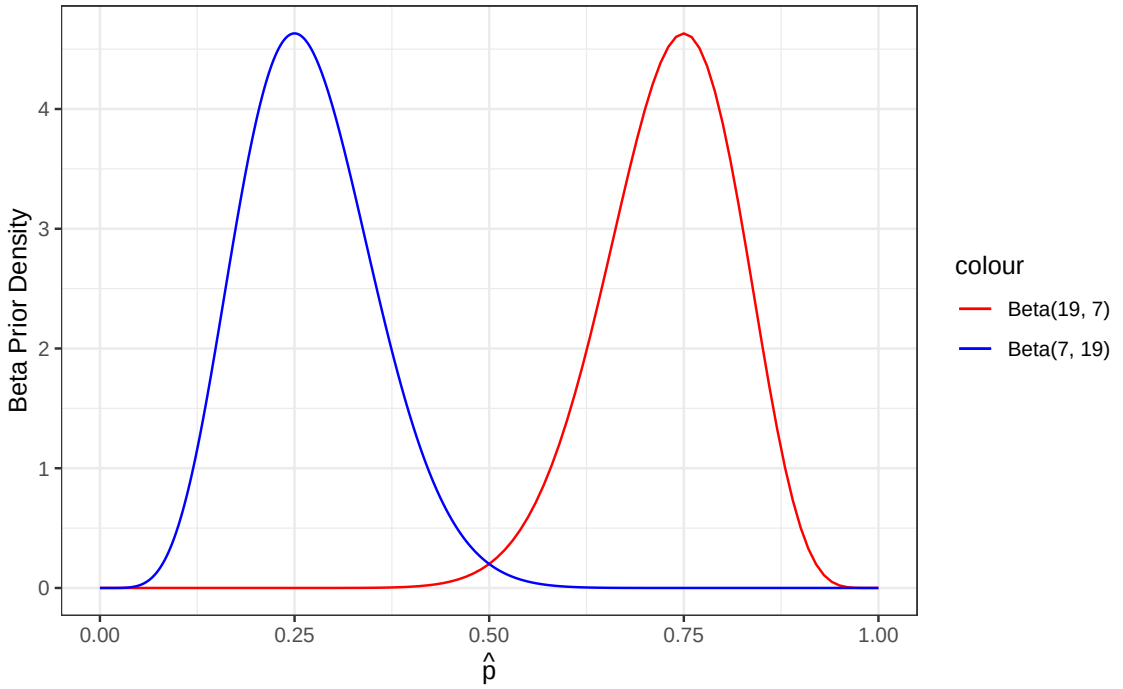

Figure 3.1: Densities of beta(19, 7) prior (in blue) and beta (7, 19) prior (in red).

sufficiently close in score for the amount of time remaining, the prior distribution is flat, or "uninformative," meaning it gives equal weight to both teams winning the game.

Figure 3.2 shows the maximum likelihood estimates (MLE) $\bar{p}_{t,\ell}$ from equation (3.1) and Figure 3.3 the Bayesian estimates $\hat{p}_{t,\ell}$ from equation (3.2) for the same data. The graph of the MLE is not as smooth as the graph of the Bayesian estimates. This is easily explained by noting that the MLE are computed on each (t, ℓ) point, and not across a rectangular bin. There are more points (t, ℓ) with no games (and so no estimate) than rectangular bins with no games. Moreover, since each MLE is computed based on the number of games at point (t, ℓ), the standard error of the MLE is likely to be greater than the corresponding Bayesian estimate.

The Bayesian prior parameters have an interesting interpretation, first noted by Deshpande and Jensen (2016). The parameter $\alpha_{t,\ell}$ can be interpreted as the

number of "pseudo-wins" in that cell; likewise $\beta_{t,\ell}$ as the number of "pseudo-losses." Through this interpretation, the two parameteres can be seen as a way of increasing the number of games in a specific (t, ℓ) cell. If the home team is ahead, then first scale parameter being large effectively acts to increase the number of wins in that cell. On the other hand, if the home team is behind, the second scale parameter is large, and acts to increase the number of losses. These ideas contribute to the smoother appearance of the Bayesian estimates.

Table 3.1: Imputed parameters for beta prior

Elapsed Time (t) (sec.)	Home Team Lead (ℓ)	$\alpha_{t,\ell}$	$\beta_{t,\ell}$
0 – 360	0 – 9	1	1
0 – 360	10 – 14	18	9
0 – 360	≥ 15	54	6
361 – 720	0 – 9	1	1
361 – 720	10 – 19	19	7
361 – 720	≥ 20	34	3
721 – 1440	0 – 9	1	1
721 – 1440	10 – 19	18	5
721 – 1440	≥ 20	51	2
1441 – 2160	0 – 9	1	1
1441 – 2160	10 – 14	22	6
1441 – 2160	15 – 19	15	2
1441 – 2160	≥ 20	71	2
2161 – 2520	0 – 9	1	1
2161 – 2520	10 – 14	22	3
2161 – 2520	15 – 19	25	2

Elapsed Time (t) (sec.)	Home Team Lead (ℓ)	$\alpha_{t,\ell}$	$\beta_{t,\ell}$
2161 – 2520	≥ 20	133	2
2521 – 2700	0 – 9	1	1
2521 – 2700	10 – 14	46	3
2521 – 2700	15 – 19	48	1
2521 – 2700	≥ 20	133	1
2701 – 2820	0 – 4	1	1
2701 – 2820	5 – 9	10	2
2701 – 2820	10 – 14	104	3
2701 – 2820	≥ 15	328	2
2821 – 2879	0 – 2	1	1
2821 – 2879	3, 4	10	2
2821 – 2879	5 – 9	17	1
2821 – 2879	≥ 10	167	1

3.3.2 Binning Procedure for NBA Games

The approach in Maddox et al. (2022a) fixes bin area; in other words, regardless of time remaining and score differential, the length and width of all bins are equal. However there are practical issues with fixed area binning, especially closer to the end of the game. Consider that during the middle of the game, a two point difference in score should not have a major effect on the win probability. However, closer to the end of a close game a two point difference in score could have a very large effect

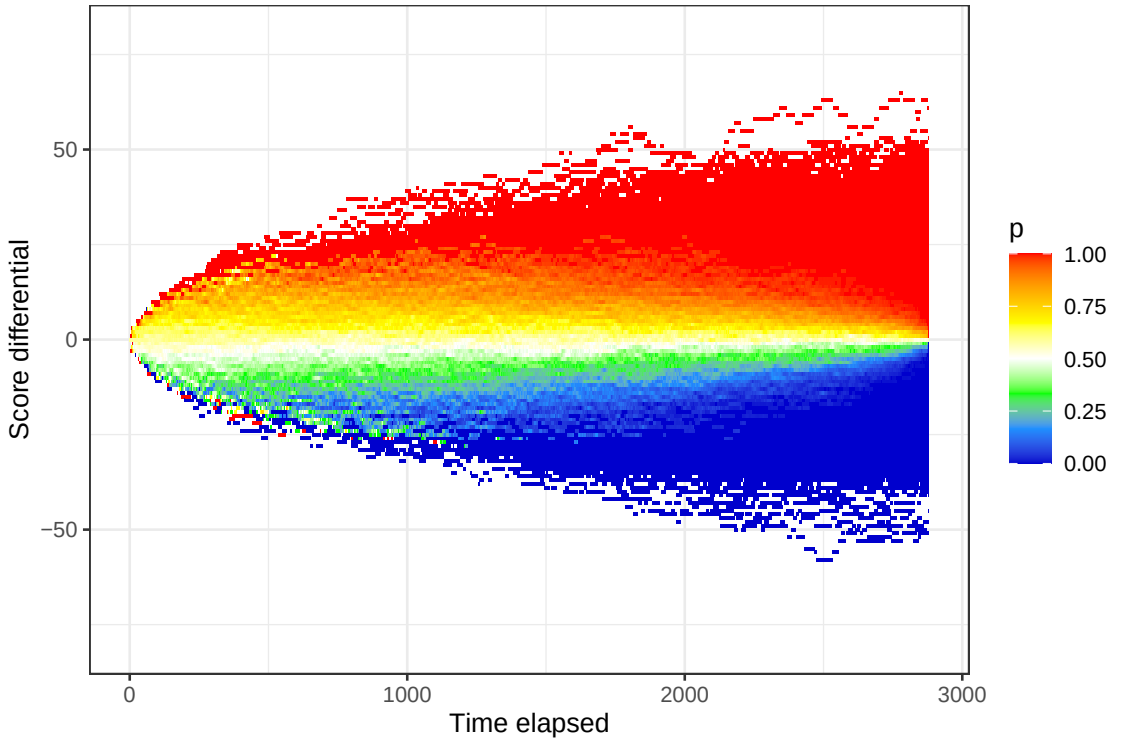

Figure 3.2: Maximum likelihood estimate of home team win probability.

on win probability. For example, suppose there are five seconds remaining and a tie score with a jump ball to take place. Intuitively there should be about a 50% probability that either team wins the game. However, with the same amount of time if one of those teams is up by two points, that team should have a greater than 50% probability of winning. In this scenario, it is not reasonable to use the same bin width on score differential at the end of the game as at the beginning. We propose that between 2700 and 2820 seconds into the game, or three to one minutes remaining, the width of the score differential bins is shortened to $[\ell - 1, \ell + 1]$, and after 2820 seconds into the game, or in the last minute, there is no binning on score differential. The bin widths on time remain the same at the end of the game, because with a specific score differential a small shift in time should not have a large effect on win probability even at the end of the game.

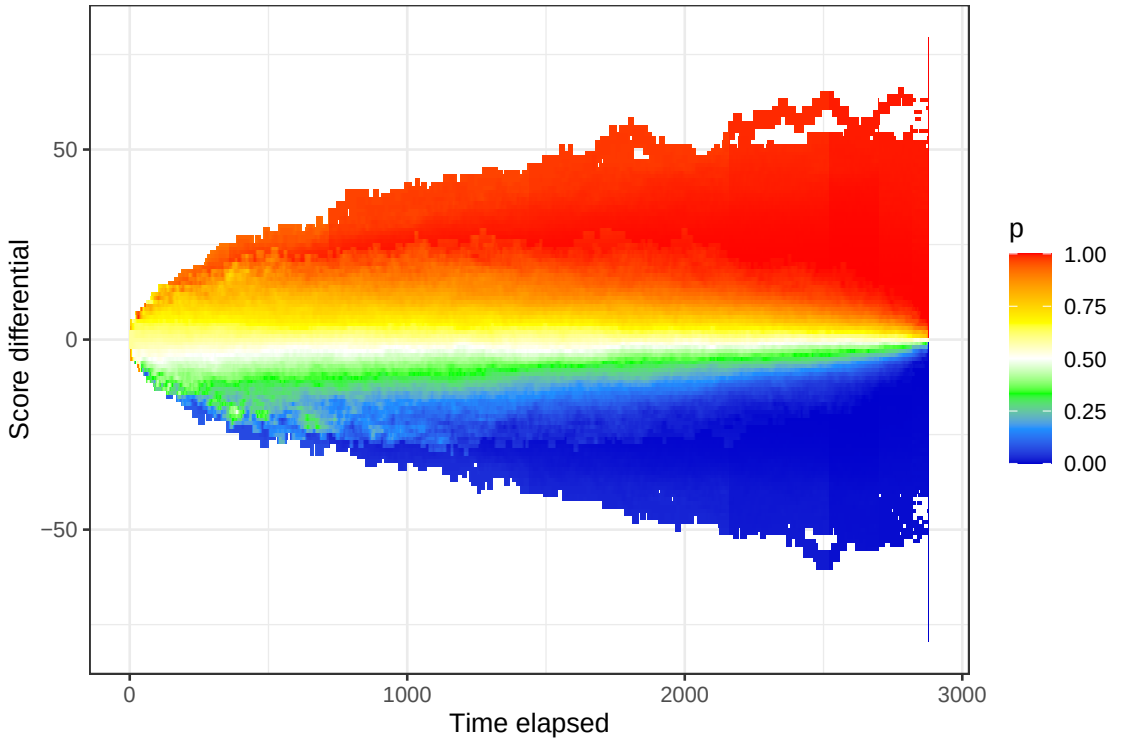

Figure 3.3: Dynamic Bayesian estimate of home team win probability.

3.4 Adjusted Dynamic Bayesian Estimator

In-game win probability is certainly a function of time and score differential during the game. However, it is also affected by the skill of the teams playing the game. Incorporating some measure of team ability in to the model was also discussed in Maddox et al. (2022a). The normal distribution quantile function was used to convert the pregame point to a pregame home team win probability $\hat{p}_p$ for each team. The pregame probability, $\hat{p}_p$, was added into the model for predicting in-game win probability so that the weight of $\hat{p}_p$ decreased linearly as a function of time remaining in the game to get a final adjusted Bayesian estimate of in-game home-team win probability. In what follows, the model for finding $\hat{p}_p$ and including it in the model is refined by allowing for more complex models for including $\hat{p}_p$.

The resulting final "adjusted dynamic Bayesian" estimates are compared with the non-adjusted dynamic Bayesian estimates and the ESPN counterparts.

3.4.1 Brier's Score

Brier's score is a statistic used to compare the performance of different methods for estimating probabilities. Brier's score is the average of the square of the difference between the estimated probability and the observed binary outcome. In the context of in-game home team win probabilities, this observed binary outcome is denoted y_i and is the observed value of Y_i as defined in Section 3.3. To interpret Brier's score, if $Y_i = 1$ for all i, and the predicted probability is also one for every i, then Brier's score will be zero, indicating perfect prediction. On the other hand, if for all $i, Y_i = 0$, and the estimated probability is one, then Brier's score will be one, the worst possible Brier's score.

To compute Brier's score, if $\rho_{t,\ell}$ represent the estimated in-game home team win probability. For each (t, ℓ) cell, let $N^*_{t,\ell}$ represent the number of games in the cell in which the home team led by ℓ points at time t; that is, $N^*_{t,\ell}$ is the number of games observed in that cell. Then for non-missing estimated probabilities, Brier's score is

$$B = \frac{1}{Q} \sum_{t=0}^{2879} \sum_{\ell=-58}^{58} \sum_{j=1}^{N^*_{t,l}} \left(\rho_{t,\ell} - y_j\right)^2,$$

where Q is the sum of $N^*_{t,\ell}$ in cells without missing $\rho_{t,\ell}$. When evaluating the models using Brier's Score, the 2018-19 and 2019-20 seasons are used as testing data, each having $Q = 6,590,576$ observations.

3.4.2 Pregame Win Probabilities

Maddox et al. (2022a) adjustment to dynamic Bayesian estimator by including a measure of pregame win probabilities, linearly shifting the weight from pregame

win probabilities to in-game win probabilities across using

$$\hat{p}^*_{t,\ell} = \left(\frac{S-t}{S}\right)\hat{p}_p + \left(\frac{t}{S}\right)\hat{p}_{t,\ell}, \tag{3.3}$$

where S is the number of seconds in the game, and is 2,880 for NBA games[2] , $\hat{p}_{t,\ell}$ is given in equation (3.2) using the binning procedure described in Section 3.3.2 and the prior structure given in Table 3.1, $\hat{p}^*_{t,\ell}$ is the final predicted win probability and $\hat{p}_p$ is the pregame win probability. Incorporating pregame win probabilities into the model helps to improve predictive accuracy because team quality plays a role in predicting who may win a game. The linear adjustment is simple, and only a function of time remaining.

In what follows, we investigate three different functions for incorporating pregame probabilities. The first is a linear function of only time remaining. The second is a linear function of time remaining and score differential. The third is linear in time, but quadratic in score differential. Other variations to the three weight functions were considered. For example, including a quadratic term for time was attempted. However, the more complicated models would not converge appropriately for R to optimize the performance accurately.

Let the A be some set of numbers. Then for a specific number a, the indicator function $\mathbb{1}_A(a)$ is defined as

$$\mathbb{1}_A(a) = \begin{cases} 1 & \text{if } a \in A, \text{ and} \\ 0 & \text{otherwise.} \end{cases}$$

The three weight functions specifically considered here are

$$B_1 = bt,$$

$$B_2 = c_1 t + c_2|\ell|,$$

$$B_3 = d_0 + d_1\mathbb{1}_0(\ell) + d_2 t + d_3|\ell| + d_4\ell^2.$$

[2] In Maddox et al., $S = 2,400$, the number of seconds in the NCAA basketball game.

Each one of these weight functions yields a competing model for including pregame win probabilities, which can be represented

$$p^*_{t,\ell,j} = \begin{cases} \hat{p}_p, & B_j \leq 0 \\ (1 - B_j)\hat{p}_p + B_j\hat{p}_{t,\ell}, & 0 < B_j < 1 \\ \hat{p}_{t,\ell} & B_j \geq 1 \end{cases}, \qquad j = 1, 2, 3.$$

The model for $p^*_{t,\ell,1}$ is equivalent to the model in equation (3.3) for $b = 1/2880$ of Maddox et al.. The weight function B_2 includes a linear dependence on both time and score differential. Additionally, if the value of B_2 is greater than one, only $p^*_{t,\ell,2} = \hat{p}_{t,\ell}$, the Bayesian estimator in equation (2.2). Finally, $p^*_{t,\ell,3}$ specifies a linear effect in time remaining, but quadratic effect of score differential as well as an intercept term, d_0. As with B_2, if $B_3 > 1, p^*_{t,\ell,3} = \hat{p}_{t,\ell}$. Since ℓ enters B_3 through only $|\ell|$ and ℓ^2, B_3 is more likely to be greater than 1.

Values for $b, c_1, c_2, d_0, d_1, \ldots, d_4$ are estimated by choosing those values which minimize Brier score for each $p^*_{t,\ell,j}$ tested on data from the 2018–19 and 2019–20 seasons. The Brier score for each $p^*_{t,\ell,j}$ is shown in Table 3.2. The model $p^*_{t,\ell,3}$ is the most accurate predictor when using seasons 2018-20 to evaluate the model built on seasons 2012-2018. The fitted model for B_3 that provided the minimal Brier score is

$$B_3 = -1.10633 - 0.02313\mathbb{1}_0(\ell) + 0.00027t + 0.06618|\ell| - 0.00139\ell^2.$$

Using this expression for B_3, the model

$$p^*_{t,\ell,3} = \begin{cases} \hat{p}_p & B_3 \leq 0 \\ (1 - B_3)\hat{p}_p + B_3\hat{p}_{t,\ell}, & 0 < B_3 < 1 \\ \hat{p}_{t,\ell}, & B_3 \geq 1 \end{cases}$$

is the "adjusted dynamic Bayesian estimator."

Table 3.2: Brier Scores for models determining pregame probability proportion.

Proportion Model	Brier Score
Linear Time (B_1)	0.1622
Linear Time & Score (B_2)	0.1613
Quadratic (B_3)	0.1598

Figures 3.4 through 3.6 illustrate the estimated in-game home team win probabilities using the adjusted Bayes estimator, $p^*_{t,\ell,3}$. Because the estimated probability is affected by the pregame home team win probability, three values of $\hat{p}_p$ were selected to illustrate the performance of the newly proposed estimator. Of the 7,376 games in the training data, approximately 59% were won by the home team, motivating the choice of $\hat{p}_p = 0.59$. The other two choices of pregame home team win probability are $\hat{p}_p = 0.59 \pm 0.3$.

3.4.3 Model Comparison

For each second of each game from the 2018-19 and 2019-20 seasons, Brier's score was computed. The Brier Scores for each of the three models are shown in Table 3.3. As mentioned previously, ESPN's win probability model is not publicly available, and the model itself cannot be reproduced. However, ESPN does release

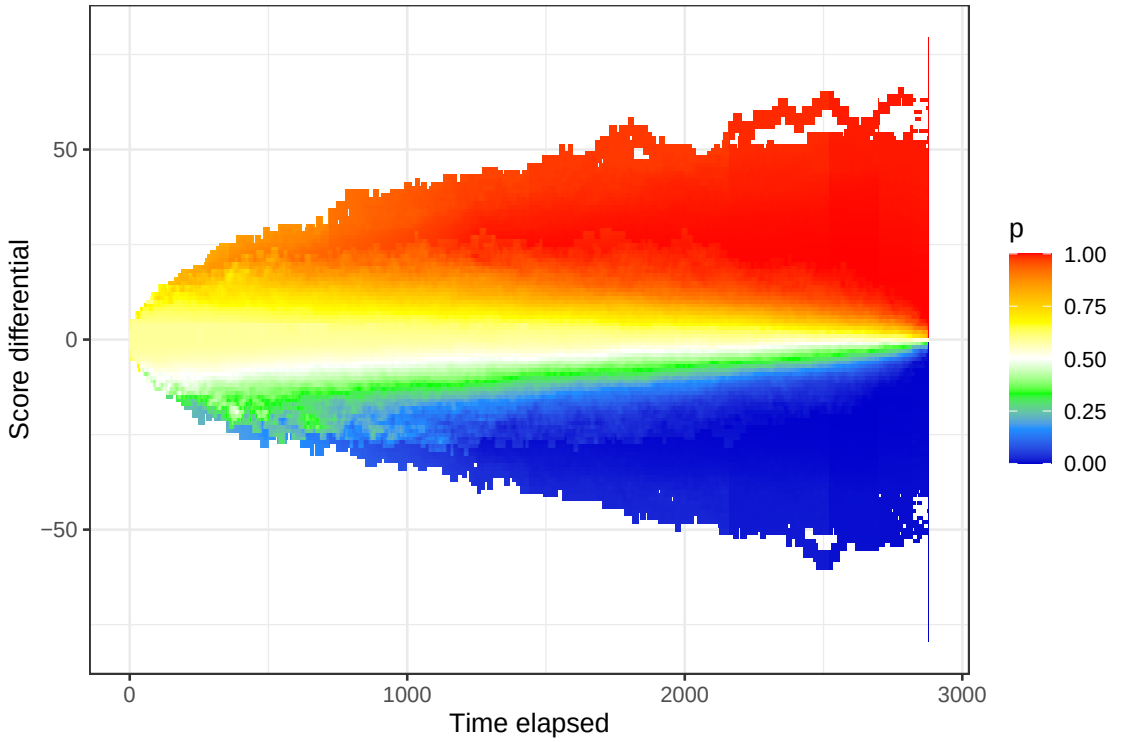

Figure 3.4: Adjusted dynamic Bayesian estimates for average home team pregame win probability.

the results of the model, which were scraped from their website. Using Brier Scores, ESPN's win probability model does outperform the new adjusted dynamic Bayesian model for both seasons in the test data set. However, Tables 3.4 and 3.5 shows the performance of each model at different times during the game. The adjusted dynamic Bayesian model performs better than the ESPN model at several points in the fourth quarter in both seasons. This may indicate that future improvements could be made to the adjusted dynamic Bayesian model by adjusting the pregame portion of the model.

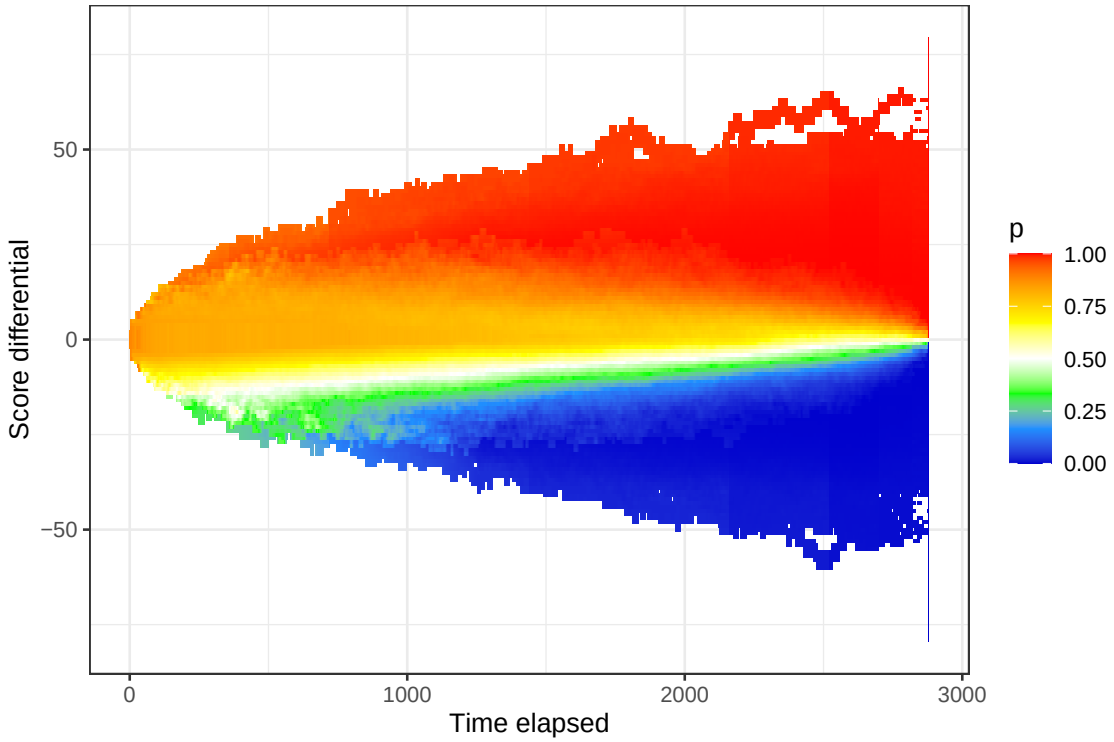

Figure 3.5: Adjusted dynamic Bayesian estimates for above average home team pregame win probability.

Table 3.3: Brier Scores for predictive performances for the 2018-2019 and 2019-2020 seasons.

	2018-2019 Season	2019-2020 Season	Total
Bayes with dynamic prior	0.1663	0.1736	0.1697
Adjusted dynamic Bayes	0.1568	0.1635	0.1598
ESPN win probabilities	0.1550	0.1621	0.1582

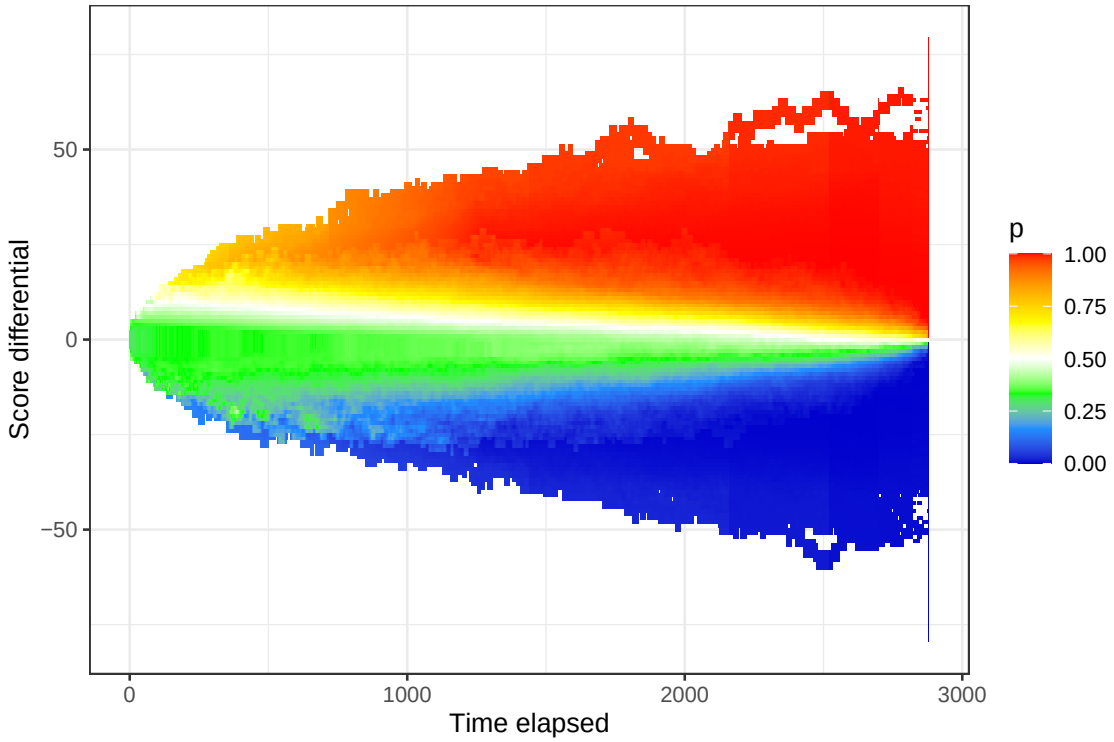

Figure 3.6: Adjusted dynamic Bayesian estimates for below average home team pregame win probability.

Table 3.4: In-game evaluation of predictive performances for the 2018-2019 season.

	Time remaining (in minutes)				
Model	24	12	6	3	1
Dyn Bayes	0.1761	0.1194	0.0935	0.0775	0.0537
Adj Dyn Bayes	0.1702	0.1168	0.0927	0.0773	0.0537
ESPN	0.1666	0.1181	0.0931	0.0769	0.0536

Table 3.5: In-game evaluation of predictive performances for the 2019-2020 season.

	Time remaining (in minutes)				
Model	24	12	6	3	1
Dyn Bayes	0.1840	0.1334	0.1071	0.0822	0.0588
Adj Dyn Bayes	0.1762	0.1300	0.1063	0.0819	0.0588
ESPN	0.1738	0.1311	0.1079	0.0829	0.0584

3.4.4 `TeamRankings.com` *vs. Elo*

While various metrics can be used to arrive at pregame win probabilities, short of adjusting for starting lineups, power rankings should work well in determining the pregame win probabilities. Metrics from TeamRankings.com might not be superior to others. However, many others do not have public daily ratings. Another common and well-researched win probability system, called the Elo ranking system, is often used to predict team performance for a game which can be used to compare the differences between various power ranking systems. The website FiveThirtyEight.com provides historical Elo ratings for each team for each day they played a game dating back to the 1946-47 season.

We can compare the performance of TeamRankings.com win probabilities and Elo win probabilities by both comparing their Brier scores or by comparing their performance when applied to the model introduced in this paper. These results are found in Table 3.6. The lower Brier score for TeamRankings.com pregame predictions compared to that of the Elo model shows TeamRankings.com performs better overall. Additionally, when the Elo pregame win probabilities are applied to

the methodology outlined in this paper the Brier score for the adjusted dynamic Bayesian model is 0.1598 using TeamRankings.com compared to 0.1605 for Elo.

Table 3.6: Brier Scores for TeamRankings.com compared to Elo.

Pregame probability model	Pregame model	In-game model
TeamRankings.com	0.2167	0.1598
Elo	0.2179	0.1605

3.5 Application to a Game

On November 23, 2019 the Chicago Bulls travelled to play at the Charlotte Hornets, in what was expected to be a trivial, low-profile regular season game between two relatively unsuccessful teams. However, the game quickly turned into one of the most exciting games of the season, with Zach LaVine setting a career high of 49 points, including a game-winning 3-point shot with less than one second left in the game[3] . The Bulls ended the game on a 16-7 run in the last minute of the game. The win probabilities for this game produced from each of the three models are displayed in Figure 3.7. The largest differences in estimated win probability for the three models occur early on during the game. This is to be expected, as early on in the game there will be more variance in win probability due to more time allowing for many possible unpredictable occurrences in the game. Throughout the first three quarters of the game, ESPN predicts a larger probability of the Bulls winning the game than the other two models. However, many of the rises and falls in the three models occur at the same time in the game. During the fourth quarter, the three models are close to indistinguishable from each other, each giving very similar results for this particular game. Lastly, at the end of the game all three models had

[3] Since that game, LaVine scored a new career high 50 points on April 9, 2021 against the Atlanta Hawks.

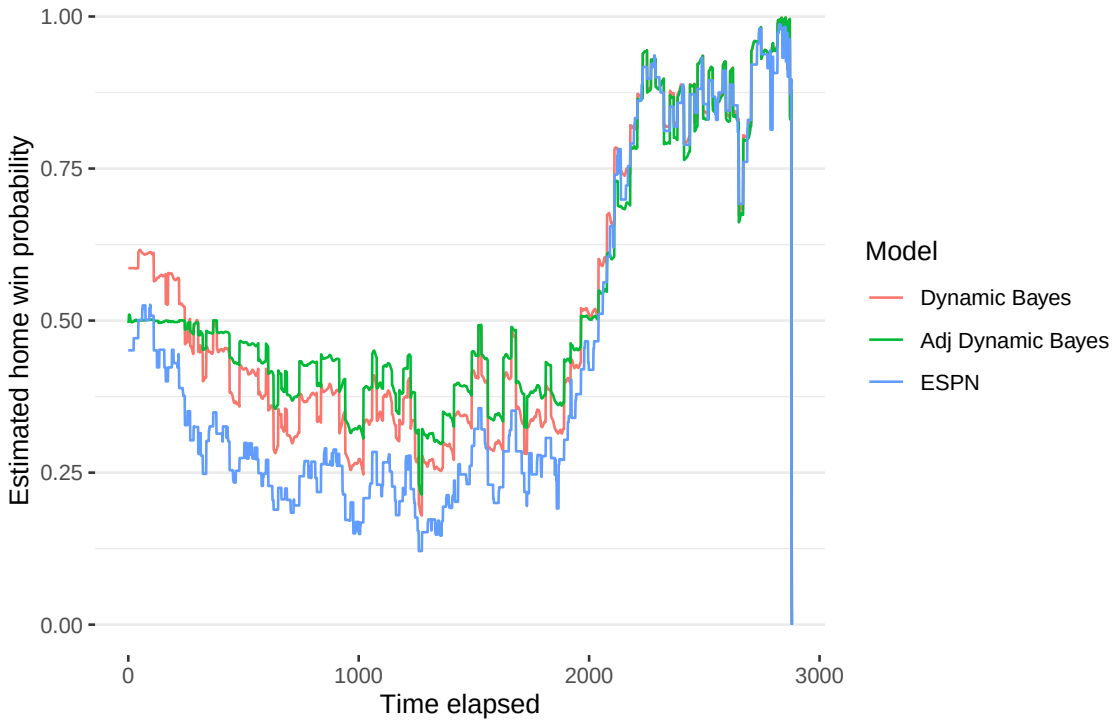

Figure 3.7: In-game win probabilities for Chicago at Charlotte.

an extremely high probability that the Hornets would win, signifying the remarkable run the Bulls went on in the last minute, culminating in a buzzer beating three point shot that swung all three models from predicting a Hornets' win to a Bulls' win. The graph shows this by what appears to be an almost perfectly vertical blue line at the end of the game. However, that line is actually all three models having home win probability close to 1, then plummeting to zero once LaVine's shot is made.

3.6 Conclusion

Two new methods are proposed for estimating in-game win probability for NBA games. Both are an extension and enhancement of the methods in Maddox et al. (2022a), which provide a number of models for estimating in-game home-team win probabilities for NCAA basketball. The first proposed "dynamic Bayesian estimator," uses a prior that has been calculated based on the distribution of pre-

dicted win probabilities from 14 NBA field experts, including several anonymous front office associates within the NBA. The second method, referred to as the "adjusted dynamic Bayesian estimator," adjusts the dynamic Bayesian estimator based on pregame win probabilities obtained from TeamRankings.com. The adjustment is optimized over a function of both time and score so that as the game moves on or the score differential increases, the adjusted dynamic Bayesian estimator will begin to approach the dynamic Bayesian estimator rather than the pregame win probability. These two methods are then compared to the win probability model that ESPN uses for seasons 2018-19 and 2019-20. The ESPN model performs the best overall, but there are times during the game the adjusted dynamic Bayesian model performs the best, indicating that there are some features of the model that estimate probabilities well and some that could be improved upon in the future, such as the calculation of pregame win probability.

CHAPTER FOUR

Bayesian Estimation of In-Game Home Team Win Probability for Division-I FBS College Football

4.1 Introduction

Sports analytics has become a well-established area of research. Work spans more than 30 years and a large range of difficulty. Since the early 2000s, research in statistical methods for sports analytics has risen dramatically. The review articles of Kubatko et al. (2007), Santos-Fernandez et al. (2019), and Terner and Franks (2021) provide a fairly comprehensive review for sports analytics for a wide variety of sports, including football. One problem of interest is predicting the probability that the home team wins during the course of the game, or predicting "in-game win probability."

Speaking broadly, models for predicting the outcome of a sporting event can be classified based on two objectives: (1) pregame prediction or (2) in-game, or in-play, prediction. Pregame prediction involves determining the outcome of a game before play begins. Once play begins, the process of predicting the outcome ends. In contrast, in-game prediction attempts to use the progress during a game to determine win probabilities that vary as a function of in-game variables, for example, elapsed game time or score difference. The focus of this paper is in-game prediction for college football.

One early paper on predicting in-game win probability for baseball is attributed to Lindsey (1963), who used a conditional maximum likelihood estimator to determine the probability the home team wins given the current inning the game and the home team's lead. Lindsey used historical information to determine percentages the home team had won in a variety of scenarios, and applied those to the current

game to determine current team's estimated win probability as the game progressed. Up until this work, baseball decisions were based on what would maximize the scoring output for an inning. In contrast, Lindsey's groundbreaking research focused on determining how each decision would affect the probability that a team wins instead of only the change in expected score. A more recent development in estimating in-game win probability was approached by Benz (2019) who modeled college basketball in-game win probability by using a series of logistic regressions at different times throughout games on score differential and pregame win probability. He then smooths the multiple logistic regression models into a single smooth function. Maddox et al. (2022b) propose a Bayesian model for the National Basketball Association (NBA) based on time and score differential as the two predictors. Much of their methodology can be extended to college football, but different predictors must be considered. In football, score and time are not as informative for estimating win probability. A simple consideration of the two sports makes clear a different model is required for each. For example, in basketball, teams score often and quickly – often within seconds of each other – and one possession can result in zero up to five points. However, it is unusual for a team to score in seconds in a football game, and the possible values for scores are not sequential integers. Additionally, in football, many other variables contribute to the likelihood of a score, such as field position or down-and-distance.

Only a few attempts have been proposed for in-game prediction of win probability for (American) football. Lock and Nettleton (2014) use a random forest of regression trees with up to ten predictor variables to combine pre-play variables to estimate in-game win probability before any play of an National Football League (NFL) game. Pro Football Reference (2012) create a quasi "black box" model, where they go into some detail about creating their win probability model using expected points along with pregame win probability and the known standard deviation of

the end of the game score differential. However, many of the details used for their model are not provided. Later, Ruscio and Brady (2021) compare the performance of the random forest model by Lock and Nettleton and the model put forth by Pro Football Reference when applied to the NFL. Their findings were that there was no discernable difference between the two models. Ruscio and Brady were able to obtain the Pro Football Reference model for their paper to reproduce the results. For Australian rules football, Ryall (2011) used play-by-play data with pregame Elo rankings to develop a model for Australian Rules football. In what follows, a new approach that uses Bayesian methods is proposed. Explanatory variables for predicting in-game win probability include expected number of possessions remaining and expected score differential and with pregame power rankings.

The remainder of the paper is organized as follows. Section 4.2 details the process of gathering and cleaning the data. Section 4.3 presents a method for modeling the expected number of possession remaining in a college football game. In Section 4.4, multiple models are proposed for best estimation of the expected score and compared to each other. The resulting estimates of expected remaining possessions and expected score are used as predictors for the overall win probability model described in Section 4.5. The models for win probability are evaluated and applied to a specific game in Section 4.6. Section 4.7 provides closing remarks.

4.2 Data Collection

The primary goal of the proposed models is effective practical prediction of in-game home team win probability for a single college football game or a collection of college football games during the regular season. The data collected for investigating the proposed models performances were taken from ESPN. Specifically, play-by-play data from ESPN was scraped using *R* (R Core Team, 2021) and the package `rvest` (Wickham, 2021). For football games, ESPN does not display all plays on each

game's webpage by default. Instead, they store the data in a back end server that can be accessed through the ESPN Developer Center. [1] There is a general form of the URL for college football back-end data providing access to the play-by-plays. For each game, the game id changes in the URL. The game ids are scraped for all games in each season, then input into this URL to scrape the play-by-plays for each game. The data was collected starting with the beginning of the 2004 college football season through the 2021 season, excluding the 2020 season due to the uniqueness of that season resulting from the COVID-19 pandemic.

There were some issues that occurred working with the raw data from ESPN's back-end server. As when working with any raw data, there can be typographical errors or mislabels that must be evaluated and corrected, if correction is possible. Clear identifiers were not always reliable; for example, errors were found on which team had possession of the ball. The data was combed through many times to ensure that specific identifiers for the game state were correctly labeled. For some games, play-by-play data was not available in any capacity. This was true especially for games during or close to the 2004 season. There was no clear pattern to which games had no play-by-play data. However, enough games were collected on all seasons that the missing games should not have a negative impact on the analysis.

4.3 Possessions Remaining Model

Within the game of college football, pace or tempo of play has been a key discussion point for the last 20 years. Mike Leach is considered a modern day football pioneer. While the head football coach at Texas Tech University, Leach ushered in a new offensive play style which he carried with him to Washington State University and then Mississippi State University. The style of play is commonly called the "air-raid" or "up-tempo" offense. A primary feature of the play style is that as

[1] The description of the ESPN Developer Center may be found at www.espn.com/apis/devcenter/overview.html.

little time as possible is used between successive offensive plays. The philosophy is minimizing time between snaps prevents the opponent from successfully setting up their defense.

The most common measure of a team's pace is the team's average number of plays per game. However, Troch (2016) introduces time between consecutive offensive plays as a viable alternative. He argues that average number of plays per game does not take into account the tempo of the opponent or the number of run versus pass plays a team calls. Because of incomplete passes, pass plays will stop the clock more frequently than run plays. Therefore, the more pass plays a team attempts, the more plays that team will run during a game, without it necessarily being due to that team's tempo.

For the purposes of the win probability model, interest lies more in the number of possessions that remain in the game than the time between consecutive offensive plays. The more possessions there are remaining, the more snaps, and the more potential points are left for teams to score in the game. Troch's critique of plays per game not accounting for the opponent is valid. A new method is proposed that is similar to the method put forth by Pomeroy (2012) for college basketball. He calculates pace based on the number of possessions a team would expect to have in a game against a team that plays with average tempo.

In what follows, the term "pace" is defined as the expected number of possessions against an opponent that plays with average tempo. Pace is calculated by a recursive algorithm as described with the following steps for each season. For iteration m and team $k = 1, 2, \ldots, n$, let $\xi_{k,m}$ represent pace. For a given season, n is the number of Football Bowl Subdivision (FBS) teams.

(1) For all teams $k = 1, \ldots, n$, initialize $\xi_{k,0} \equiv 0$.

(2) Choose $\delta > 0$ to be small.[2] While $\max_k\{|\xi_{k,m} - \xi_{k,m-1}|\} > \delta$,

[2] The value $\delta = 0.0001$ was used.

(a) Calculate the average pace for all teams,

$$\mu_m = \frac{1}{n}\sum_{k=1}^{n} \xi_{k,m-1}.$$

(b) For each team $k = 1, \ldots, n$, let κ_k represent the collection of indices corresponding to opponent teams played during the regular season. For each team k, sum the paces of the opponent teams,

$$\psi_{k,m} = \sum_{j \in \kappa_k} \xi_{j,m-1}, \quad k = 1, \ldots, n.$$

(c) Find the average difference between the total number of possessions played $x_{k,m}$ and the expected possessions played $\psi_{k,m}$,

$$\epsilon_{k,m} = \frac{x_{k,m} - \psi_{k,m}}{w_k}, \quad k = 1, \ldots, n,$$

where w_k is the number of games the team k played in the given season.

(d) Assign the updated pace for teams based on their number of possessions above expected,

$$\xi_{k,m} = \mu_m + \epsilon_{k,m}, \quad k = 1, \ldots, n.$$

(3) When $\max_k\{|\xi_{k,m} - \xi_{k,m-1}|\} \leq \delta$, the convergence criteria is satisfied and the pace ξ_k for each team is taken to be the final iterative value of $\xi_{k.m}$.

The fastest and slowest team paces for Power 5 FBS teams for the 2021 season are shown in Table 4.1. The Power 5 teams with the highest pace are teams that are known for having up-tempo, pass heavy offenses. For example, Pitt and Tennessee had two of the highest scoring offenses in the nation due to their fast pace. On the other end, teams known for their slow-it-down, run-based offensive style have the lower pace. Kansas State and Boston College are known for their grind-it-out style that does not translate into many possessions or points.

Table 4.1: Fastest and slowest paces for Power 5 teams for the 2021 regular season.

Rank	Team	Pace (ξ_k)
11	Oklahoma State Cowboys	30.11
16	Duke Blue Devils	29.69
21	Pitt Panthers	29.13
22	Colorado Buffaloes	29.04
23	Tennessee Volunteers	28.96
124	Boston College Eagles	23.96
125	Kentucky Wildcats	23.76
127	Oregon Ducks	23.19
128	BYU Cougars	22.90
129	Kansas State Wildcats	22.42

To compute the value of expected possessions remaining after a play, denoted τ, the two teams' paces are averaged and the average is weighted with a decreasing linear function of the time left in the game. If ξ_1 and ξ_2 represent the pace for the two teams in the game, and t the number of seconds elapsed during the game, then τ is given by

$$\tau = \left(\frac{3600 - t}{3600}\right)\left(\frac{\xi_1 + \xi_2}{2}\right). \tag{4.1}$$

4.4 Point Value Model

There have been many attempts to accurately predict the point value of a drive in football. Many models are explained in generalities, but the details are vague. For example, one of the more notable expected point models was posted by a user named

"Mike" and is featured on `Sports-Reference.com` Pro Football Reference (2012). A linear model is used to predict the average point value for a possession given down, distance, and yards to end zone. However the website does not disclose the specifics of the model. The author of this paper contacted `SportsReference.com`. They did not wish to provide any details regarding the model.

Many of the models predict the point value of the current drive. However, they fail to account for the dependency of the success of the next drive on the outcome of the current drive. Consider, for example, the situation when a team has a fourth down 99 yards away from the end zone. That team will almost certainly punt, which will typically result in the opposing team receiving the ball with good field position, making that opponent more likely to score. On the other hand, if a team has the ball just a few yards from the end zone, they are likely to score. They will then kick the ball off and the opponent would likely start their following possession about 75 yards from the end zone. These two contrasting scenarios illustrate the effect of a prior possession on the likelihood of the subsequent possession resulting in a score and thus on the in-game win probability. Therefore, instead of modeling the point value of only the current possession, the proposed model is built on the combined score for both the current possession and the following opponent possession.

Four competing models are used to predict the point value of the current and ensuing possessions using the predictor variables in Table 4.2. The first model is a linear regression model with no interactions. The second is a linear regression that includes all $2^6 = 64$ interactions between the nine predictors, with the exception of any interactions involving offensive pace, defensive pace, or number of possessions. The third and fourth models are a random forest model and an extreme gradient boosting (XGBoost) model, respectively. Literature on linear regression is prevalent, and the technique is commonly applied. However, random forests and the XGBoost models are more current, lesser known techniques. Random forest and XGBoost

models have become more important with advances in computation. Sections 4.4.1 and 4.4.2 provide overviews of these techniques. For more detail on random forest see James et al. (2013), and for more detail on XGBoost, see Chen and Guestrin (2016) and Jain (2022).

Table 4.2: Predictor variables in competing models for predicting the point values of the current and ensuing possessions.

Time remaining $(3600 - t)$	Offensive score	Defensive score
Down	Distance	Yards to end zone
Number of possessions played by time t	Offensive pace	Defensive pace

4.4.1 Random Forest Model

The random forest model is built on many independent decision trees. A decision tree is a machine learning algorithm that takes the data as a whole, finds the "best" predictor/value of predictor pair to split the data into two groups, or child nodes. What is "best" is determined by the predictive power, that is, the one that would minimize root mean square error (RMSE) or mean absolute error (MAE) of the model, using the average response in each node as the prediction for the observations that fall in that node. Each of these child nodes are then split again by the same process, with splitting continuing to occur until some stopping criteria is reached. This stopping criteria could be the maximum number of splits in a tree, the maximum depth of the tree, or the maximum number of terminal nodes.

To build a random forest, a sample with replacement is taken from the data, typically the same size as the data. This sample is used to build a decision tree, then another sample, independent from the previous, is taken and another tree is built. This is repeated many times to gather many independent trees. To make a prediction, an observation is inserted into each tree, with the prediction being the

average response from its terminal node. These predictions from each of the trees are then averaged to obtain the overall prediction for the random forest.

Machine learning models, such as random forests, do not require assumptions such as statistical models do. Instead, they have parameters that require "tuning." For example, the maximum number of splits in a tree is a parameter that could be tuned for each random forest model. The process would be to run the model on several different values for the maximum number of splits, then the number that minimizes the MAE of the test data would be chosen for this model. The fewer number of splits in the tree, the less fit each tree becomes to the training data, causing a greater risk of underfitting the data. Likewise, the larger number of splits in each tree, the greater the risk of overfitting the training data. Parameters for random forests can be split into two categories, tree-based parameters and model-based parameters. Tree-based parameters are applied to each tree individually and include, but are not limited to maximum number of splits, minimum number of observations in a terminal node, and maximum parameters considered at each split. Model-based parameters are values pertaining to the random forest as a whole and include, but are not limited to number of trees in the model and probabilities attached to each observation of being included in each sample for building the tree.

4.4.2 XGBoost Model

An extreme gradient boosting model is similar to a random forest in that it is a machine learning algorithm that is built on numerous decision trees. The difference lies in that instead of the trees being independent from another, each tree is built one at a time on the errors of the previous tree in XGBoost. Each tree is also weighted down by a shrinkage factor, η, $0 < \eta < 1$, to help prevent the tree from learning too quickly and over-fitting the training data. In the training set, denote $\mathbf{x}$ as the collection of all predictor variables in the training set and x_i as the vector

of predictor values of observation i, y_i be the response variable value, and $\hat{f}(x_i)$ be predicted value from the model $\hat{f}(\mathbf{x})$ at x_i. The XGBoost algorithm can be set up as follows:

(1) Initialize $\hat{f}^0(x_i) \equiv 0$ and $r_i^0 = y_i$ for all i in the training set, where r_i^0 is the residuals for each observation before any iterations in the algorithm.

(2) For $b = 1, 2, \ldots, B$, where B is the total number of trees in the algorithm, repeat the following:

   (a) Fit a tree $\hat{f}^b(\mathbf{x})$ to r^{b-1}, the residuals from the most recently updated model.

   (b) Update $\hat{f}^b(\mathbf{x})$ by adding a shrunken version of the new tree,

$$\hat{f}(\mathbf{x}) \leftarrow \hat{f}(\mathbf{x}) + \eta \hat{f}^b(\mathbf{x}).$$

   (c) For each i, update the residuals,

$$r_i^b \leftarrow r_i^{b-1} - \eta \hat{f}^b(x_i).$$

(3) Output the boosted model,

$$\hat{f}(\mathbf{x}) = \sum_{b=1}^{B} \eta \hat{f}^b(\mathbf{x}).$$

Choice of B is critical, since for larger values of B, that is, a model with more trees, η must be closer to zero to prevent over-fitting the training data. If η is large, say close to 1, then the first tree built will start to fit the training data well, and trees later on in the algorithm will be fitting residuals that are more likely due to noise than to an actual signal.

Tuning of the XGBoost model can often be time consuming due to the number of parameters to tune around and the amount of time it can take to run each iteration of the model for large data sets. XGBoost has the same tree-based parameters that are present in random forest models, most notably the depth of the tree, because the

depth of each tree is the largest possible interaction depth the model can detect. The additional model-based parameters that need tuning are B and η which are tuned simultaneously. There are many methods for tuning parameters of the XGBoost model, one of which is performed using the following steps.

(1) Choose a relatively large η, somewhere between 0.05 and 0.2.

(2) Optimize B for this shrinkage value, keeping B to a value where a machine can run the model relatively quickly.

(3) Tune tree-based parameters using the values of η and B obtained in the first two steps.

(4) Decrease η and increase B proportionally until the model's performance improves minimally. Measures of the model's performance include statistics like MAE or RMSE.

4.4.3 Test and Training Data for Models

To fit the point value model, and later the win-probability model, the data was separated into a test data set and two training data sets. In particular, the last five of the 17 seasons were designated as the test data for the win probability model described in Section 4.5 and evaluated in Section 4.6. For the remaining 12 seasons, half of the data were randomly selected to act as the overall data for the point value model. This data will be referred to as the "point value data." The other half were used to build the win probability model.

4.4.4 Point Value Model Selection

To determine which of the four point value models is best, the point value data was further separated into test and training data sets. Each of the four point value models were built on the training set that was randomly pulled from the point

value data and then applied to the point value test data. The models were compared using mean absolute error (MAE). The results are shown in Table 4.3. For both linear regression models, the average difference between the predicted point difference for the current and following possession and the actual point difference was slightly more than three points per play. The random forest model slightly outperforms the regression models, lowering the MAE to slightly below three points. The XGBoost model performs the best, with the lowest MAE by a margin at least 0.2949 points. Therefore, the XGBoost model is adopted to predict the point differential for the current and subsequent possessions. Conclusions about model performance are similar based on root mean square error.

Table 4.3: Performance of point value models using test MAE.

Model	MAE
Linear regression	3.0805
Regression w/ inter	3.0614
Random forest	2.9751
XGBoost	2.6802

4.5 Win Probability Model

For a specific game, consider the random process that is the expected lead for the home team following the current and succeeding possession, which will now be referred to as expected score, ω. The expected score will shift as time in the game moves forward, or equivalently as the expected possessions remaining in the game, τ, decreases, where τ is calculated as in (4.1). For specific values of τ and ω, letD $p_{\tau,\omega}$ represent the in-game probability that the home team will win the game. When considering multiple games $i = 1, 2, \ldots, M$, let $Y_i = 1$ if the home team wins game i and 0 otherwise. Consider $p_{\tau,\omega}$ as a continuous function of τ and ω. Maddox

et al. (2022a) introduce a Bayesian estimator of in-game win probability, $p_{t,\ell}$, based on current time t and score differential ℓ, in college basketball. Their methods are extended and adapted to the NBA in Maddox et al. (2022b). In both papers, Maddox et al. argue that the nature of a basketball game makes time and score differential excellent predictors for in-game win probability. For college football, instead of using time and score differential as predictors, τ and ω are preferred for two reasons. First, and most obviously, the nature of the games of basketball and football are inherently different. As opposed to basketball, the predictors τ and ω contain more information about current the state of the game, and account for which team has the ball, how physically close (on the field) that team is to scoring, etc. We now extend the methods of Maddox et al. (2022b) to estimate $p_{\tau,\omega}$ for college football.

4.5.1 Naive Estimator of In-game Win Probability

For each combination of τ and ω rounded to the nearest whole number, consider the (τ, ω) "cell." On a specific cell, the number of wins by the home team, $n_{\tau,\omega}$, follows a binomial$(N_{\tau,\omega}, p_{\tau,\omega})$ distribution, where $N_{\tau,\omega}$ is the total number of games that are observed in the (τ, ω) cell. Due to the total number of possible cells, one for each combination of τ and ω, some cells may have a small value of $N_{\tau,\omega}$, resulting in a large standard error for any estimator for $p_{\tau,\omega}$. Deshpande and Jensen (2016) and Maddox et al. (2022a,b) suggest a binning approach to address the small sample sizes. Windows centered around (τ, ω) can be created in such a way that the in-game win probability remains relatively constant across the window. For creating the interval around expected score, since scores were considered for each two possessions, a shift of two possessions will rarely have a major affect on win probability. Moreover within college football for similar score differentials, two or fewer possessions should not have a large effect on the win probability, especially early

in the game. Therefore a reasonable interval for expected possessions remaining is defined as $[\tau - 2, \tau + 2]$. To determine an interval for the expected score, note that when an offense scores, the fewest points attainable is three points from a field goal. Therefore, the interval on expected score is taken to be $[\omega - 3, \omega + 3]$. The same notation will be adopted for any $[\tau - 2, \tau + 2] \times [\omega - 3, \omega + 3]$ window; that is, $N_{\tau,\omega}$ is the number of games in the window in which the home team has an expected lead by any value in $[\omega - 2, \omega + 2]$ points with any expected possessions remaining in $[\tau - 2, \tau + 2]$, and given the specific value of $N_{\tau,\omega}$, $n_{\tau,\omega} = \sum_{i=1}^{N_{\tau,\omega}} Y_i$, distributed as a binomial$(N_{\tau,\omega}, p_{\tau,\omega})$ random variable. Based on the binomial distribution, a simple estimator for in-game home team win probability for for each (τ, ω) window is the maximum likelihood estimator

$$\bar{p}_{\tau,\omega} = \frac{n_{\tau,\omega}}{N_{\tau,\omega}}. \tag{4.2}$$

As a given game approaches the end of regulation, each individual point and possession will have a larger impact on the in-game win probability. Therefore the windows should be modified at the end of the game to reflect this. Starting from fifteen expected possessions remaining, the proposed method shortens window lengths and widths; that is, the length of interval around each of expected score and expected number of possessions remaining will gradually decrease until there is an expected two possesses remaining, when the intervals' widths become zero.

4.5.2 Dynamic Bayesian Estimator

To elicit a prior distribution for in-game win probability, Maddox et al. (2022b) suggest polling a sample of industry experts. The same can be done for college football. The authors contacted a panel of 14 college football experts, including coaches from a major Division I college football team and respected media personnel. Each provided their estimate of the probability a team wins for each combination of score differential, ℓ, and time elapsed, t, in Table 4.4, regardless of which team

is the home team. Note that t and ℓ are the actual time and score differential in the hypothetical game and are not the same as the previously defined τ and ω. To reduce the level of complexity and increase the likelihood of response, the authors choose to ask the question of win probability in terms of t and ℓ as opposed to τ and ω.

The Bayesian prior parameters have an interesting interpretation, first noted by Deshpande and Jensen (2016). The parameter $\alpha_{t,\ell}$ can be interpreted as the number of "pseudo-wins" in the (t, ℓ) cell; likewise $\beta_{t,\ell}$ as the number of "pseudo-losses." Through this interpretation, the two parameters can be seen as a way of increasing the number of games in a specific (t, ℓ) cell. If the home team is ahead, then first scale parameter being large effectively acts to increase the number of wins in that cell. On the other hand, if the home team is behind, the second scale parameter is large, and acts to increase the number of losses.

The sample mean $\tilde{p}_{t,\ell}$ and sample variance $s^2_{t,\ell}$, of the probabilities were computed. The two scale parameters were estimated via a method-of-moments type approach. The system of equations

$$\begin{aligned}\tilde{p}_{t,\ell} &= \frac{\alpha_{t,\ell}}{\alpha_{t,\ell} + \beta_{t,\ell}},\\ s^2_{t,\ell} &= \frac{\alpha_{t,\ell}\beta_{t,\ell}}{\left(\alpha_{t,\ell} + \beta_{t,\ell}\right)^2 \left(\alpha_{t,\ell} + \beta_{t,\ell} + 1\right)},\end{aligned}$$

is solved for $\alpha_{t,\ell}$ and $\beta_{t,\ell}$, yielding

$$\begin{aligned}\alpha_{t,\ell} &= -\frac{\tilde{p}_{t,\ell}\left(\tilde{p}^2_{t,\ell} - \tilde{p}_{t,\ell} + s^2_{t,\ell}\right)}{s^2_{t,\ell}}\\ \beta_{t,\ell} &= \frac{\left(\tilde{p}_{t,\ell} - 1\right)\left(\tilde{p}^2_{t,\ell} - \tilde{p}_{t,\ell} + s^2_{t,\ell}\right)}{s^2_{t,\ell}},\end{aligned}$$

as long as $\left(\tilde{p}_{t,\ell} - 1\right)\tilde{p}_{t,\ell}\left(\tilde{p}^2_{t,\ell} - \tilde{p}_{t,\ell} + s^2_{t,\ell}\right) \neq 0$.

In Table 4.4 the score differential is presented when the home team has the lead. The larger the lead, the more the prior is left-skewed. On the other hand, if the visiting team has the lead, then the roles of $\alpha_{t,\ell}$ and $\beta_{t,\ell}$ are reversed, causing the

prior to be right-skewed. At any time, if the game is tied ($\ell = 0$), or is sufficiently close in score for the amount of time remaining, the prior distribution is a diffuse beta$(1, 1)$ prior, allowing the likelihood to drive the results in the posterior.

Table 4.4: Imputed parameters for beta prior.

Elapsed Time (t) (sec.)	Home Team Lead (ℓ)	$\alpha_{t,\ell}$	$\beta_{t,\ell}$
$[0, 900]$	$[0, 7]$	1	1
$[0, 900]$	$(7, 14]$	21	10
$[0, 900]$	$(14, 28]$	16	3
$[0, 900]$	$(28, \infty)$	59	1
$(900, 1800]$	$[0, 7]$	1	1
$(900, 1800]$	$(7, 14]$	15	7
$(900, 1800]$	$(14, 28]$	11	2
$(900, 1800]$	$(28, \infty)$	44	1
$(1800, 2700]$	$[0, 7]$	1	1
$(1800, 2700]$	$(7, 14]$	14	4
$(1800, 2700]$	$(14, 28]$	13	1
$(1800, 2700]$	$(28, \infty)$	126	1
$(2700, 3300]$	$[0, 3]$	1	1
$(2700, 3300]$	$(3, 7]$	28	16
$(2700, 3300]$	$(7, 14]$	27	8
$(2700, 3300]$	$(14, 21]$	11	1
$(2700, 3300]$	$(21, \infty)$	98	1
$(3300, 3480]$	0	1	1
$(3300, 3480]$	$(0, 3]$	29	18
$(3300, 3480]$	$(3, 7]$	17	6

Elapsed Time (t) (sec.)	Home Team Lead (ℓ)	$\alpha_{t,\ell}$	$\beta_{t,\ell}$
$(3300, 3480]$	$(7, 10]$	13	2
$(3300, 3480]$	$(10, 14]$	16	1
$(3300, 3480]$	$(14, \infty)$	98	1
$(3480, 3600)$	0	1	1
$(3480, 3600)$	$(0, 3]$	21	10
$(3480, 3600)$	$(3, 7]$	18	5
$(3480, 3600)$	$(7, 10]$	16	1
$(3480, 3600)$	$(10, \infty)$	91	1

The maximum likelihood estimator in (4.2) is based on τ and ω. However, the elicited prior is based on t and ℓ. Consequently, the in-game win probability will be written $p_{t,\ell,\tau,\omega}$. Since the beta family of distributions is a conjugate prior for the binomial distribution, the beta-binomial connection is used to estimate $p_{t,\ell,\tau,\omega}$ with the mean of the posterior beta distribution, specifically

$$\hat{p}_{t,\ell,\tau,\omega} = \frac{n_{\tau,\omega} + \alpha_{t,\ell}}{N_{\tau,\omega} + \alpha_{t,\ell} + \beta_{t,\ell}}. \tag{4.3}$$

4.5.3 Adjusted Dynamic Bayesian Estimator

During a game, the in-game win probability is clearly a function of expected possessions remaining and expected score. However, it is also affected by the overall skill of the teams playing the game. The skill level can be incorporated using pregame win probabilities for each game. The normal distribution quantile function was used

to convert the pregame point spread to a pregame home team win probability $\hat{p}_p$ for each game using the method outlined by Maddox et al. (2022a).

Maddox et al. (2022b) introduce three different functions to incorporate pregame probabilities. The first is a linear function of only time remaining. The second is a linear function of time remaining and score differential. The third is linear in time, but quadratic in score differential. Other variations to the three weight functions were considered. For example, including a quadratic term for time was attempted. However, the more complicated models would not converge appropriately for R to optimize the performance accurately. The three weight functions specifically considered here are

$$\begin{aligned} D_1 &= bt, \\ D_2 &= c_0 + c_1 t + c_2 |\ell|, \\ D_3 &= d_0 + d_1 t + d_2 |\ell| + d_3 \ell^2. \end{aligned}$$

Each of the weight functions yields a competing model for including pregame win probabilities, which can be represented as

$$p^*_{t,\ell,\tau,\omega,j} = \begin{cases} \hat{p}_p, & D_j \leq 0 \\ (1 - D_j)\,\hat{p}_p + D_j \hat{p}_{t,\ell,\tau,\omega}, & 0 < D_j < 1\,, \\ \hat{p}_{t,\ell,\tau,\omega}, & D_j \geq 1 \end{cases} \qquad j = 1, 2, 3,$$

where $p^*_{t,\ell,\tau,\omega,j}$ is the final predicted win probability associated with weight function D_j.

Values for $b, c_0, \ldots, d_3$ are estimated by minimizing Brier score for each $p^*_{t,\ell,\tau,\omega,j}$ computed from the holdout test data. The Brier scores for each $p^*_{t,\ell,\tau,\omega,j}$ are shown in Table 4.5. The model $p^*_{t,\ell,\tau,\omega,2}$ is the most accurate predictor for the test data. The fitted model for D_2 that provided the minimal Brier score is

$$D_2 = 0.09589 + 0.00018t + 0.02523\ell.$$

Using this expression for B_2, the model

$$p^*_{t,\ell,\tau,\omega,2} = \begin{cases} \hat{p}_p, & D_2 \leq 0 \\ (1 - D_2)\,\hat{p}_p + D_2 \hat{p}_{t,\ell,\tau,\omega}, & 0 < D_2 < 1 \\ \hat{p}_{t,\ell,\tau,\omega}, & D_2 \geq 1 \end{cases}$$

is the "adjusted dynamic Bayesian estimator." The percent of in-game win probability accounted for in the final predicted win probability is displayed in Figure **??**. As desired, early on in any game the majority of the final predicted win probability comes from the pregame win probability. However, as the game draws to an end or one team takes a large lead, the final predicted win probability gets closer and closer to the dynamic Bayesian estimator.

Table 4.5: Brier scores for models determining in-game probability proportion.

Proportion Model	Brier Score
Linear Time (D_1)	0.1272
Linear Time & Score (D_2)	0.1250
Quadratic (D_3)	0.1265

4.6 Model Evaluation and Application

For each play from each game from the 2017 through 2021 seasons, excluding the 2020 season, Brier score was computed. The Brier scores of the three models, dynamic Bayesian estimator, adjusted dynamic Bayesian estimator, and the random forest model proposed by Lock and Nettleton (2014) are shown in Table 4.6. Brier's score is a statistic used to compare the performance of different methods for estimating probabilities. Brier's score is the average of the square of the difference between the estimated probability and the observed binary outcome. In the context

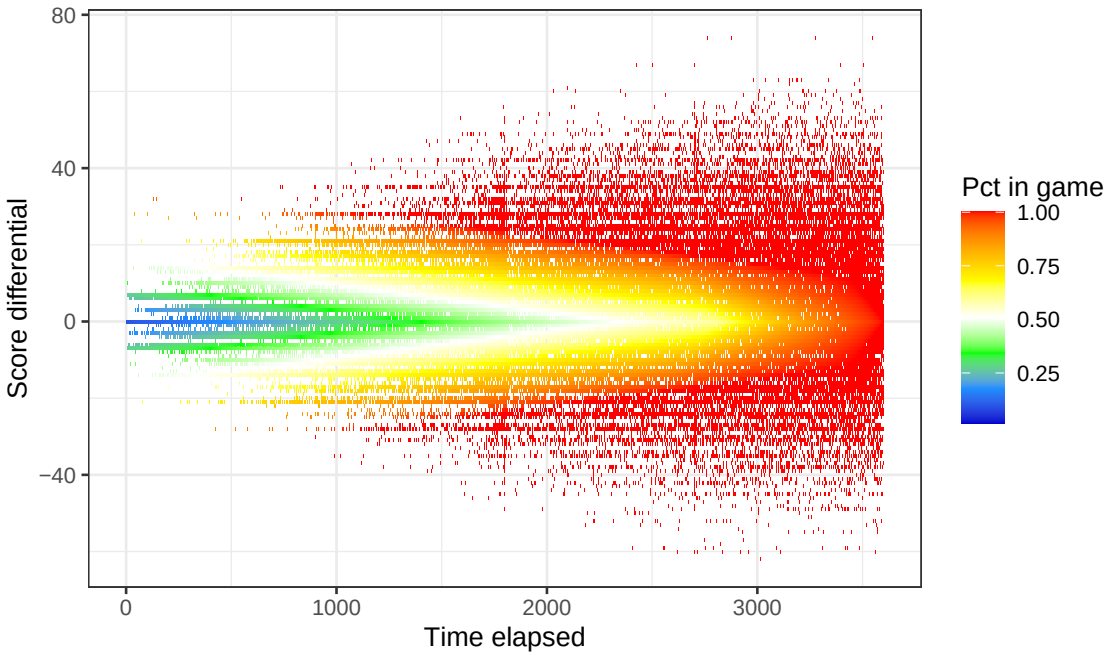

Figure 4.1: Graphical function of D_2.

of in-game home team win probabilities, this observed binary is Y_i as defined in Section 4.5. To interpret Brier's score, if $Y_i = 1$ for all i, and the predicted probability is also one for every i, then Brier's score will be zero, indicating perfect prediction. On the other hand, if for all $i, Y_i = 0$, and the estimated probability is one, then Brier's score will be one, the worst possible Brier's score. Both Bayesian models outperform the random forest model. The dynamic Bayesian model with the pregame adjustment is the model that performs best overall.

Table 4.6: Brier scores for predictive performances for 2017 through 2021 seasons.

Model	Brier Score
Dynamic Bayes	0.1453
Adjusted dynamic Bayes	0.1250
Random forest	0.1705

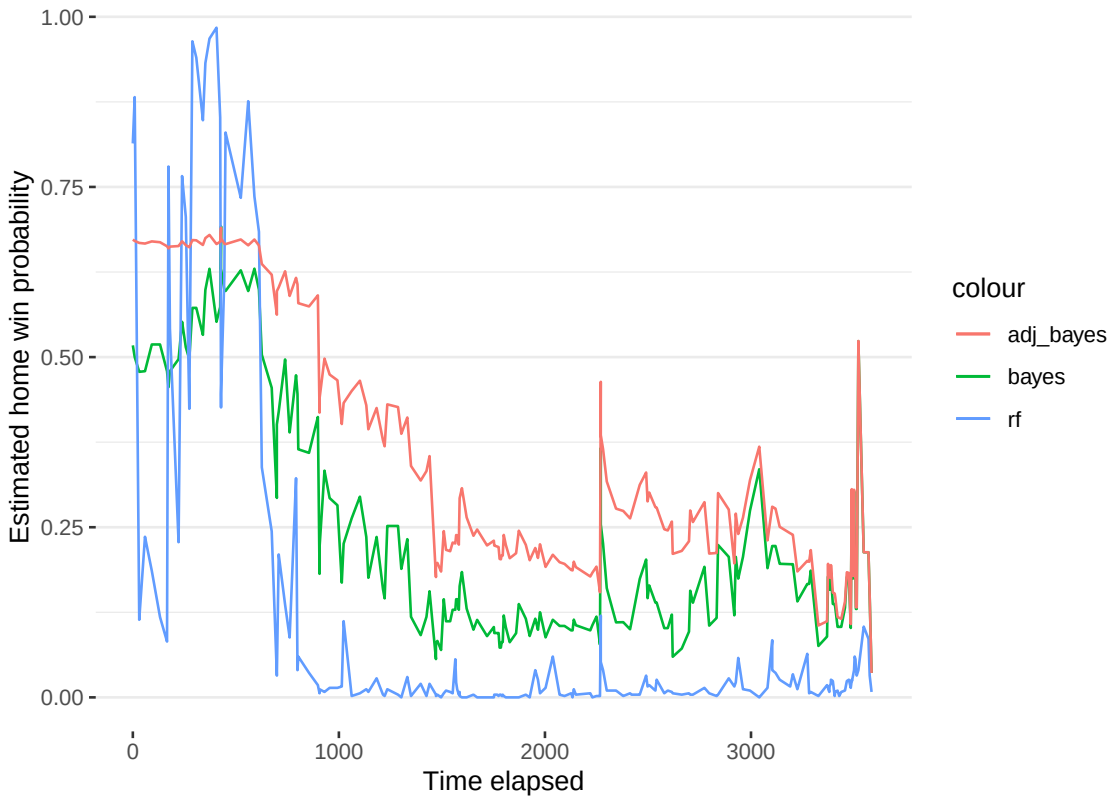

Figure 4.2: In game win probability for 2021 Big 12 Championship.

To observe the performance of each model further, each model can be applied to a specific game, observing the features of the models as the game progresses. The models are run for the 2021 Big 12 Championship game between Baylor University and Oklahoma State University. The results can be seen in Figure 4.2.

On December 4, 2021, the Baylor Bears and Oklahoma State Cowboys met at AT&T Stadium in Dallas, Texas for the Big 12 Championship game. Oklahoma State was the home team. During the regular season, Oklahoma State had beaten Baylor, and were considered the favorite to win this game. The probability traces for the game for each of the three models is seen in Figure 4.2. The adjusted dynamic Bayes model (red curve) shows that Oklahoma State is favored by starting at well over 50% chance that Oklahoma State would win. Early in the game, one

of the features that leads to the Bayesian models outperforming the random forest (blue line) is can be seen. The random forest model is too quick to jump to a win probability close to 0 or 1, especially early in the game. The game got off to a relatively slow start, but by halftime Baylor had surprised many by jumping out to a 21-6 lead. At that time, all three models predicted Baylor was most likely to win. In the second half, Oklahoma State started to make a come back. Halfway through the third quarter, Oklahoma State was able to score a touchdown and make the score 21-13, significantly raising their probability of winning. The Cowboys then had another significant uptick in their win probability with ten minutes to go when they earned a first-and-goal from the Baylor one-yard line, appearing to be on the verge of scoring a touchdown with the potential to tie the game. However, Baylor was able to hold Oklahoma State to just a field goal, maintaining their lead and their edge in win probability. The climax of the game came with three minutes left when Oklahoma State started a drive from their own 10 yard line. They methodically drove the ball down the field and once again ended up with a first-and-goal, this time from the two-yard line with 80 seconds left. Being down by 5 points, they were not going to kick a field goal. Instead, they had four attempts to score a touchdown to win the game. With so many attempts from so close, it appeared likely that Oklahoma State would be able to score, driving their win probability in the two Bayesian models over 50% despite trailing. Once again the Baylor defense made an impressive goal line stand, keeping Oklahoma State out of the end zone, ultimately less than half of a yard short, sealing the victory for Baylor.

4.7 Conclusion

Two new methods are proposed for estimating in-game win probability for college football games. Both are an extension and enhancement of the methods in Maddox et al. (2022a,b), which provide a number of models for estimating in-

game home-team win probabilities for NCAA basketball and NBA basketball respectively. The first proposed "dynamic Bayesian estimator," uses expected score differential and expected possessions remaining as predictors, which are modeled from the data themselves, and a prior that has been calculated based on the distribution of predicted win probabilities from 14 college football field experts. The second method, referred to as the "adjusted dynamic Bayesian estimator," adjusts the dynamic Bayesian estimator based on pregame win probabilities obtained from TeamRankings.com. The adjustment is optimized over a function of both time and score so that as the game moves on or the score differential increases, the adjusted dynamic Bayesian estimator will begin to approach the dynamic Bayesian estimator rather than the pregame win probability. These two methods are then compared to the random forest win probability model proposed by Lock and Nettleton (2014). Both new models outperform the standard random forest model, with the adjusted dynamic Bayesian estimator performing the best out of all of the models.

CHAPTER FIVE

Summary and Conclusions

5.1 College Basketball Conclusion

For NCAA basketball games, two new methods are proposed for estimating or predicting in-game home team win probabilities. The first newly proposed method is a Bayesian estimator with a prior distribution that changes as a function of lead differential and time elapsed, which was called the Bayesian estimator with a dynamic prior. The second method adds to the original estimate a time-weighted adjustment based on pregame win probability computed from daily ratings. In this paper, the adjustment was applied only to the Bayesian estimate with dynamic prior. It is reasonable to conclude the adjustment would improve the performance of the other estimators, just as it did the dynamic Bayesian estimator. A comparison of the methods for the purpose of estimation shows that the two proposed estimates outperforms the estimates from the three standard methods, and is competitive with the logistic model of Benz. For prediction, the adjusted dynamic Bayesian method out performs the other, based on a comparison of both Brier Score and misclassification rates. There are a number of additional problems to be investigated. First note the methodology can be easily restructured based on the total time of the game to apply to other basketball leagues, most notably the National Basketball Association (NBA). Also of interest is the effect on the estimators for different window choices. Another consideration is that the adjustment resulted in a substantial improvement on the prediction of the adjusted dynamic Bayesian estimator. The adjusted dynamic Bayes model performed almost as well as Benz's model. Another area of investigation that remains is to determine a function of time that gives pregame win

probability a more quickly decreasing role than the linear function considered here. It is also likely that other pregame metrics, or some combination of pregame metrics, in place of the pregame win probability derived from the power rankings (found on `teamrankings.com`) might improve the outcome. These other metrics include different power rankings, ELO rating, score-differential, and other team statistics; any of these can be used singularly, as in this paper, or combined. Finally, the development of these types of models for other sports also presents unique challenges that are worth investigating.

5.2 NBA Conclusion

Two new methods are proposed for estimating in-game win probability for NBA games. Both are an extension and enhancement of the methods in Maddox et al. (2022a), which provide a number of models for estimating in-game home-team win probabilities for NCAA basketball. The first proposed "dynamic Bayesian estimator," uses a prior that has been calculated based on the distribution of predicted win probabilities from 14 NBA field experts, including several anonymous front office associates within the NBA. The second method, referred to as the "adjusted dynamic Bayesian estimator," adjusts the dynamic Bayesian estimator based on pregame win probabilities obtained from TeamRankings.com. The adjustment is optimized over a function of both time and score so that as the game moves on or the score differential increases, the adjusted dynamic Bayesian estimator will begin to approach the dynamic Bayesian estimator rather than the pregame win probability. These two methods are then compared to the win probability model that ESPN uses for seasons 2018-19 and 2019-20. The ESPN model performs the best overall, but there are times during the game the adjusted dynamic Bayesian model performs the best, indicating that there are some features of the model that estimate probabilities

well and some that could be improved upon in the future, such as the calculation of pregame win probability.

5.3 CFB Conclusion

Two new methods are proposed for estimating in-game win probability for college football games. Both are an extension and enhancement of the methods in Maddox et al. (2022a) and Maddox et al. (2022b), which provide a number of models for estimating in-game home-team win probabilities for NCAA basketball and NBA basketball respectively. The first proposed "dynamic Bayesian estimator," uses expected score differential and expected possessions remaining as predictors, which are modeled from the data themselves, and a prior that has been calculated based on the distribution of predicted win probabilities from 14 college football field experts. The second method, referred to as the "adjusted dynamic Bayesian estimator," adjusts the dynamic Bayesian estimator based on pregame win probabilities obtained from TeamRankings.com. The adjustment is optimized over a function of both time and score so that as the game moves on or the score differential increases, the adjusted dynamic Bayesian estimator will begin to approach the dynamic Bayesian estimator rather than the pregame win probability. These two methods are then compared to the random forest win probability model proposed by Lock and Nettleton. Both new models outperform the standard random forest model, with the adjusted dynamic Bayesian estimator performing the best out of all of the models.

BIBLIOGRAPHY

Tom Adams. *Improving your NCAA Bracket with Statistics*. CRC Press, Boca Raton, 2019.

Mark Bashuk. Using cumulative win probabilities to predict NCAA basketball performance. In *MIT Sloan Sports Analytics Conference*, February 2012. URL http://www.sloansportsconference.com/wp-content/uploads/2012/02/Using-Cumulative-Win-Probabilities-to-Predict-NCAA-Performance-Bashuk.pdf.

Luke Benz. A new ncaahoopr win probability model, December 2019. URL https://lukebenz.com/post/ncaahoopr_win_prob/. [Online; posted December 26, 2019].

Tao Chen and Qingliang Fan. A functional data approach to model score difference process in professional basketball games. *Journal of Applied Statistics*, 45:112–127, 2018. doi: https://doi.org/10.1080/02664763.2016.1268106.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system, Aug 2016. URL https://dl.acm.org/doi/pdf/10.1145/2939672.2939785.

Harris Cooper, Kristina M. DeNeve, and Frederick Mosteller. Predicting professional game outcomes from intermediate game scores. *CHANCE*, 5(3-4):18–22, 1992. doi: 10.1080/09332480.1992.10554981.

Sameer K. Deshpande and Shane T. Jensen. Estimating an NBA player's impact on his team's chances of winning. *Journal of Quantitative Analysis in Sports*, 12 (2):51–72, January 2016. doi: 10.1515/jqas-2015-0027.

ESPN/Associated Press. Drexel makes history with 34-point rally to beat delaware. *ESPN*, February 2018.

Aarshay Jain. Complete guide to parameter tuning in XGBoost, Jun 2022. URL https://www.analyticsvidhya.com/blog/2016/03/complete-guide-parameter-tuning-xgboost-with-codes-python/.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer, 2013. URL https://faculty.marshall.usc.edu/gareth-james/ISL/.

Justin Kubatko, Dean Oliver, Kevin Pelton, and Dan T Rosenbaum. A starting point for analyzing basketball statistics. *Journal of Quantitative Analysis in Sports*, 3(3), 2007. doi: doi:10.2202/1559-0410.1070.

George R. Lindsey. An investigation of strategies in baseball. *Operations Research*, 11(4):477–501, 1963. doi: 10.1287/opre.11.4.477.

Dennis Lock and Dan Nettleton. Using random forests to estimate win probability before each play of an NFL game. *Journal of Quantitative Analysis in Sports*, 10(2):197–205, 2014. doi: 10.1515/jqas-2013-0100.

Jason T. Maddox, Ryan Sides, and Jane L. Harvill. Bayesian estimation of in-game home team win probability for college basketball. *Journal of Quantitative Analysis in Sports*, 2022a.

Jason T. Maddox, Ryan Sides, and Jane L. Harvill. Bayesian estimation of in-game home team win probability for national basketball association games. *Journal of Sport Analytics*, 2022b.

Ken Pomeroy. Ratings glossary: The kenpom.com blog, Aug 2012. URL https://kenpom.com/blog/ratings-glossary/.

Pro Football Reference. The p-f-r win probability model. https://www.sports-reference.com/blog/2012/03/features-expected-points/, Mar 2012.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL https://www.R-project.org/.

John Ruscio and Kevin Brady. Estimating win probability for nfl games, Jan 2021. URL https://ruscio.pages.tcnj.edu/files/2021/01/NFL-Win-Probability.pdf.

Richard Ryall. *Predicting Outcomes in Australian Rules Football*. PhD thesis, Royal Melbourne Institute of Technology University, 2011.

Edgar Santos-Fernandez, Paul Wu, and Kerrie L. Mengersen. Bayesian statistics meets sports: a comprehensive review. *Journal of Quantitative Analysis in Sports*, 15(4):289–312, 2019.

Jian Shi and Kai Song. A discrete-time and finite-state markov chain based in-play prediction model for NBA basketball matches. *Communications in Statistics: Simulation and Computation*, 2019. doi: 10.1080/03610918.2019.1633351.

Kenny Shirley. Markov model for basketball. In *Proceedings of the New England Symposium for Statistics in Sports*, Boston, MA, 2007.

Kai Song and Jian Shi. A gamma process based in-play prediction model for National Basketball Association games. *European Journal of Operational Research*, 283: 706–713, 2020. doi: https://doi.org/10.1016/j.ejor.2019.11.012.

Kai Song, Yiran Gao, and Jian Shi. Making real-time predictions for NBA basketball games by combining the historical data and bookmaker's betting line. *Physica A*, 547:1–8, 2020.

Hal S. Stern. A Brownian motion model for the progress of sports scores. *Journal of the American Statistical Association*, 89(427):1128–1134, 1994. doi: 10.1080/01621459.1994.10476851.

Erik Štrumbelj and Petar Vračar. Simulating a basketball match with a homogeneous markov model and forecasting the outcome. *International Journal of Forecasting*, 28:532–542, 2012. doi: https://doi.org/10.1016/j.ijforecast.2011.01.004.

Zachary Terner and Alexander Franks. Modeling player and team performance in basketball. *Annual Review of Statistics and Its Applications*, 8(1):1–23, 2021. doi: https://doi.org/10.1146/annurev-statistics-040720-015536.

Jake Troch. A new look at college football tempo, Nov 2016. URL https://www.footballstudyhall.com/2016/11/7/13549506/college-football-tempo-pace-increasing.

Petar Vračar, Erik Štrumbelj, and Igor Kononenko. Modeling basketball play-by-play data. *Expert Systems with Applications*, 44:58–66, 2016. doi: https://doi.org/10.1016/j.eswa.2015.09.004.

Peter H Westfall. Graphical presentation of a basketball game. *The American Statistician*, 44(4):305–307, 1990.

Hadley Wickham. *rvest: Easily Harvest (Scrape) Web Pages*, 2021. URL https://CRAN.R-project.org/package=rvest. R package version 1.0.2.