2016

Statistical methods in sports with a focus on win probability and performance evaluation

Dennis Lock
Iowa State University

Statistical methods in sports with a focus on win probability and performance evaluation

by

Dennis Lock

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Statistics

Program of Study Committee:

Dan Nettleton, Major Professor

Jarad Niemi

Philip M. Dixon

Kenneth J. Koehler

Stephen Vardeman

Iowa State University

Ames, Iowa

2016

TABLE OF CONTENTS

LIST OF TABLES iv

LIST OF FIGURES v

ACKNOWLEDGEMENTS viii

ABSTRACT ix

CHAPTER 1. INTRODUCTION 1

CHAPTER 2. USING RANDOM FORESTS TO ESTIMATE WIN PROBABILITY BEFORE EACH PLAY OF AN NFL GAME 3

2.1 Introduction 3

2.2 Training Data 6

2.3 Random Forest Method 7

2.4 Win Probability Prediction Accuracy 9

2.5 Assessing Variable Importance and Relationships with Win Probability 10

2.6 Examining Changes in Win Probability 13

2.7 Win Probability Analysis of Coaching Decisions 14

2.8 Discussion 16

CHAPTER 3. USING RANDOM FORESTS TO ESTIMATE IN-GAME WIN PROBABILITY 21

3.1 Introduction 21

3.2 Estimating Win Probability Values 23

3.2.1 Datasets 25

3.2.2 Implementation 26

3.2.3 Performance 27

3.3 Using Win Probability Values 28

3.3.1 Wins Added 29

3.4 Discussion 34

CHAPTER 4. USING INFORMATION UNDERLYING MISSING DATA TO IMPROVE ESTIMATION OF NFL FIELD GOAL KICKER ACCURACY 39

4.1 Introduction 39

4.2 A Model for Informative Missingness in Response Values 42

4.3 The Dataset and Its Connection to the IM Model 43

4.4 Results 46

4.4.1 Performance Comparison 47

4.4.2 Kicker Evaluation 51

4.5 Discussion 53

CHAPTER 5. SUMMARY AND DISCUSSION 60

LIST OF TABLES

Table 2.1 Description of predictor variables 6

Table 2.2 Test set MSE by quarter 9

Table 2.3 Variable importance (overall and by quarter) 12

Table 3.1 Top 5 NBA players from 2009 - 2012 by Wins Added. 31

Table 3.2 Lebron James $\hat{\beta}_j$ and WA in Cleveland for the 2009/2010 season and in Miami for the 2010/2011 season. 32

Table 3.3 Median and 95% credible intervals for α_0 and α_1 32

Table 4.1 Posterior means and 95% credible intervals for the β parameters in (4.2) and the α parameters in (4.5). 47

Table 4.2 Posterior standard deviations for intercept and distance coefficients (averaged over kickers) by method. 49

Table 4.3 Test set mean absolute error (MAE) and mean squared error (MSE) for each method. 49

Table 4.4 Top 5 NFL kickers by adjusted FG%, with their raw FG% included. 52

Table 4.5 Comparing the success probability for S. Janikowski and G. Hartley at a distance of 25 yards and 50 yards on a grass surface with no precipitation, and comparing the two on overall adjusted FG% (posterior means with 95% credible intervals in parentheses). 52

LIST OF FIGURES

Figure 2.1 Estimated Baltimore Ravens win probability from every play of Superbowl 47. Highlighted plays with score and Ravens win probability: 1 = Start of game (0-0, $WP = 0.382$), 14 = Ravens score first (7-0, $WP = 0.552$), 57 = Ravens intercept the ball on the first play following their second touchdown (14-3, $WP = 0.743$), 84 = Ravens Open half with 109 Yard kickoff return (28-3, $WP = 0.935$), 115 = 49ers recover a fumble after back to back touchdowns (28-20, $WP = 0.572$), 159 = 49ers first and goal with 2 minutes remaining (28-23, $WP = 0.524$). 4

Figure 2.2 Binned estimated win probability and proportion of games won in each bin (line = a perfect fit). 10

Figure 2.3 Changing one variable at a time with others held constant $Down = 1$, $YTG = 10$, $Yardline = 20$, $TOTp = 28$, $Seconds = 300$, $Score = 0$, $Spread = 0$, $TIMO = TIMD = 3$ unless otherwise specified. Variable changed in plot (a) *Score*, (b) *Seconds* (blue: $Score = -7$, Light Blue: $Score = -3$, Black: $Score = 0$, Orange: $Score = 3$, red: $Score = 7$), (c) *Spread* ($Seconds = 3600$), (d) *Down*, (e) *YTG* ($Down = 3$), and (f) *Yardline* (black: $Seconds = 1700$, red: $Seconds = 120$). Note the y-axis is focused on a narrower range in plots (d) and (e). . . 11

Figure 2.4 Variable importance of *Spread* by seconds from start (smoothed using a loess smoother). 13

Figure 2.5 Estimated win probability by play for Superbowl 44, blue vertical lines represent Indianapolis touchdowns and gold vertical lines represent New Orleans touchdowns. 14

Figure 2.6 Estimated win probability by play for Superbowl 42, blue vertical lines represent New York touchdowns and red vertical lines represent New England touchdowns. 15

Figure 2.7 Estimated win probability before (red) and after (black) a successful 27 yard field goal on fourth down and 3 when the offense is trailing by a score of 14 to 10 prior to the kick (Other variables: $TIMO = TIMD = 3$, $Spread = 0$). Dotted lines indicate the changing quarters. 16

Figure 2.8 Estimated win probability by play for Superbowl 45 with both random forest estimation (Black) and Brian Burke's estimation (Red). 17

Figure 3.1 Actual proportion of wins vs. estimated WP for the (a) NHL, (b) NFL, and (c) NBA. The diagonal line represents what would constitute ideal performance. . 28

Figure 3.2 Chicago Blackhawks win probability over the course of the game. The influential points are goals by Boston, Chicago, then Boston again, and finally the most influential point at Chicago's two goals 14 seconds apart that changed the game from 1-2 to 3-2. 29

Figure 3.3 Cleveland Cavalier win probability over the course of fourth quarter of the 2016 NBA finals game 7. The two vertical lines mark the possession where Cleveland trailed by 4 with less than 6 minutes remaining (blue), and the possession where Kyrie Irving made a 3 point shot to break the 89-89 tie (red). . 30

Figure 3.4 Comparing the β_j estimates from the model fit with a $N(0, \sigma^2/\lambda)$ (x axis) to those fit with a $N(\alpha_1 + \alpha_2(S_j), \sigma^2/\lambda)$ (y axis). 33

Figure 4.1 Comparing the distribution of attempts by attempt distance for John Carney, Garret Hartley, Rob Bironas, and Sebastian Janikowski 45

Figure 4.2 Estimates of the probability of taking the kick as a function of the probability of making the kick for kicks of varying importance, ranging from not very important (change in win probability = 0.01, Black) to very important (change in win probability = 0.50, Red). 48

Figure 4.3 Comparing the estimated probability of making an attempt as a function of kick distance with Garret Harlety (red) and Sebastian Janikowski (black) for the (a) SLR, (b) HPR, and (c) IM analyses. 51

Figure 4.4 Comparing raw FG% to adjusted FG% for all 41 qualifying kickers (dotted line indicates equality between raw and adjusted FG%). 53

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my thanks to those who helped me with various aspects of conducting research and the writing of this thesis. First and foremost, Dr. Dan Nettleton for his guidance throughout the whole process, willingness to work outside his research area, and most of all for allowing, encouraging, and assisting me in pursuing my dream despite the difficulties and work it has caused. I would also like to thank my committee members for their efforts and contributions to this work: Dr. Jarad Niemi, Dr. Philip Dixon, Dr. Ken Koehler, and Dr. Stephen Vardeman. Specifically I would like to thank Dr. Philip Dixon for assisting me in the early stages of my graduate studies and teaching me how to be a successful statistical consultant. Finally I would like to thank my wife Dr. Amy Lock and my entire family for their constant encouragement and support.

ABSTRACT

Within this dissertation are 3 papers application of statistical analyses to data in sport. We discuss the common methods of estimating in-game win probability values and present an approach using random forests that is uniformly applicable to all head-to-head competitions. The random forest is a non-parametric machine learning methodology common in big data regression and classification problems. We demonstrate the performance and usefulness of our method to the NHL, NBA and NFL. We also introduce a new methodology to account for missing values that are associated with the linear predictor in order to improve the estimation of NFL field goal kicker accuracy. Due to its flexibility, we believe the that the framework for incorporating information underlying missing values could be useful in a wide array of applications.

CHAPTER 1. INTRODUCTION

While statistics in sports has existed for many years (Bill James' first Baseball Almanac was published in 1977) it is more recent that statistical analyses of sport has become prevalent in the professional and academic setting. There are now multiple peer reviewed journals that seek out innovative methodologies for analyzing data in sport, multiple high profile conferences each year dedicated specifically to presenting statistics in sport research, and a growing Section on Statistics in Sports within the American Statistical Association. As with many other fields, the amount and accessibility of sports data being collected is currently experiencing an influx, creating the opportunity for advanced statistical analyses such as those presented in this dissertation. The unifying aspect of each of the three major chapters within this dissertation is the application of statistical analyses to data in sport, with a focus on in-game win probability estimation and performance evaluation.

One of the more useful and interesting sports statistics is the estimated win probability between two teams or individuals during a match. This dissertation examines the estimation of in-game win probability on play-by-play data. Play-by-play data is data that gives a detailed state of the match at various stages from start to finish. As an example in a football game play-by-play data will give the state of the game before each play, in basketball it gives the state of the game before each possession. Estimating the win probability on play-by-play data is not a novel concept, companies such as the Entertainment and Sports Programming Network (ESPN) and Major League Baseball currently use win probability analyses as part of the fan experience. We review existing win probability methodologies and win probability research. We found that existing win probability methodologies within previously established research are highly variable, with each typically designed exclusively for one sport. One of the valuable results of this dissertation is presenting a win probability estimation procedure that can be applied to any head-to-head based competition with minimal adjustment required.

In Chapter 2, we introduce an empirical methodology for estimating in-game win probability values on play-by-play data within a National Football League (NFL) game using the random forest. The methodology combines information from several pre-play situations and game status variables along with the Las Vegas point spread to estimate the win probability. We also introduce a way to measure the performance of a win probability method, accounting for the fact that there is no observed in-game win probability value. Once performance is established we discuss several ways to use the estimated win probability values on NFL data, including examination of which variables are important during different facets of the game, producing game plots and identifying the influential plays, and evaluating various in-game coaching decisions.

In Chapter 3, we expand the methodology introduced in chapter 2 by applying the random forest method to the National Hockey League (NHL) and National Basketball Association (NBA) along with the NFL. With these three sports providing evidence we demonstrate that the method could be a uniform way of estimating win probability in any head-to-head based competition. We also revisit the uses of estimated win probability values providing examples from each of the three sports. One such use is a hierarchical model to evaluate the performance of NBA players by predicting the change in win probability during portions of a match using the 5 players on the court for each team as the predictor variables.

Within Chapter 4, we examine situations where missing values of a response variable are related to values of the predictor variables. Motivation for this work came from analyzing Bernoulli distributed NFL field goal data, where the existence of a trial should be related to the probability of success. We introduce a new methodology that uses the linear predictor's relationship with the missing values to improve estimation of fixed and random effects in a general linear mixed model. Logically a coach's decision whether or not to attempt a field goal should relate to the kicker's probability of success, implying that missing trials should be informative about the abilities of each kicker. By applying our method we are able to account for the coach's decision and incorporate information from the missing values into estimation of kicker accuracy. Since a missing value may depend on variables other than the linear predictor's relationship with the response variable the method is designed such that additional variables can be applied to estimate the probability of a missing value.

Chapter 5 summarizes the results presented in Chapter 2, 3 and 4, while making broad conclusions and discussing possibilities for future research.

CHAPTER 2. USING RANDOM FORESTS TO ESTIMATE WIN PROBABILITY BEFORE EACH PLAY OF AN NFL GAME

A paper accepted by *Journal of Quantitative Analyses in Sports*

Dennis Lock and Dan Nettleton

Abstract

Before any play of a National Football League (NFL) game, the probability that a given team will win depends on many situational variables (such as time remaining, yards to go for a first down, field position and current score) as well as the relative quality of the two teams as quantified by the Las Vegas point spread. We use a random forest method to combine pre-play variables to estimate Win Probability (WP) before any play of an NFL game. When a subset of NFL play-by-play data for the 12 seasons from 2001 to 2012 is used as a training dataset, our method provides WP estimates that resemble true win probability and accurately predict game outcomes, especially in the later stages of games. In addition to being intrinsically interesting in real time to observers of an NFL football game, our WP estimates can provide useful evaluations of plays and, in some cases, coaching decisions.

2.1 Introduction

The probability that a particular team will ultimately win an NFL game can be difficult to estimate at a specific moment. Undoubtedly fans, coaches, and players alike implicitly consider this probability as a game unfolds. We develop a statistical method for estimating this win probability (WP) prior to any play of an NFL game. As an example of what our methodology can produce, Figure 2.1 illustrates our WP estimates of a Baltimore Ravens victory in Super Bowl 47 prior to every play of the game.

In addition to being of interest to fans as they watch a game in progress, our WP estimates could be used to evaluate specific plays and coaching decisions. For instance by comparing WP estimates

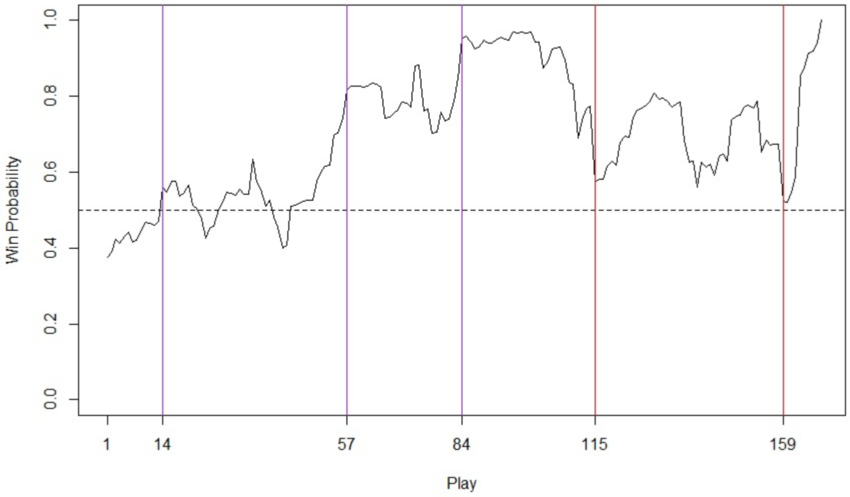


Figure 2.1 Estimated Baltimore Ravens win probability from every play of Superbowl 47. Highlighted plays with score and Ravens win probability: 1 = Start of game (0-0, $WP = 0.382$), 14 = Ravens score first (7-0, $WP = 0.552$), 57 = Ravens intercept the ball on the first play following their second touchdown (14-3, $WP = 0.743$), 84 = Ravens Open half with 109 Yard kickoff return (28-3, $WP = 0.935$), 115 = 49ers recover a fumble after back to back touchdowns (28-20, $WP = 0.572$), 159 = 49ers first and goal with 2 minutes remaining (28-23, $WP = 0.524$).

in Section 2.7, we examine whether certain penalties should be accepted or declined and whether an offensive team should kick a field goal on fourth down or attempt to get the first down. While we discuss only a few specific examples, similar analyses can be used by coaches to strengthen decisions and enhance game strategy.

The idea of WP estimation for major sports is not new. Early uses of win probability were primarily in Major League Baseball but have existed since the beginning of the 1960's (Lindsey (1961)). Recent books on baseball analytics dedicate entire sections or chapters to the topic of win probability (Schwartz (2004), Tango et al. (2006)). Analysts and fans of the other major sports have also begun to examine and use win probability more recently. For NBA and NHL examples, see Stern (1994) and Buttrey et al. (2011), respectively.

Motivation for this paper came partially from Brian Burke's NFL win probability metric found at *www.advancednflstats.com*. Burke constructs a play-by-play win probability using mostly empirical estimation. His win probability focuses on in-game variables score, time remaining, down, yards to go for a first down and field position. His general strategy is to partition the observations of a training dataset into bins based on values of his predictor variables score, time remaining, and field position. The proportion of training observations in a bin that correspond to a win for the team on offense provides an estimate of the win probability for the offensive team whenever the score, time remaining, and field position of a new situation are consistent with the bin. Adjustments to WP based on first down conversion probabilities are included to account for down and yards to go. Some extrapolation and smoothing are used to incorporate information from situations in other bins similar to the situation for which a prediction is desired.

We attempt to enhance Burke's approach in several ways. First, rather than subjectively binning the training observations, we let the data define a partitioning using a method that attempts to minimize prediction error. Second, we include the pre-game point spread to measure the quality of both teams competing. Thus, unlike Burke's approach, our method provides WP estimates that differ from 50% for each team at the beginning of a game. Third, our method permits the use of additional variables and provides a natural assessment of the importance of each variable. Finally, the approach we propose can be applied in a largely automatic and straightforward manner to other sports when sufficient training data are available.

The heart of our WP estimation methodology is the random forest (Breiman (2001)). The random forest is a good candidate for our prediction function for many reasons. First, and perhaps most important is the well documented predictive ability of a random forest (see, for example, Breiman (2001b), Svetnik et al. (2013), Diaz-Uriarte and de Andres (2006), Genuer et al. (2008), Genuer et al. (2010)). Second, a random forest can combine many predictor variables with unknown interactions in a non-linear data driven manner. Third, the random forest method provides natural and effective assessment of variable importance (Breiman (2001b)). Finally, the method runs on minimal assumptions, handles outliers well, and predicts based on empirical evidence (Breiman (2001), Liaw and Wiener (2002), Cutler et al. (2007)).

Successful use of the random forest has begun to appear in sports analytics recently. Some examples include predicting major league success in minor league baseball players (Chandler and Stevens (2012)), predicting hall of fame voting in baseball (Freiman (2010), Mills and Salaga (2011)), and predicting game outcome in non-American football matches (Hucaljuk and Rakipovic (2011)). It should be noted each of these examples used a random forest of classification trees. Our approach differs somewhat because our WP estimates are generated by a random forest of regression trees.

In Section 2.2, we discuss the training data used to construct our WP estimates. In Section 2.3, we describe the random forest estimation method. In Section 2.4, we examine the performance of our estimator. Sections 2.5 through 2.7 evaluate the importance of variables, and their effects on win probability, changes in win probaiblity during the course of games, and using win probability estimates to analyze coaching decisions. The paper concludes with a discussion including alternative approaches, future considerations, limitations, and conclusions in Section 2.8.

2.2 Training Data

The analyses in this paper are based on all play-by-play data from NFL seasons 2001 through 2012 obtained from *ArmChairAnalysis.com*. Except where noted otherwise, data from the 2012 season were set aside as a test set, and only data from 2001 to 2011 were used as a training set. This training set consists of $n = 430,168$ plays from $2,928$ games with $p = 10$ predictors for each play extracted or constructed from the play-by-play data. A list and description of each predictor variable is included in Table 1. We use Y to denote the $n \times 1$ response vector and X to denote the $n \times p$ matrix of predictor values. Each row of the data matrix $[Y,X]$ corresponds to one pre-play situation observed with respect to the offensive team. The response is an indicator of victory so that the i^{th} element of Y is 1 if the team on offense before play i won the game and 0 otherwise.

Table 2.1 Description of predictor variables

Variable Name	Variable Description
Down	The current down (1st, 2nd, 3rd, or 4th)
Score	Difference in score between the two teams
Seconds	Number of seconds remaining in the game
AdjustedScore	$Score/\sqrt{Seconds+1}$
Spread	Las Vegas pre-game point spread
TIMO	Time outs remaining offense
TIMD	Time outs remaining defense
TOTp	Total points scored
Yardline	Yards from own goal line
YTG	Yards to go for a first down

2.3 Random Forest Method

Random forests generate predictions by combining predicted values from a set of trees. Each individual tree provides a prediction of the response as a function of predictor variable values. We use a forest of regression trees, where each individual regression tree is generated as follows.

1. Draw a bootstrap sample of observations from the training dataset and group all sampled observations in a single node N_0.

2. Randomly select m predictor variables from all p predictors.

3. For each x among the m selected predictors and for all cut points c, compute the sum of squared errors
$$\sum_{k=1}^{2} \sum_{i \in N_k} (y_i - \bar{y}_k)^2,$$
where N_1 is the set of training observations with $x \leq c$, N_2 is the set of training observations with $x > c$, and $\bar{y}_k$ is the response mean for training observations in N_k $(k = 1,2)$.

4. Choose x and c to minimize the sum of squared errors in step 3, and split the training observations into two subnodes accordingly.

5. Repeat steps 2 through 4 recursively at each resulting node until one of two conditions are met:

The final nodes that result from this recursive partitioning process are referred to as terminals nodes. This series of splits can be presented graphically as a binary tree, where each split constructs the "branches" and the final "leaves" represent the terminal nodes. Once the tree is constructed from the training data, a predicted response for a future observation can be found by tracing the observation's path down the branches of the tree to a terminal node (based on the observation's predictor variable values) and computing the average of the training responses in that terminal node. The prediction of the forest is then obtained by averaging the predictions of all trees in the forest.

As discussed by Lin and Jeon (2006) and Xu et al. (2014), random forests are similar to adaptive nearest-neighbors methods, which predict the response for a target observation by averaging the responses of the "nearest" training observations. Such methods are adaptive in the sense that the definition of "nearest" is based on a concept of distance in the predictor variable space that accounts for the relationship between the predictors and the response inferred from the training data. Predictor variables unrelated to the response are ignored while predictor variables strongly associated with the response play a major role when evaluating the distance between observations. In our application, the random forest win probability estimate for a given target play is a weighted average of game outcomes associated with past plays that are judged by the random forest algorithm to be similar to the target play. These similar training set plays are those that make up the terminal nodes of trees in the forest that contain the target play. The game outcomes for the training set plays most similar to the target play (i.e., those that occur most often in terminal nodes associated with the target play) get heavily weighted while outcomes for dissimilar plays (i.e., those seldom in a terminal node with the target play) receive little or no weight.

We construct our random forest using the function *randomForest* in the R package *randomForest*. The random forest has two tuning parameters, m the number of candidate predictors at each split and *nodesize* the maximum terminal node size. We chose both parameters using a cross-validation strategy described as follows. Play-by-play data from the 2011 season were set aside, and WP estimates for plays from the 2011 season were generated using random forests constructed from plays in 2001 through 2010 with various choices of m and *nodesize*. Based on the resulting misclassification rates, we chose $nodesize = 200$ (well above the R *randomForest* regression default of 5) and $m = 2$ (slightly below the default value of ($\lfloor p/3 \rfloor = 3$). The *randomForest* default of 500 regression trees were constructed with these two tuning parameter choices.

The decisions to use regression trees and to use the constructed variable *AdjustedScore* were also based on our cross-validation performance. For our data, the main difference between regression and classification trees is the predicted response in the terminal node of a regression tree is the proportion of response values equal to 1, while a classification tree reports a 1 if the proportion is greater than 0.50 and a 0 otherwise. The variable *AdjustedScore* was included to improve the performance of the method primarily in the later stages of games. Because we know *a priori* that a nonzero lead increases in value as the seconds remaining in a game decreases, we considered using

$$AdjustedScore(\gamma) = \frac{Score}{(Seconds + 1)^{\gamma}}$$

for $\gamma = 0, 0.1, 0.5, 1, 1.5$ or 2. We ultimately selected $\gamma = 1/2$ because this value of γ minimized our cross-validation misclassification. Note that this cross-validation analysis favors using *AdjustedScore* and *Score* over *Score* alone because

$$AdjustedScore(\gamma) = Score \text{ for } \gamma = 0,$$

which was one value in our candidate set from which $\gamma = 1/2$ was selected.

2.4 Win Probability Prediction Accuracy

Measuring the accuracy of estimated win probabilities is a difficult task. For instance, when it appeared the 49ers were about to score late in Super Bowl 47 we estimated a 48% chance of victory. This may have been an accurate estimate of their win probability despite the fact that they lost the game 9 plays later. One basic way to measure accuracy is to calculate mean squared error (0.156) from all plays in our test set, where the example above contributes $(0-0.48)^2$ to the numerator of that mean squared error. Another option is to look at the mean squared error as the game progresses (Table 2). Mean squared error should decrease as the game progresses because we gain more information and move closer to the final response.

Table 2.2 Test set MSE by quarter

Quarter	1st	2nd	3rd	4th
MSE	0.201	0.177	0.143	0.107

Possibly a better way to measure accuracy is to bin the plays in the test set by estimated win probability and then calculate the proportion of wins in each bin. This proportion of wins is a representation of the unknown true win probability for the plays in a given bin. If a method performs well, we would expect estimated win probabilities that define bins to be similar to the actual proportion of wins within bins. For example, among plays with an estimated WP ≈ 0.75, aproximately 75% should be associated with an offensive win. Figure 2.2 shows a plot of estimated win probability (binned in 5% increments) for plays in the test set and the proportion of offensive wins among plays in each bin. Correlation between proportion of wins and the WP at the center of each bin is extremely high ($r = 0.998$), and the random forest WP estimates are clearly well calibrated.

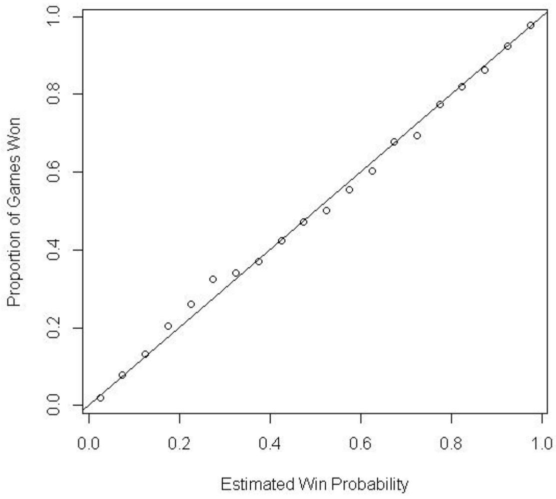
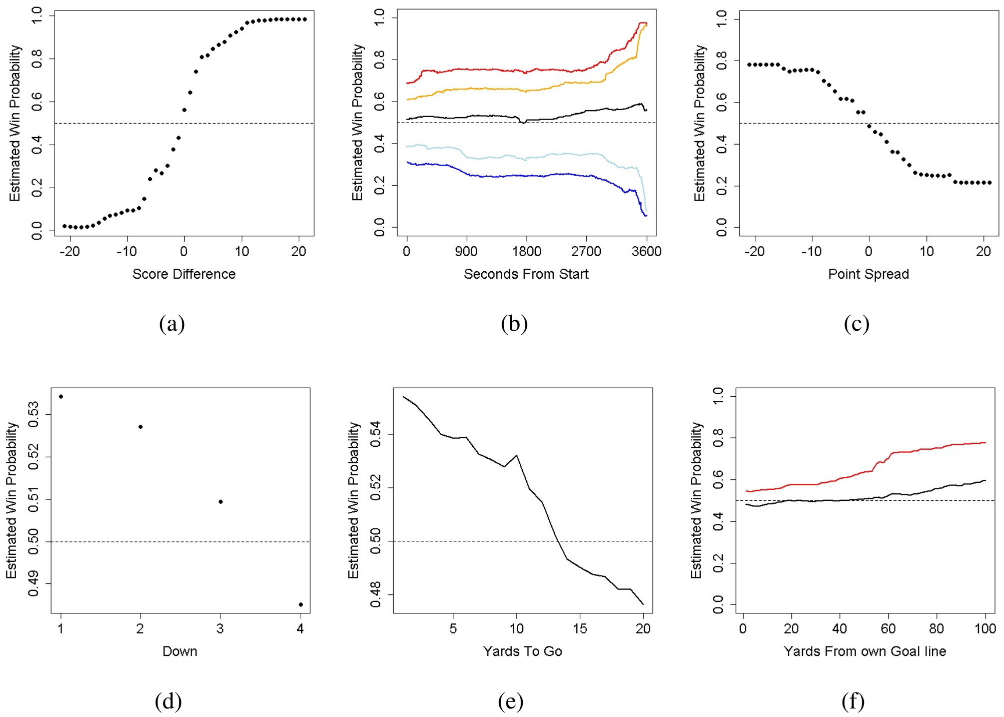


Figure 2.2 Binned estimated win probability and proportion of games won in each bin (line = a perfect fit).

2.5 Assessing Variable Importance and Relationships with Win Probability

In addition to the performance measures discussed in Section 2.4, it is interesting to examine how WP estimates change when one variable expected to have an effect on win probability is changed while holding the others constant. For example, Figure 3.1(a) shows how WP changes as the difference in score changes while holding all other variables constant. The other plots in Figure 2.3 show the effect of varying seconds (b), spread (c), down (d), yards to go (e), or yards from own goal line (f). Primarily of note is that each variable changes win probability in the direction we would expect, with *Score* having the greatest effect. That being said their are many other interesting features to note. For example, in Figure 2.3(b), we see that WP changes little over time until around the 4th quarter for each of the score differences examined. The black line in Figure 2.3(b) shows that having the ball in a tied game at your own 20 is advantageous (WP $>$ 0.5) until just before halftime when it provides no advantage to either team (WP $\approx$ 0.5). When varying point spread in Figure 2.3(c) we see many plateaus, especially in the more extreme point spreads where the random forest is primarily grouping an interval of point spreads together as equivalent. Also no team is given a pre-game win probability greater than 80%, regardless of the point spread. Figure 2.3(f) shows that improving field position is noticeably more important with less time remaining, with a considerable increase around the opposing 40 yard line (entering field goal range).



(a)

(b)

(c)

(d)

(e)

(f)

Figure 2.3 Changing one variable at a time with others held constant $Down = 1$, $YTG = 10$, $Yardline = 20$, $TOTp = 28$, $Seconds = 300$, $Score = 0$, $Spread = 0$, $TIMO = TIMD = 3$ unless otherwise specified. Variable changed in plot (a) *Score*, (b) *Seconds* (blue: $Score = -7$, Light Blue: $Score = -3$, Black: $Score = 0$, Orange: $Score = 3$, red: $Score = 7$), (c) *Spread* ($Seconds = 3600$), (d) *Down*, (e) *YTG* ($Down = 3$), and (f) *Yardline* (black: $Seconds = 1700$, red: $Seconds = 120$). Note the y-axis is focused on a narrower range in plots (d) and (e).

The importance of score difference is apparent graphically from the plots in Figure 2.3, but we can also numerically estimate variable importance. We chose to calculate importance for the kth variable as follows.

1. Randomly permute the values of predictor variable k within the test set and re-predict WP.

2. For each play i calculate the squared error after permuting the values of variable k minus the

original squared error.

3. Repeat steps 1 and 2 many times (we chose 100 repetitions), and find the average increase in squared error for each play *i*.

This provides a play-wise variable importance for all plays in the test set. Overall variable importance can be found by averaging across all plays, and variable importance for specific types of plays can be found by averaging over plays of given type. Table 2 shows overall and quarter-specific measures of variable importance for each of our 9 variables (note that the variable *AdjustedScore* is just a function of *Seconds* and *Score* so was not permuted separately but recalculated for permutations of either *Seconds* or *Score*). Overall and in three of the four quarters, *Score* is the most important variable, however in the first quarter *Spread* is actually more important than *Score*.

Table 2.3 Variable importance (overall and by quarter)

Variable	Overall	Qtr 1	Qtr 2	Qtr 3	Qtr 4
Score	0.13653	0.04697	0.09773	0.15348	0.23539
Spread	0.02462	0.05361	0.02919	0.01436	0.00459
Seconds	0.00657	0.01105	0.00570	0.00341	0.00643
Yardline	0.00265	0.00276	0.00139	0.00208	0.00428
TOTp	0.00160	0.00195	0.00000	0.00152	0.00334
Down	0.00031	0.00038	0.00017	0.00018	0.00045
TIMO	0.00019	0.00040	0.00000	0.00005	0.00062
TIMD	0.00013	0.00023	0.00000	0.00017	0.00062
YTG	0.00009	0.00010	0.00002	0.00006	0.00018

One major advantage of calculating variable importance in this way is that we can examine how two variables interact by observing the importance of one variable at specific values of the other variable. For example, in Table 3 we can see that the difference in score becomes more important as the game progresses while the point spread becomes less important. Figure 2.4 shows a plot of the interaction between *Spread* and *Seconds*, looking at the importance of *Spread* at each second. Not surprisingly the importance of *Spread* is relatively high at the beginning of games (when not much other information is available) but diminishes to near irrelevance in the closing seconds.

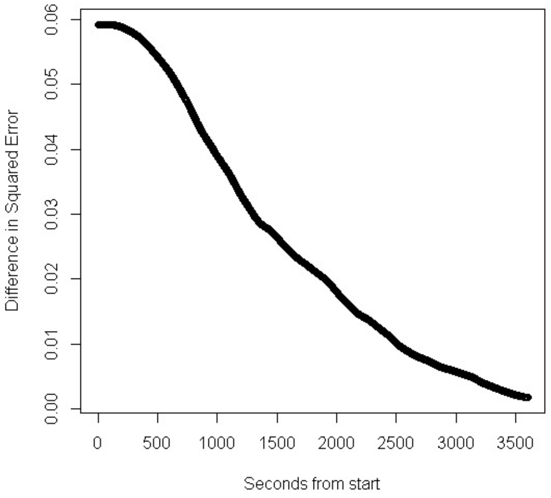


Figure 2.4 Variable importance of *Spread* by seconds from start (smoothed using a loess smoother).

2.6 Examining Changes in Win Probability

We can use change in win probability (ΔWP) to judge the most influential plays within a specific set of plays. For example, David Tyree's catch in Super Bowl 42 (ΔWP = 0.113), rated the greatest Super Bowl play in NFL history by Fox Sports, was actually not even the most influential play of that game. The touchdown pass to Plaxico Burress 4 plays later had a much greater increase in win probability (ΔWP = 0.389). If we were to choose the greatest Super Bowl play based on ΔWP, it would be James Harrison's 100 yard interception return for a touchdown just before halftime in Super Bowl 43 (ΔWP = 0.511). Similarly, we can judge the best play of the entire 2012 season to be Cecil Shorts' 39 yard touchdown reception to take a 1 point lead over the Vikings with 27 seconds remaining in the 4th quarter (ΔWP = 0.710).

Using the predicted win probability values from an entire game, we can plot how win probability changed as the game progressed. The plots of two of the more exciting Super Bowls from the last 12 seasons are presented below. Figure 2.5 shows Super Bowl 44 between the Indianapolis Colts and New Orleans Saints, and Figure 2.6 shows Super Bowl 42 between the undefeated New England Patriots and New York Giants.

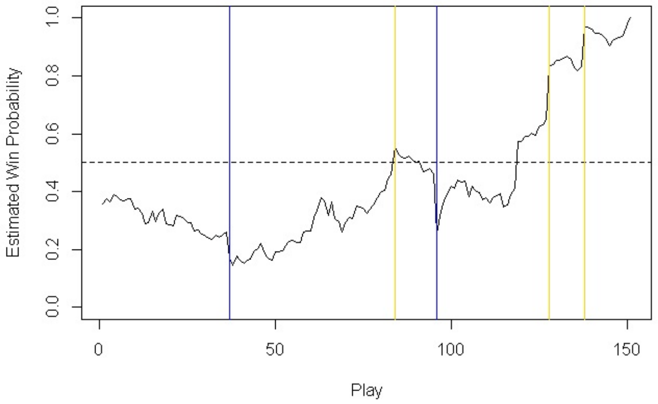


Figure 2.5 Estimated win probability by play for Superbowl 44, blue vertical lines represent Indianapolis touchdowns and gold vertical lines represent New Orleans touchdowns.

2.7 Win Probability Analysis of Coaching Decisions

WP and ΔWP can be used to evaluate some coaching decisions. Suppose, for example, that the offense is flagged for a holding penalty while throwing an incomplete pass on third and 10 from their own 20 yard line. In a game between evenly matched teams (*Spread*= 0) with the score tied 7 to 7 at the beginning of the second quarter, should the defense decline the penalty and force a fourth and 10 situation, or accept the penalty to put the offense at third and 20 from their own 10 yard line? Random forest WP calculations can provide guidance and favor accepting the penalty even though declining would almost surely guarantee a punt and a change of possession. The WP for the offense at fourth and 10 from the 20 is estimated to be 46% compared to 43% at third and 20 from the 10.

It is possible, of course, that the coach on the field will base his decision on additional information not available to the random forest. If the offense is facing a strong headwind, for example, accepting the penalty may be even more favorable than indicated by the random forest WP estimates. On the other hand, early-game injuries in the defensive backfield might make declining the penalty better than asking an inexperienced defense to make another stop. However, regardless of additional information, the random forest does provide useful baseline information that indicates that, in the past, teams facing third and 20 from their own 10 have lost more often than those in situations like fourth and 10 from the 20.

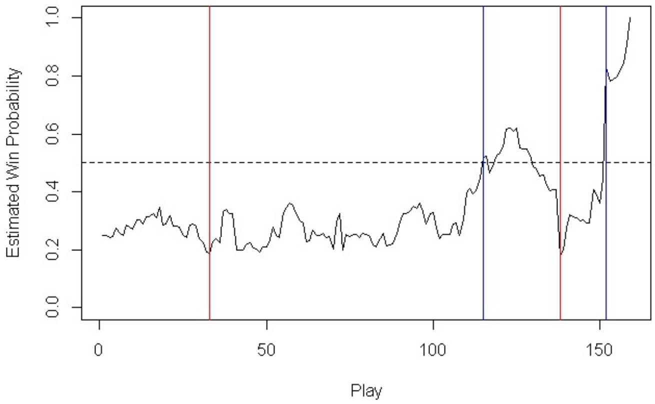


Figure 2.6 Estimated win probability by play for Superbowl 42, blue vertical lines represent New York touchdowns and red vertical lines represent New England touchdowns.

A very similar accept vs. decline WP analysis can be used in other situations, such as when a punting team commits a penalty that, if accepted, would almost surely result in a re-punt. The receiving team's WP at first and 10 from their current field position can be compared to their WP if the penalty were assessed and the punting team were to face fourth down again closer to their own goal line. Because the best choice may depend on the capabilities of the special teams involved, the WP analysis may not be able to give a definitive answer. However, the WP calculations can serve as a useful starting point for making an informed decision.

As another example, suppose the offense trails 14 to 10 and faces a fourth and 3 from the opponent's 10 yard line. Should they take the almost certain 27 yard field goal to cut their deficit to 14–13 or try for a first down? When the offense's WP at fourth and 3 is greater than their WP would be following a successful field goal, going for the first down is likely the better choice. Figure 2.7 shows that kicking the field goal is a good decision in the first half while going for the first down is better with about 10 or fewer minutes to go in the game. The two options are approximately equivalent throughout the third quarter.

As noted for the first examples of this section, the WP analysis provides baseline guidance constructed from past performance that could be overridden when special circumstances or specific strengths, weaknesses, and tendencies of the competing teams are taken into account. It is important to remember that our training data are observational rather than experimental. Teams facing similar situations in the

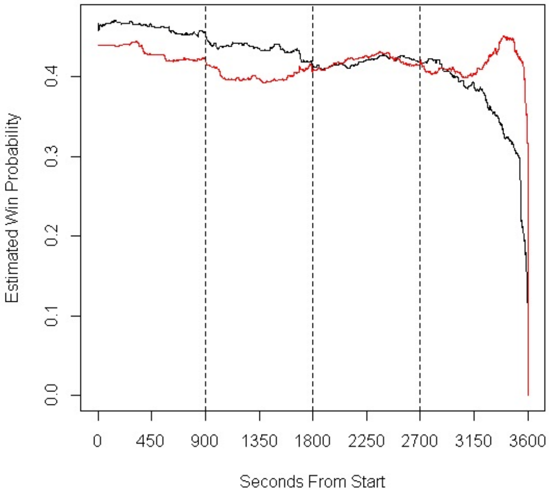


Figure 2.7 Estimated win probability before (red) and after (black) a successful 27 yard field goal on fourth down and 3 when the offense is trailing by a score of 14 to 10 prior to the kick (Other variables: $TIMO = TIMD = 3$, $Spread = 0$). Dotted lines indicate the changing quarters.

past have not been randomly assigned to courses of action by an experimenter. Thus, we cannot be certain that decisions to go for the first down rather than kicking the field goal with, say, 8 minutes to go in a game *caused* the higher success rate experienced by teams that chose to go for the first down. However, in general, we believe coaches should avoid attempting plays that, even if successful, will result in a decrease in their team's WP.

2.8 Discussion

Win probabilities estimated through our random forest method are similar to those calculated by Brian Burke. Figure 2.8 shows estimated WP for Superbowl 45 ($Spread = 2.5$) using both a random forest and Brian Burke's estimation. In general, because we include the point spread variable, our methodology provides better WP estimates near the beginning of games, especially in games with a clear favorite.

One major advantage of our approach is that it is fairly simple and could easily be replicated in other sports provided sufficient data is available. Due to its nature and performance, random forest methodology offers a unified approach to predict in-game win probability across many sports. In other work in progress, we have used random forests to estimate WP in the NHL and NBA, with success

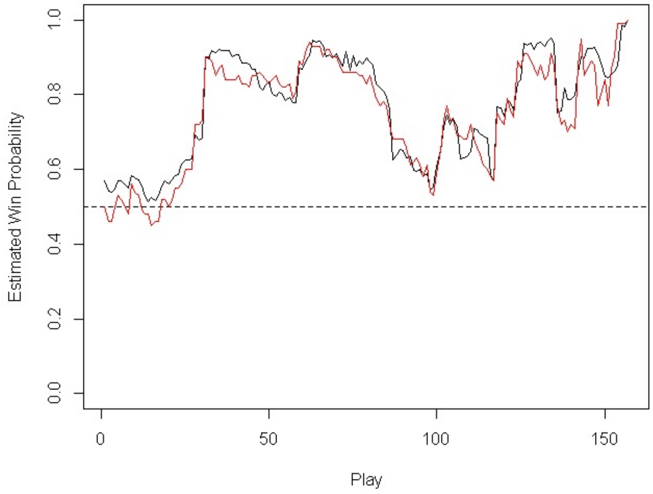


Figure 2.8 Estimated win probability by play for Superbowl 45 with both random forest estimation (Black) and Brian Burke's estimation (Red).

similar to that reported here for the NFL.

In addition to the WP calculator that served as motivation for our NFL work, Burke has also developed WP calculation methods for the NBA and NHL. While both a random forest and Burke's method predict win probability through historical values, there also exist methods which predict through simulation (some good examples include *accuscore.com* and *predictionmachine.com*). These methods differ in that WP is estimated utilizing win proportions from simulated game outcomes, rather than win proportions from historical situations. Because the details of the methods underlying the simulations are proprietary, it is not possible for us to evaluate the performance of these methods relative to our random forest approach.

Using either a random forest method or Brian Burke's method, specific WP values are estimated without regard to previous plays; only the current situation rather than the events leading up to that situation is considered. This may not be detrimental as there are many papers that discount the effect of momentum in the NFL (Johnson et al. (2012), Fry and Shukairy (2012)). A related issue is that each game has approximately 150 sequential observations all associated with the same response, and the random forest treats these observations as independent.

We attempted multiple adjustments to our random forest approach, some of which were meant to account for potentially detrimental effects of dependence in the data. To account for momentum, we attempted including the win probability estimates at the end of each quarter as predictors for later stages in the game. To account for multiple observations within each game, we constructed 3,600 separate random forests for each second, where the set of plays in each forest contains one observation per game, chosen as the closest observation to that second. This guaranteed that each forest consisted of 2,928 independent observations (under the assumption that plays from different games are independent). Future predictions were found by generating a prediction from the forest that corresponds with the current second. Methods with separate random forests by down or quarter were also attempted, such that, for instance, 4th down plays were estimated only from other 4th down plays. With each of these adjustments, we either saw no improvement in performance with added complexity (Momentum adjustment, separate forests by quarter) or a decrease in performance with added complexity (separate forests by second, separate forests by down), so no adjustment was included. To be thorough we constructed a few simple models unrelated to random forests, such as logistic regression on win/loss and linear regression on final score difference, each with large decreases in performance.

We also considered information other than *Spread* to account for team quality. The most basic was just adding variables such as current record, points per game, yards allowed per game, etc. to the variables in the forest. The most advanced was combining team quality variables using either a logistic regression or pre-game random forest to come up with a pre-game win probability variable to include as another predictor in a subsequent within-game forest. None of these alternative approaches showed an improvement over a model with only the point spread.

Throughout this paper, WP estimates for a play in a given game *i* were generated from training data that did not include plays from game *i*. For example, as previously discussed, data from 2001 through 2011 were used as the training set for generating predictions in 2012. Similarly, WP estimates for Super Bowls prior to 2012 were generated by using plays from 2001 through 2011 excluding the 11 Super Bowls. In future application, our random forest could be retrained after each week of NFL games to include play-by-play data from 2001 up through the most recently completed NFL games.

In conclusion, we have developed a method of estimating win probability that performs well and is simple to replicate. Regardless of how pre-play win probabilities are estimated, the uses of these values are numerous and could improve the way we look at the game.

Bibliography

Breiman, L. (2001a). Random forests. *Machine Learning*, 45:5–32.

Breiman, L. (2001b). Statistical modeling: the two cultures. *Statistical Science*, 16:3:199–231.

Buttrey, S. E., Washburn, A. R., and Price, W. L. (2011). Estimating NHL scoring rates. *Journal of Quantitative Analysis in Sports*, 7:3.

Chandler, G. and Stevens, G. (2012). An exploratory study of minor league baseball statistics. *Journal of Quantitative Analysis in Sports*, 8:4.

Cutler, D. R., Edwards, Jr., T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., and Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology*, 88:11:2783–2792.

Diaz-Uriarte, R. and de Andres, S. A. (2006). Gene selection and classification of microarray data using random forest. *Bioinformatics*, 7:3.

Freiman, M. H. (2010). Using random forests and simulated annealing to predict probabilities of election to the baseball hall of fame. *Journal of Quantitative Analysis in Sports*, 6:2.

Fry, M. J. and Shukairy, F. A. (2012). Searching for momentum in the NFL. *Journal of Quantitative Analysis in Sports*, 8:1.

Genuer, R., Poggi, J., and Tuleau, C. (2008). Random forests: some methodological insights. *arXiv*.

Genuer, R., Poggi, J., and Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, 31:14:2225–2236.

Hucaljuk, J. and Rakipovic, A. (2011). Predicting football scores using machine learning techniques. *MIPRO, 2011 Proceedings of the 34th International Convention*, pages 1623–1627.

Johnson, A. W., Stimpson, A. J., and Clark, T. K. (2012). Turning the tide: big plays and psychological momentum in the NFL. *MIT Sloan Sports Analytics Conference 2012*.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomForest. *R News*, 2:3:2225–2236.

Lin, Y. and Jeon, Y. (2006). Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, 101(474):578–590.

Lindsey, G. R. (1961). The progress of the score during a baseball game. *Journal of the American Statistical Association*, 56:703–728.

Mills, B. M. and Salaga, S. (2011). Using tree ensembles to analyze National Baseball Hall of Fame voting patterns: an application to discrimination in BBWAA voting. *Journal of Quantitative Analysis in Sports*, 7:4.

Schwartz, A. (2004). *The numbers game*. Thomas Dunne Books, New York.

Stern, H. (1994). A brownian motion model for the progress of sports scores. *Journal of the American Statistical Association*, 89:1128–1134.

Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., and Feuston, B. P. (2013). Random forest: a classification and regression tool for compound classification and QSAR modeling. *Journal of Chemical Information and Modeling*, 53:8.

Tango, T., Lichtman, M., Dolphin, A., and Palmer, P. (2006). *The Book: Playing the Percentages in Baseball*. TMA Press, New York.

Xu, R., Nettleton, D., and Nordman, D. J. (2014). Predictor augmentation in random forests. *Statistics and Its Interface*, 8:2.

CHAPTER 3. USING RANDOM FORESTS TO ESTIMATE IN-GAME WIN PROBABILITY

A paper submitted to the *Journal of Quantitative Analyses in Sports*

Dennis Lock and Dan Nettleton

Abstract

In-game win probability is the estimated probability of victory for each competitor in the midst of a competition. We discuss the common methods of estimating in-game win probability values and present an approach using random forests that is uniformly applicable to all head-to-head competitions. The random forest is a non-parametric machine learning methodology common in big data regression and classification problems. Applying our method to the NHL, NBA and NFL demonstrates the performance and consistency of the approach. We also discuss applications of win probability values, including a wins added metric that utilizes win probability values to evaluate the overall contributions of NBA players.

3.1 Introduction

The probability of victory (win probability) over the course of a sporting event has recently become a more common and useful statistic. The way win probability (WP) values are calculated can vary greatly from sport to sport and even within a sport. Generally each approach to estimating WP is based primarily on either simulation, formal statistical modeling, or raw empirical evidence. We present a unified empirical approach to estimating win probabilities in any head-to-head competition using random forests. Performance, usefulness, and simplicity of our random forest approach are demonstrated for three of the major North American professional leagues, the National Football League (NFL), National Basketball Association (NBA), and National Hockey League (NHL).

In contrast to our proposed approach, media outlets and most proprietary examples of win probability estimation predominantly utilize a simulation approach. The simulation approach consists of repeatedly simulating the remainder of a game from a specific moment to determine the winner, and calculating the proportion of simulations each team won. Simulation is most useful in sports with a predetermined set of plays where each play has a specific set of possible outcomes, such as at bats in baseball, holes in golf, or points in tennis. As a simulation example the following steps estimate the win probability at any moment of a tennis match between Serena Williams and Simona Halep:

1. Calculate the proportion of points Williams has won while serving (p_1), and the proportion of points Williams has won while receiving (p_2) over the course of the match leading up to the specific moment.

2. Simulate the remaining games and sets by simulating each point as a draw from the Bernoulli distribution with $p = p_1$ when Williams is serving and $p = p_2$ when Williams is receiving.

3. Repeat step (2) 1,000 times and calculate the proportion of simulated outcomes that resulted in a Serena Williams victory.

There are ways to adjust and improve this simple example, such as incorporating prior information on the respective abilities of Serena Williams and Simona Halep. Prominent examples of a simulation approach to win probability estimation include *accuscore.com* and *predictionmachine.com.*

Win Probability estimation through formal statistical modeling has been address by some researchers in academia; examples include Ryder (2004), Kaplan et al. (2014), and Asif and McHale (2014). Within statistical modeling is a subset of methodologies that model the final score, and use the estimated probability distribution of final score to develop an estimated win probability (Buttrey et al. (2011)). These methodologies are useful in sports with no clear possession or play, such as hockey or soccer. As a modeling example the following steps estimate both teams' win probabilities at any moment of a soccer match between Arsenal and Manchester United:

1. Model the number of goals each team will score in the remainder of the match as independent draws from team-specific Poisson distributions, where the Poisson mean for each team is a log-linear function of team offense/defense prowess and time remaining.

2. Calculate the discrete probability of each final score possibility by combining the Poisson distributions from (1) with the current score.

3. Use the final score probabilities to calculate the win probability for each team.

Again this is a simplified illustrative example that could easily be improved upon.

An empirical approach to WP estimation involves binning historical data to find situations from previous games similar to the current situation, and calculating win probability as the proportion of "wins" in the bin. This approach is most common in sports where there is a defined play or possession but a wider range of possible outcomes, such as American football. Brian Burke's NFL win probability found at *AdvancedFootballAnalytics.com* uses an approach relying on empirical estimation. As part of his formula Burke chose to bin data based on variables such as field position, time remaining, and score difference using informed but subjective opinion.

While the simulation approach to finding win probabilities in the tennis match above is simple, straightforward, and can yield accurate results, it would require substantial adjustments for various complications to work in baseball, and is not applicable in a sport such as soccer or hockey. Similarly the example of a modeling approach could work effectively for soccer or hockey, but cannot be easily adapted to sports such as baseball, football, or tennis. Finally Burke's approach to NFL win probability works well within the confines of the NFL but is not directly applicable to any other sport. Designing each approach specifically around the respective sport is a common theme in win probability research. We argue that accurate win probability values estimated using a random forest are possible in any head-to-head competition for which adequate training data are available.

Regardless of how they were calculated, accurate win probability values are becoming increasing useful and common to aid in game management, personnel evaluation, and enhancing the fan experience. In section 3.3 of this paper we discuss some of the more common uses and introduce a wins added (WA) metric to evaluate the overall impact of NBA players using win probability values.

3.2 Estimating Win Probability Values

Empirical win probability estimation is essentially a nearest neighbors problem, where the primary challenge is determining how similar each observation in the training data is to a current situation.

Subjectively binning the data is one approach that relies heavily on sport expertise and opinion. The random forest (Breiman (2001)) is an alternative approach, which can objectively measure the similarity between past situations and any current situation of interest from a set of predictor variables with the pre-determined goal of estimating win probability.

The random forest is a method of combining predictions from regression trees, where each regression tree is constructed from a greedy splitting algorithm for predicting a response Y using a p-dimensional vector of predictor variables X. While many variations of the random forest exist, we use the methodology constructed by Liaw and Wiener (2002), available in the *R* package *randomForest*, where each tree is generated using the following the algorithm:

1. Select a with-replacement bootstrap sample of observations from the training dataset and group all sampled observations in a single node N_0.

2. Draw a simple random sample of m predictor for all p predictors.

3. For each predictor variable x among the m selected predictors and for all possible division points d, compute the sum of squared errors

$$\sum_{k=1}^{2} \sum_{i \in N_k} (y_i - \bar{y}_k)^2,$$

   where N_1 is the subset of N_0 with $x \leq x$, N_2 is the subset of N_0 with $x > d$, and $\bar{y}_k$ is the response mean for observations in N_k $(k = 1, 2)$.

4. Find x and d that minimize the sum of squared errors in step 3. Then split the training observations into two sub-nodes accordingly.

5. Recursively repeat steps 2 through 4 at each resulting node. During this process, a node is split no further if either

   (a) the number of observations in the node is less than or equal to a chosen tuning parameter $nodesize$,

   (b) the response value is identical for all observations in the node, or

   (c) the predictor variables selected in step 2 are each constant within the node.

The algorithm contains two tuning parameters, m the number of variables chosen at each split and *nodesize* the maximum terminal node size, which can be adjusted to maximize performance using cross validation. Once each tree (T_j) is constructed from the training data using the steps above, a predicted response value for a given X can be found by tracing the observation's path down the branches of the tree to a terminal node and computing the average of the training response values in that terminal node ($\hat{Y}_j$). The prediction of the forest is then obtained by averaging the predicted values from the full set of trees, $\hat{Y} = \sum_{j=1}^{J} \hat{Y}_j / J$, where $J =$ the total number of trees.

Steps 1 and 2 above ensure that a variety of different trees will be grown from the training data, reducing variation of prediction relative to a single tree and enhancing predictive performance (Breiman (2001)). The result is equivalent to predicting the response for a given vector X using a weighted average of the response values from the training data, where the weights are determined by how often the training cases appear in the terminal nodes containing X. This approach to win probability was first introduced exclusively for predicting win probability in National Football League games (Lock and Nettleton (2014)).

3.2.1 Datasets

Win probability estimation utilizing a random forest requires in-game observations from a large database of games. Each observation consists of values of p predictor variables that characterize an in-game situation, along with a response variable that indicates the winning team for the game from which the observation originated. Any of several equivalent codings of the response indicator could be used. For example, for NFL observations, we define the response as 1 if the team on offense for the observation ultimately won and 0 if the team on defense won. For the NHL and NBA, we arbitrarily define the response indicator as 1 if the home team won and 0 if the away team won. Other conventions, such as indicating whether the team that comes first alphabetically ultimately won, would produce equivalent results. Appendix A contains the list of predictor variables we used to construct win probabilities for the NFL, NHL, and NBA. As illustrated in section 3.2.3, we have found these variables to be effective for producing WP estimates that are well calibrated with observed win proportions. The selection of these particular variables, as opposed to other potentially useful variables, was driven by a combination of utility and availability.

Data to construct the NFL random forest were generated from plays in both regular season and playoff games from the 2001-2012 seasons, with the 2012 data set aside as the test data. Raw data were downloaded from the website *ArmChairAnalysis.com*. The observational unit is a pre-play situation observed with respect to the offensive team. This resulted in 430,168 pre-play situations from 2,937 games in the training data and 39,990 pre-play situations from 267 games in the test data.

Data to construct the NHL random forest were generated from all regular season games in the 2010-2011, 2011-2012, and 2012-2013 seasons, with data from the most recent season set aside as the test data. Raw data were collected using the package *nhlscraper* (Thomas and Ventura (2014)) within the *R* software environment for statistical computing. Each observation corresponds to a face-off. We chose to concentrate only on face-offs because, unlike in football, a “play” is not specifically defined in hockey. This resulted in 141,937 faceoffs from 2,457 games in the training data and 42,896 faceoffs from 720 games in the test data.

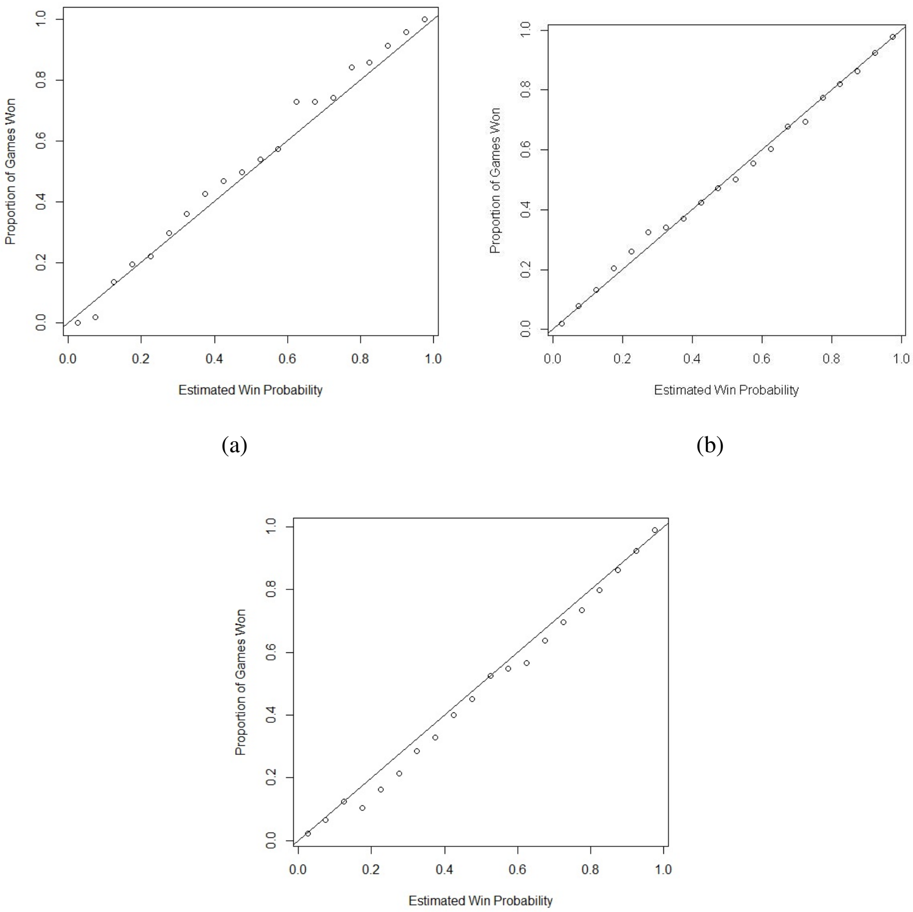
Data to construct the NBA random forest were downloaded from *BasketballValue.com* and derived from both regular season and playoff games in the 2009-2010, 2010-2011, and 2011-2012 seasons, with the 2011-2012 data set aside as the test data. For simplicity overtime games were not included in either the training or test dataset. An observation was recorded at every substitution. We focused on substitutions intentionally in order to fit the framework of a wins added performance metric, discussed in section 3.3.1, but the methodology would be identical using each possession as an observational unit. This resulted in 70,550 substitutions from 2,452 games in the training data and 25,300 substitutions from 883 games in the test data.

3.2.2 Implementation

For each training dataset (NHL, NFL, and NBA) a random forest was constructed using the *R* randomForest package (Liaw and Wiener (2002)). Tuning parameter m, the number of variables chosen at each split, was given the value of $\lfloor p/3 \rfloor$, the default value in the *randomForest* package. The tuning parameter *nodesize* for the point at which to stop splitting a node was set to 100 for the NFL and 200 for the NBA and NHL, with each choice supported by cross validation within each sport. A much larger than default *nodesize* of 5 was shown to improve the NFL random forest (Lock and Nettleton (2014)), likely by accounting for the lack of independence among observational units (an issue discussed more fully in Section 3.4). A total of 500 trees were constructed for each forest, and subsequent performance was measured on the test seasons as defined in section 3.2.1.

3.2.3 Performance

Measuring performance of any prediction methodology commonly consists of comparing the estimated value to the observed value on a set of data not included in the original analysis. When estimating win probability, however, the observed value is the final result rather than the win probability we are attempting to estimate. The actual win probability is unknown for all data points, and using the observed final result will lead to inflated error estimates (i.e. at the start of a game two evenly matched teams should have win probabilities close to 0.50, despite the observed response of 0 or 1). We overcome this difficulty by binning observations in each test set by estimated win probability values and calculating the proportion of wins in each bin. This proportion of wins is a representation of the unknown win probability for the observations in that bin. Agreement between the estimated win probability values and the proportion of wins can then be used to evaluate performance, where identical values would indicate ideal performance. Figure 3.1 shows the performance of our random forest method evaluated in this way using bins of size 0.05 on the test sets from each of the three sports. The correlation between the center of each bin and the proportion of wins in that bin is close to 1 for all three sports ($r_{NHL} = 0.996, r_{NFL} = 0.998, r_{NBA} = 0.997$).



(a)

(b)

(c)

Figure 3.1 Actual proportion of wins vs. estimated WP for the (a) NHL, (b) NFL, and (c) NBA. The diagonal line represents what would constitute ideal performance.

3.3 Using Win Probability Values

One of the most interesting and common uses of win probably to date is enhancing the fan experience. Estimated win probability values are beginning to be appear in forms such as live updates, measuring influential plays, and tracking the flow of a game. Figure 3.2 shows the estimated win probability over the course of a game between the Chicago Blackhawks and the Boston Bruins in the final game of the 2013 Stanley Cup Playoffs. In one of the most influential 14 seconds in Stanley Cup Final history, the Blackhawks probability of victory went from below 0.20 to above 0.90 with back-to-back goals late in the final period. Figure 3.3 shows the estimated win probability over the course of the 4th quarter for the Cleveland Cavaliers in the final game of the 2016 NBA Finals. The most influential basket from this game was Kyrie Irving's 3 point basket to take a 92 - 89 lead with 53 seconds remaining, increasing Cleveland's win probability from below 0.5 to 0.75.

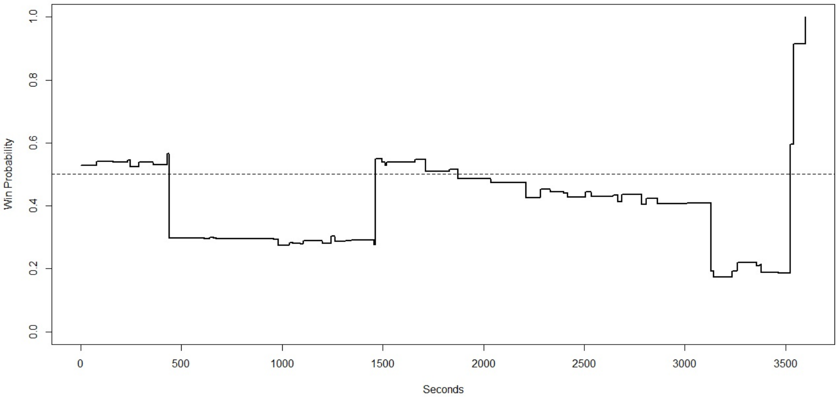


Figure 3.2 Chicago Blackhawks win probability over the course of the game. The influential points are goals by Boston, Chicago, then Boston again, and finally the most influential point at Chicago's two goals 14 seconds apart that changed the game from 1-2 to 3-2.

Win probability values can also be utilized for game management and to evaluate in-game coaching decisions. Lock and Nettleton (2014) demonstrated situations where NFL teams commonly kick field goals despite the fact that a made field goal lowers their team's win probability. Brian Burke on AdvancedFootballAnalytics looked at win probability to determine when a defense should intentionally allow a touchdown to improve their chances of winning the game. In basketball win probability can be utilized to examine when a team should intentionally foul the opposition, comparing the effect of free throws and a stopped clock to time lost without fouling. In hockey win probability analyses could suggest the best time to pull the goalie and utilize six skaters when behind late in the game.

3.3.1 Wins Added

Recently the use of win probability values has begun to expand to player evaluation (Pettigrew (2015)). A Wins Added (WA) method presented here is designed to measure the overall contribution and importance of an NBA player to his team, estimating the number of wins he added or cost his team by playing in the "shifts" that he played. A similar methodology was introduced in Deshpande and

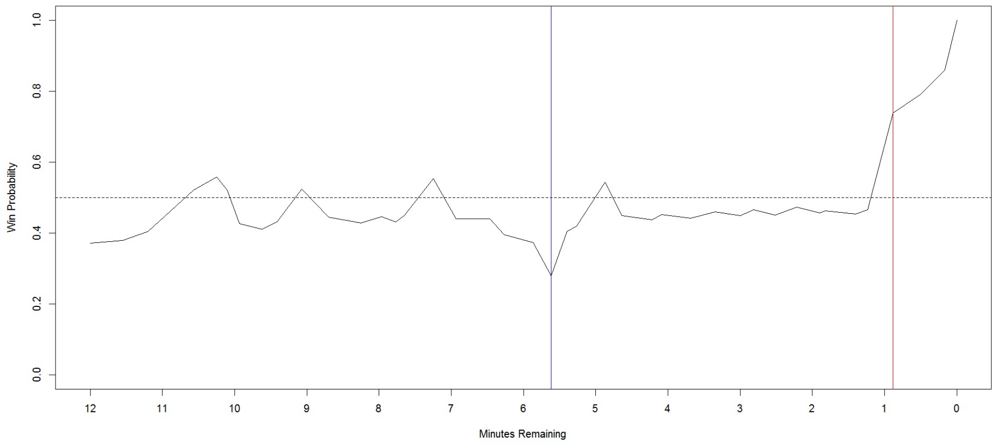


Figure 3.3 Cleveland Cavalier win probability over the course of fourth quarter of the 2016 NBA finals game 7. The two vertical lines mark the possession where Cleveland trailed by 4 with less than 6 minutes remaining (blue), and the possession where Kyrie Irving made a 3 point shot to break the 89-89 tie (red).

Jensen (2016). We define a shift as the period of time between one substitution and the next, noting that each shift ends on a substitution from either team. We model the change in win probability over the course of each shift using indicator variables for the individual players as predictor variables as follows. Let $i = 1, \ldots, n$ index shifts with n the number of total shifts from all games included in the sample, and let $j = 1, \ldots, p$ index players with p the number of players who played at least 1 shift. For each shift i,

- ΔWP_i is the change in the home team's win probability from the start of that shift to the end,
- X_i^H is a p length row vector of 0/1 indicator functions, indicating which 5 players were on the court for the home team,
- X_i^V is a p length row vector of 0/1 indicator functions, indicating which 5 players were on the court for the visiting team.

We fit a hierarchical model

$$\Delta WP_i = H + X_i^H \beta - X_i^V \beta + \varepsilon_i \tag{3.1}$$

where H is an unknown parameter to account for home court advantage, β is a p length vector with elements $\beta_1, ..., \beta_p \overset{iid}{\sim} N(0, \sigma^2/\lambda)$ and independent of $\varepsilon_1, ..., \varepsilon_p \overset{iid}{\sim} N(0, \sigma^2)$, σ^2 is an unknown positive variance component, and λ is a user-specified shrinkage parameter.

We gave shrinkage parameter λ a value of 100, chosen through a combination of cross validation and β_j shrinkage trace plots. This hierarchical aspect of the model improves test set accuracy by shrinking β_j parameter estimates towards 0 for players with fewer shifts, and stabilizing estimates of β_j coefficients of multicollinear indicator variables corresponding to teammates. Finally, we assume a diffuse uniform prior on σ^2 and a standard normal prior on H. This model design is comparable to fitting a ridge regression with penalty $\lambda = 100$. The use of a hierarchical model or ridge penalty has improved estimation in a variety of sports performance metrics (Macdonald (2012); Jensen et al. (2009)).

The resulting β_j estimates ($\hat{\beta}_j$) serve as an approximation of each individual player's average contribution to win probability each shift, accounting for the impact of teammates and opponents. Wins Added for each player j is defined as the player's estimated contribution to win probability on each shift ($\hat{\beta}_j$) times the player's total number of shifts. The top 5 players from 2009 to 2012 by Wins Added are presented in Table 3.1.

Table 3.1 Top 5 NBA players from 2009 - 2012 by Wins Added.

Rank	Player	β_j	Shifts	WA
1	Lebron James	0.0111	4936	54.9
2	Dirk Nowitzki	0.0119	3966	47.5
3	Dwight Howard	0.0089	4280	38.1
4	Chris Paul	0.0080	3824	30.4
5	Kevin Durant	0.0059	4989	29.7

This is an extension of the Adjusted +/-, originally introduced at 82games.com, replacing the difference in points with the difference in win probability from each shift and adding a hierarchical aspect. With either Adjusted +/- or Wins Added the values measure impact of each player on his respective team rather than overall quality of the player. For instance, Lebron James' Wins Added was much greater during the last season of his first stint in Cleveland than his first season in Miami (Table 3.2). This does not necessarily indicate that he was a better player during the end of his first run in Cleveland, but rather that he was more important to Cleveland's success than Miami's.

Table 3.2 Lebron James $\hat{\beta}_j$ and WA in Cleveland for the 2009/2010 season and in Miami for the 2010/2011 season.

Team	$\hat{\beta}_j$	Shifts	WA
Cleveland (09-10)	0.0144	1794	25.8
Miami (10-11)	0.0064	1828	11.5

Fitting the β_j parameters in this way shrinks players with limited shifts towards 0, which by design should represent an average NBA player. However, if a player is only asked to play in a limited number of shifts, chances are they are not an average NBA player. The coach's decision whether or not to play an individual for a shift likely provides information on that player's value, information that is being ignored when shrinking all players to 0. By modeling each of the β_j values as independent $N(\alpha_0 + \alpha_1(S_j), \sigma^2/\lambda)$, where S_j represents the number of shifts which player j participated, β_j values are shrunk towards players who played a comparable number of shifts. Table 3.3 shows 95% credible intervals for the two alpha parameters obtained by fitting model (1) with this adjusted hierarchical aspect. Figure 3.4 shows the effect of this adjustment, comparing the $\hat{\beta}_j$ estimated from the model fit with each $\beta_j \sim N(0, \sigma^2/\lambda)$ to those with each $\beta_j \sim N(\alpha_1 + \alpha_2(S_j), \sigma^2/\lambda)$. Estimates for players who played in a large number of shifts are minimally affected by the adjustment, but estimates for players who played in a small number of shifts can be strongly affected. This adjustment does improve β_j estimation as measured through test set accuracy. However, for a coach or front office, evaluating players by shrinking to average may be preferable.

Table 3.3 Median and 95% credible intervals for α_0 and α_1

Parameter	Posterior Median (95% Credible Intervals)
α_0	$-0.0021(-0.0027, -0.0015)$
α_1	1.8×10^{-6} $(1.3 \times 10^{-6}, 2.4 \times 10^{-6})$

Wins Added has multiple advantages over the simpler adjusted +/-. First, WA measures what is most important in assessing the value of a player: the player's effect on the probability of winning a game and consequently number of wins in a season. Second, not all equal changes in point difference should be given equal credit. For instance a shift with a change in score difference of +30 to +35 is not nearly as beneficial as a shift with a change of 0 to +5, especially in late game situations. Using the adjusted +/- approach these two shifts are equivalent in value, while using the Wins Added approach

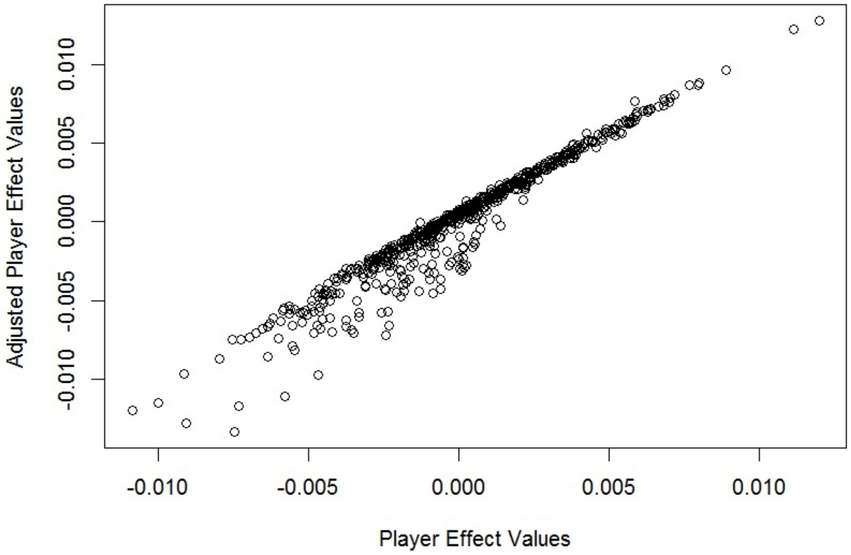


Figure 3.4 Comparing the β_j estimates from the model fit with a $N(0,\sigma^2/\lambda)$ (x axis) to those fit with a $N(\alpha_1+\alpha_2(S_j),\sigma^2/\lambda)$ (y axis).

the difference in win probability from +30 to +35 is approximately 0, while the difference from 0 to +5 could be quite large, especially in late game situations. A contrasting disadvantage is that, unlike point differential, change in win probability is an estimate of an unknown value, which introduces an additional source of error.

3.4 Discussion

In this paper, we summarized existing approaches and demonstrated a uniformly applicable methodology to estimate win probability in any head-to-head competition that performs well provided sufficient in-game training data are available. Comparing overall performance of the random forest approach to existing approaches is difficult because most are either proprietary or difficult to replicate or use to produce more than a few WP estimates. While constructing the forest is initially computationally intensive, once constructed, subsequent predictions are nearly instantaneous. Utilization of the R package *randomForest* also substantially simplifies the construction process. Adjustments would be necessary to generate win probability values in sporting events that are not head-to-head such as golf or many Olympic sports.

Across all sports, estimating win probability within a game consists of a series of observations all leading up to the same response. The random forest assumes these associated sequential observations are independent and stochastic. The independence assumption is clearly violated in that multiple observations all lead to the same response. This means the result from one game could factor in with a large set of observations. In order to ensure that predications are not typically based on the outcome of one or two games in the training data, we suggest a *nodesize* value much larger than the default of 5 used by Liaw and Wiener (2002). We attempted many adjustments to the random forest to account for the lack of independence, such as a version of the case specific random forest (Xu et al. (2016)), but they each yielded equivalent or worse test set accuracy while adding a layer of complexity. We also attempted random forests where the state of the game after each quarter/period was included in future predictions to account for the possibility of momentum, but these methods did not improve test set accuracy.

Regardless of the calculation methodology, we demonstrate some of the common uses of win probability and show how win probability can be used to evaluate the overall impact of NBA players by wins added to their respective teams. What's presented here constitutes just a few of the possibilities win probability values introduce. The paramount statistic in all sporting competition is wins, so accurate estimation of win probability and analyses designed to increase win probability and consequently number of wins are invaluable to sports analytics.

A. Variables in Each Random Forest

Sport	Variable	Descrption
NHL	Goal Difference	Difference in goals (Home - Away)
NHL	Total Goals	Total goals scored in the game
NHL	Team Quality	Team quality variable based on previous games
NHL	Time Remaining	Time Remaining in the game (seconds)
NHL	Shot Difference	Difference in shots on goal (Home - Away)
NHL	Shot Total	Total number of shots on goal
NHL	Power Play	skaters on each side (5 on 5, 5 on 4, etc.)
NBA	Score Difference	Difference in points (Home - Away)
NBA	Total points	Total points scored in the game
NBA	Team Quality	Team quality variable based on previous games
NBA	Time Remaining	Time Remaining in the game (seconds)
NBA	Possession	Inidicator variable for which team has possession
NBA	Total Possessions	Total number of possesssions in the game
NFL	Score Difference	Difference in points (Home - Away)
NFL	Total points	Total points scored in the game
NFL	Point Spread	Las vegas point spread
NFL	Time Remaining	Time Remaining in the game (seconds)
NFL	Down	The current down (1st, 2nd, 3rd, or 4th)
NFL	Yardline	Offensive yards from own goal line
NFL	Yards to Go	Yards to go for a 1st down
NFL	TO Offense	Offensive timeouts remaining
NFL	TO Defense	Defensive timeouts remaining

Bibliography

Asif, M. and McHale, I. G. (2014). In-play forecasting of win probability in one-day international cricket: A dynamic linear regression model. *INFO: Information Systems and Operational Research*, 52:2:39–50.

Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.

Buttrey, S. E., Washburn, A. R., and Price, W. L. (2011). Estimating NHL scoring rates. *Journal of Quantitative Analysis in Sports*, 7:3.

Deshpande, S. K. and Jensen, S. T. (2016). Estimating an nba player's impact on his team's chances of winning. *Journal of Quantitative Analysis in Sports*, 12:2:51.

Jensen, S. T., Shirley, K. E., and Wyner, A. J. (2009). Bayesball: A bayesian hierarchical model for evaluating fielding in major league baseball. *The Annals of Applied Statistics*, 3:2:491–520.

Kaplan, E. H., Mongeon, K., and Ryan, J. T. (2014). A markov model for hockey: Manpower differential and win probability added. *INFO: Information Systems and Operational Research*, 52:2:39–50.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomForest. *R News*, 2:3:2225–2236.

Lock, D. and Nettleton, D. (2014). Using random forests to estimate win probability before each play of an nfl game. *Journal of Quantitative Analysis in Sports*, 10:2:197–205.

Macdonald, B. (2012). Adjusted plus-minus for nhl players using ridge regression with goals, shots, fenwick, and corsi. *Journal of Quantitative Analysis in Sports*, 8:3:1–24.

Pettigrew, S. (2015). Assessing the offensive productivity of nhl players using in-game win probabilities. *9th Annual MIT Sloan Sports Analytics Conference*, 2:3.

Ryder, A. (2004). Win probabilities. a tour of win probability models for hockey.

Thomas, A. and Ventura, S. L. (2014). *nhlscrapr: Compiling the NHL Real Time Scoring System Database for easy use in R*. R package version 1.8.

Xu, R., Nettleton, D., and Nordman, D. J. (2016). Case-specific random forests. *Journal of Computational and Graphical Statistics*, 25:1:49–65.

CHAPTER 4. USING INFORMATION UNDERLYING MISSING DATA TO IMPROVE ESTIMATION OF NFL FIELD GOAL KICKER ACCURACY

Dennis Lock, Dan Nettleton, Jarad Niemi, and Casey Oliver

Abstract

We consider estimation and inference for generalized linear mixed models when each response value is missing with a probability that depends on the value of the linear predictor. To account for this type of informative missingness, we introduce and apply a new methodology that uses the relationship between indicators of missingness and the linear predictor to improve estimation of fixed effects and prediction of random effects. We demonstrate our approach by analyzing National Football League (NFL) field goal data, where a coach's decision on whether to attempt a field goal is related to the kicker's probability of making a field goal. Because the most skilled kickers are more likely than weaker kickers to be called on to attempt difficult kicks, our methodology is crucial for fairly assessing the quality of NFL kickers. Cross-validation and performance in specific examples are used to illustrate the advantages of our approach for handling informative missingness compared to standard approaches that rely only on observed kick data and incorrectly assume data are missing at random. Our work on NFL kicker evaluation is a special case of the general and pervasive problem of performance comparison when the degree of difficulty varies across performers and also contains indirect information about performance.

4.1 Introduction

Missing data are common in a wide variety of applications. When missing data are not missing at random, failure to account for missing data could result in biased parameter estimation and incorrect inference (Little (1982)). The majority of literature on adjusting for missing values focuses on missing covariates with a fully observed response. Ibrahim et al. (2005) summarize methodologies to account for missing values within covariates of a generalized linear model. Common practices are to impute the missing values using either an EM algorithm (Dempster et al. (1977)) or a Bayesian approach (Mason et al. (2012)).

Literature on methodologies to account for missing response values with observed covariates is more sparse. Copas and Farewell (1998) address "enthusiasm to respond" in survey data utilizing propensity scores (see Rosenbaum and Rubin (1983)) on whether or not the individual answered the question of interest to account for the effect of informative missingness in the response. Lipsitz and Ibrahim (1996) account for missing response values by constructing a missing data model where a missing value indicator depends on the possibly unobserved response value, as well as the fully observed covariates in the response model. The missing response values are treated as missing covariates in the missing value indicator model, and parameter estimation for the joint distribution of the response and missing value indicator is found using a version of the EM algorithm (Ibrahim (1990)).

We offer a new approach for addressing the challenge of missing response values in a generalized linear mixed-effects model when the probability of a missing response is associated with the value of the linear predictor. Our work is motivated by the problem of fairly evaluating National Football League (NFL) field goal kickers. Currently, the most common way to evaluate NFL kickers is with raw percentage of kicks made or field goal percentage (FG%). This metric is immediately improved upon by accounting for the difficulty of each field goal (Clark et al. (2013); Pasteur and Cunningham-Rhoads (2014)). Past research primarily incorporates kick difficulty by modeling successes and failures through a logistic regression (Berry and Berry (1985); Clark et al. (2013); Pasteur and Cunningham-Rhoads (2014)) with covariates (e.g., distance of the kick and some measure of field and weather conditions) that attempt to capture kick difficulty.

We argue that estimating kicker accuracy can be improved further by including information in a coach's decision not to attempt the field goal when a made field goal would be advantageous to the kicking team. Logically, a coach's decision whether or not to attempt a field goal should relate to the probability of success. If a coach decides not to attempt a field goal (opting instead to punt or attempt to gain a first down), the outcome is missing for the field goal kick that could have been attempted under the given conditions. The decision not to attempt a field goal, and subsequent missing value, may indicate that the coach perceives a low probability of success for that kicker in that situation based on information available to the coach (e.g., a kicker's performance in practice sessions) that is not explicitly contained in publicly available kicking data. Therefore, the missing outcomes associated with kicks not attempted are not reasonably modeled as missing at random.

To understand the problems that arise when missing data are ignored in the NFL kicker application, consider a kicker who is very accurate at short distances but cannot make long kicks. Such a kicker will take many short, relatively easy kicks but will not be called upon to attempt longer, more difficult kicks, even those that, if made, would help the kicker's team win. Examining only the data on field goals attempted will show a kicker with excellent performance for short kicks but provide no direct information about that kicker's performance for long kicks. Existing methods that ignore the information in missing data and focus only observed kick outcomes will tend to overestimate the kicker's ability to make long kicks because of the kicker's excellent performance for shorter kicks. In contrast, a kicker with a strong leg who is asked to take long and difficult kicks will tend to be inordinately penalized for long misses when using existing methods of analysis. This phenomenon is illustrated with specific examples in Section 4 of this paper.

Our proposed method is designed to account for information in coaches' decisions about whether to attempt field goals and to allow that information to contribute to the estimated probability of field goal success. The issues we face in our application are similar to those surrounding "enthusiasm to respond" in survey data (Copas and Farewell (1998)), where instead of "enthusiasm to respond" we have "enthusiasm to attempt." Since the enthusiasm to attempt may depend on variables outside of success probability, we model coaches' decisions such that additional covariates can be included. While this article illustrates our model in reference to field goal kickers, we believe the methodologies underlying the approach are applicable across a wider range of applications, such as evaluating surgeons who perform surgeries with different degrees of difficulty or evaluating products or services where an individual asked to give feedback may be less likely to respond when satisfaction is neither very low nor exceptionally high.

In Section 4.2, we present our proposed generalized linear mixed model for a response variable of interest and the indicator of response missingness. We explain the simplification of our general model for the special case of evaluating NFL kickers at the end of Section 4.2. In Section 4.3, we describe our NFL kicker dataset and explain its connection to the model of Section 4.2. In Section 4.4, we show results for our method and compare the new approach to previous analysis strategies. In Section 4.5, we discuss adjustments, extensions, and limitations and provide concluding remarks. The Appendix contains the details of our Bayesian approach to model fitting and inference.

4.2 A Model for Informative Missingness in Response Values

For each subject $k = 1,\ldots,K$, let x_{ik} and z_{ik} be vectors of covariate values associated with the ith of n_k observations. For $i = 1,\ldots,n_k$ and $k = 1,\ldots,K$, let m_{ik} equal 1 if y_{ik}, the response value for observation i and subject k, is missing, and let $m_{ik} = 0$ if y_{ik} is observed. Suppose

$$y_{ik}|\theta_{ik} \overset{ind}{\sim} f(\theta_{ik},\lambda), \tag{4.1}$$

where $f(\theta_{ik},\lambda)$ is an exponential dispersion distribution with parameters θ_{ik} and λ (Jørgenson (1987)). Examples of exponential dispersion distributions include Bernoulli, binomial, Poisson, Gaussian, inverse Gaussian, gamma, and beta distributions, where for single-parameter distributions, the dispersion-controlling parameter λ can be taken to be a known constant. For $i = 1,\ldots,n_k$ and $k = 1,\ldots,K$, define the linear predictor

$$h(\theta_{ik}) = x'_{ik}\beta + z'_{ik}v_k, \tag{4.2}$$

where $h(\cdot)$ is a known link function, β is an unknown parameter vector, and random effects vector $v_k \overset{ind}{\sim} N(0,\Sigma_v)$ for some unknown variance matrix Σ_v. Next, for $i = 1,\ldots,n_k$ and $k = 1,\ldots,K$, let

$$\omega_{ik} = (1, x'_{ik}\beta + z'_{ik}v_k, w'_{ik})', \tag{4.3}$$

where w_{ik} is a vector of values for covariates that, together with the value of the linear predictor $x'_{ik}\beta + z'_{ik}v_k$, could be associated with the probability of a missing response. Finally, suppose

$$m_{ik}|\pi_{ik} \overset{ind}{\sim} \text{Bernoulli}(\pi_{ik}), \tag{4.4}$$

where

$$g(\pi_{ik}) = \omega'_{ik}\alpha, \tag{4.5}$$

for some known link function $g(\cdot)$ and some unknown parameter vector α.

In applications involving this model, the primary goals may be to estimate β and predict v_k for all $k = 1,\ldots,K$. Note that because both β and v_k are involved in the distributions in (4.1) and (4.4), both the portion of the dataset involving the observed response and the portion of the dataset involving the unobserved response contribute information about β and v_k. Ignoring observations for which $m_{ik} = 1$ and analyzing only the subset of observations with $m_{ik} = 0$, using only the generalized linear mixed-effects model defined by (4.1) and (4.2), will suffer from loss of efficiency and potentially lead to invalid inferences.

In our application of interest, the subjects are NFL kickers. The n_k observations for the kth kicker correspond to in-game situations faced by the kth kicker's team where it may have been reasonable to call upon the kth kicker to attempt a field goal. (We discuss how such situations are determined in Section 4.3.) For the situation corresponding to observation i and kicker k, the missing indicator m_{ik} is 1 if no field goal was attempted and 0 otherwise. The response y_{ik} is 1 if the field goal was made (when $m_{ik} = 0$) or would have been made had it been attempted (when $m_{ik} = 1$). Likewise, y_{ik} is 0 for a missed field goal or a field goal that would have been missed had it been attempted. Because y_{ik} is binary, we choose $f(\theta_{ik}, \lambda)$ in (4.1) to be the Bernoulli distribution with success probability θ_{ik}. We choose the link function $h(\cdot)$ in (4.2) to be the standard normal cumulative distribution function $\Phi(\cdot)$, so (4.1) and (4.2) define a mixed-effects probit regression model. Similarly, we set $g(\cdot) = \Phi(\cdot)$ in (4.5). With these specifications, (4.1) through (4.5) simplify to the following informative missingness (IM) model for NFL kicker evaluation:

$$y_{ik}|\theta_{ik} \overset{ind}{\sim} \text{Bernoulli}(\theta_{ik}),\ \Phi(\theta_{ik}) = x'_{ik}\beta + z'_{ik}v_k,\ v_k \overset{ind}{\sim} N(0, \Sigma_v), \tag{4.6}$$

$$m_{ik}|\pi_{ik} \overset{ind}{\sim} \text{Bernoulli}(\pi_{ik}),\ \Phi(\pi_{ik}) = (1, x'_{ik}\beta + z'_{ik}v_k, w'_{ik})\alpha = \omega'_{ik}\alpha. \tag{4.7}$$

The covariates we select for x_{ik} and z_{ik} in (4.6) and w_{ik} in (4.7) are specified in Section 3.

4.3 The Dataset and Its Connection to the IM Model

The dataset analyzed in this paper was constructed from a subset of plays from the 2009, 2010, and 2011 NFL seasons. A play was included in the subset if and only if a field goal was attempted or (a) it was fourth down, (b) an attempted field goal would have been from a distance of no more than 76 yards, and (c) a made field goal would not have decreased the offense's win probability as estimated in Lock and Nettleton (2014). Criteria (a) and (b) select plays where attempting a field goal may be a reasonable decision. The 76 yard upper limit in (b) was selected because that distance matches the longest attempt in NFL history. Criterion (c) excludes plays where kicks were not attempted under circumstances where making a field goal would have reduced the offense's chances of winning. Opting not to attempt a field goal in such a situation provides no information about a coach's confidence in his kicker's ability to make a field goal; it is simply sound strategy. By excluding these plays from our dataset, we can reasonably assume that $m_{ik} = 1$ for a play in our dataset implies some doubt about the kicker's ability to make the field goal.

Applying criteria (a), (b), and (c) results in a dataset consisting of $n \equiv \sum_{k=1}^{K} n_k = 5706$ plays with $\sum_{k=1}^{K} \sum_{i=1}^{n_k} (1 - m_{ik}) = 2956$ plays where a field goal was attempted. The dataset includes $K = 41$ kickers. The fewest kicks attempted by a kicker in the dataset is $\min\{\sum_{i=1}^{n_k}(1 - m_{ik}) : k = 1, \ldots, K\} = 27$, and the most kicks taken by a kicker is $\max\{\sum_{i=1}^{n_k}(1 - m_{ik}) : k = 1, \ldots, K\} = 105$.

By far the most important covariate in our dataset is d_{ik}, the kick distance associated with kick opportunity i for kicker k. We define kick distance in the customary manner as 17 yards plus the distance between the line of scrimmage and the goal line. This is typically the same as the distance, measured parallel to the ground, between the point where the ball was kicked (or would have been kicked had a field goal been attempted) and the plane containing the crossbar and goal posts. For attempted kicks, the mean and standard deviation of kick distance is 36.6 and 10.3 yards, respectively. In addition to distance, we also have data on two categorical covariates: precipitation (with categories clear, rain, or snow) and field surface (with categories grass, synthetic, or mix).

Individual kick distributions among kickers can vary by distance, surface, and conditions. As an example John Carney had the shortest average attempt distance of 32.3, Garret Hartley had a shorter than typical average attempt distance of 35.4, Rob Bironas had a longer than typical average attempt distance of 39.3, and Sebastian Janikowski had the longest average attempt distance at 41.0. The full distribution of kicks attempts by kick distance for these kickers is presented in Figure 4.1. As an example illustrating the differences among kickers with regard to surface and conditions, 102 of Sebastian Janikowski's 105 kick attempts were on a grass surface (97%) and 95 of his 105 were in clear conditions (90%), while only 16 of Garret Hartley's 44 kick attempts were on a grass surface (36%) and all 44 were in clear conditions (kicking primarily in a dome).

To capture differences in difficulty among kick opportunities in our dataset, we use a six-dimensional

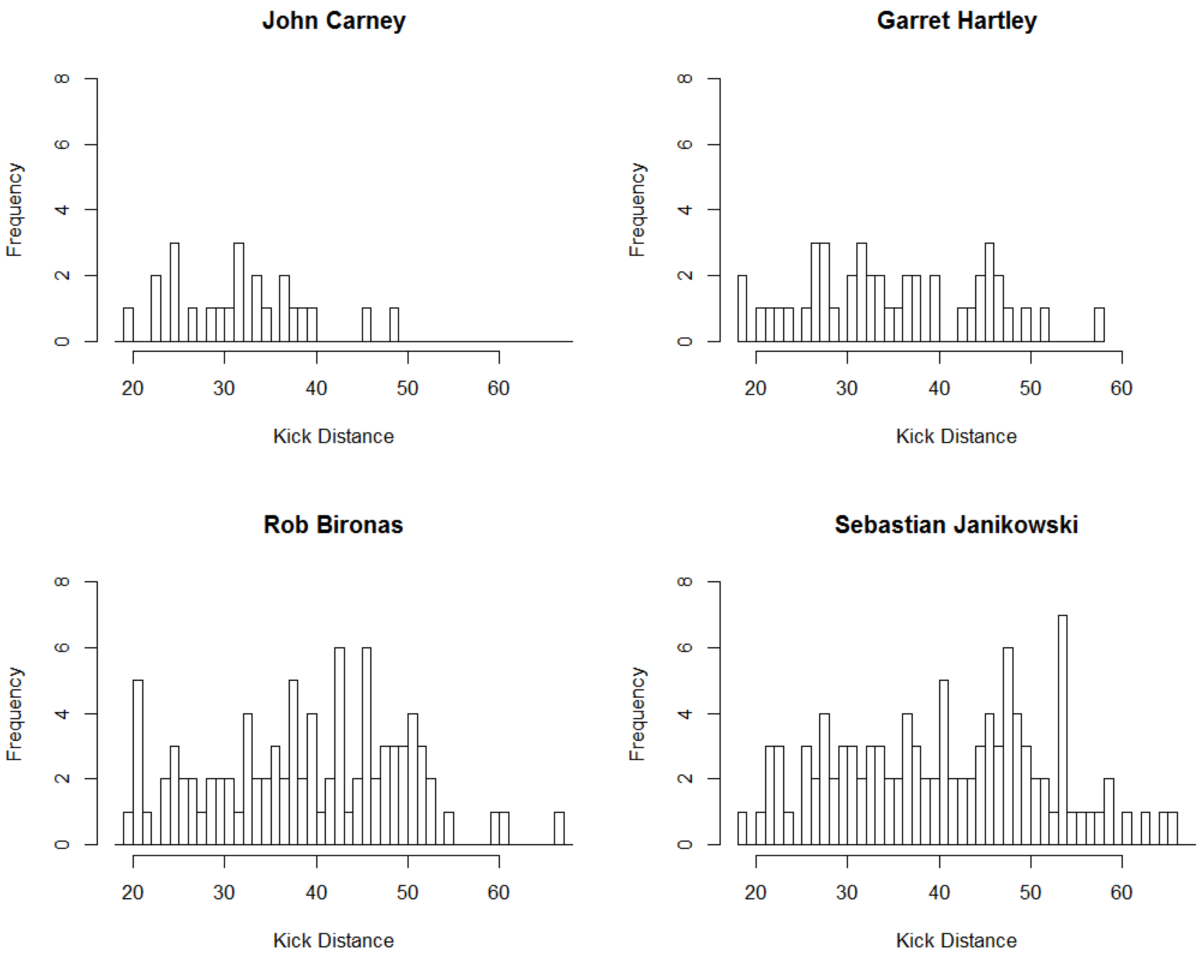


Figure 4.1 Comparing the distribution of attempts by attempt distance for John Carney, Garret Hartley, Rob Bironas, and Sebastian Janikowski

covariate vector x_{ik} in (4.6) and (4.7) with 1 and d_{ik} as the first two components, indicator variables for the precipitation conditions of rain and snow as the next two components, and indicator variables for the field surfaces synthetic and mix as the last two components. We set z'_{ik} to $(1, d_{ik})$ to allow for kicker-specific intercepts and slopes (i.e., coefficients on distance) in our probit regression model. These choices for x_{ik} and z_{ik} imply that the effect of distance on kick success probability is different for each kicker but that precipitation and surface conditions affect the probability of kick success in the same way for all kickers.

To assess the importance of making a field goal on the ith play for kicker k, we used the method from Lock and Nettleton (2014) to compute Δ_{ik}, the change in estimated win probability that would result from a made field goal. More specifically, the change in estimated win probability is

$$\Delta_{ik} = \text{WP}_{ik}(\text{Field Goal Made}) - \text{WP}_{ik}(\text{Facing Fourth Down}), \tag{4.8}$$

where $\text{WP}_{ik}(\text{Field Goal Made})$ is the estimated win probability that would result for the team considering a field goal attempt if that team were to attempt and make the field goal, and $\text{WP}_{ik}(\text{Facing Fourth Down})$ is the estimated win probability for the team with the ball facing fourth down with the option to punt, try to gain a first down, or to attempt a field goal. Large values of Δ_{ik} indicate that making a field goal in the given situation is very important, such as when the team contemplating a field goal attempt trails by 1 or 2 points with a few seconds remaining in the game. Positive values of Δ_{ik} near zero suggest that a made field goal would improve the offense's chances of winning the game but that at least one other option (punting or trying for a first down) might be nearly as good strategically.

Although coaches do not necessarily use Δ_{ik} explicitly when deciding whether to attempt a field goal, we believe that the information conveyed by Δ_{ik} is at least implicitly involved in any NFL coach's decision making process. Thus, we include Δ_{ik} as the first component of the vector w_{ik} in (4.7) so that the probability of a field goal attempt, $1 - \pi_{ik} = P(m_{ik} = 0)$, can vary with a made field goal's importance, as well as with the probability of a successful kick that is captured by the linear predictor $h(\theta_{ik}) = x'_{ik}\beta + z'_{ik}v_k$ in (4.7). As the final component of w_{ik} in (4.7), we include the product of $h(\theta_{ik})$ and Δ_{ik} to allow for interaction between the probability of making the field and the importance of the field goal. Such interaction is expected because as the importance of the kick increases, the decision to attempt the kick should depend to an increasingly stronger degree on the probability of making the kick.

4.4 Results

The posterior mean and 95% credible intervals for the β parameters in (4.6) and the α parameters in (4.7) are presented in Table 4.1. The top portion of Table 4.1 shows that the estimated effects of snow and rain are both negative, indicating that these conditions reduce the probability of a made field goal relative to kicking in clear conditions. The estimated effects from kicking off of synthetic turf or a mixture of surfaces are both positive, which implies that kicking from a natural grass surface is most challenging. However, the credible intervals for the coefficients of the snow, rain, and mixed surface indicator variables are wide due to few observed kicks under these respective conditions, so the sign of these coefficients cannot be reliably determined from our dataset. Within each kicker, success probability is most heavily affected by the distance of the kick, with posterior means for the effect of kick distance ranging from -0.051 to -0.095. All kickers are predicted to make a high percentage of kicks from very short distances regardless of conditions, where the league average is approximately 99%.

Table 4.1 Posterior means and 95% credible intervals for the β parameters in (4.2) and the α parameters in (4.5).

Variable	Parameter	Coefficient Mean	95% Credible Interval
Intercept	β_1	3.557	(3.304,3.816)
Distance	β_2	-0.069	(-0.075,-0.063)
Snow indicator	β_3	-0.076	(-0.511, 0.371)
Rain indicator	β_4	-0.132	(-0.296, 0.035)
Synthetic surface indicator	β_5	0.105	(0.017, 0.196)
Mixed surface indicator	β_6	0.039	(-0.218, 0.299)
Intercept	α_1	0.607	(0.488, 0.731)
Linear predictor for kick success	α_2	-1.055	(-1.188, -0.934)
Change in win probability	α_3	-3.044	(-4.278, -1.880)
Interaction	α_4	-7.376	(-9.478, -5.337)

According to the credible intervals in Table 4.1 for our α parameters, the probability of making the kick, the importance of the kick, and the interaction are all significantly related to the probability of taking the kick. This provides us with confirmation that a coach's decision on whether to attempt a field goal is influenced by the probability of making the kick as well as a kick's importance. Figure 4.2 illustrates the estimated relationship between the probability of taking the kick and the probability of making the kick for several changes in win probability. As we speculated in Section 3, the impact of varying success probability on the probability of attempting a kick becomes more pronounced as the importance of the kick (as measured by change in win probability) increases.

4.4.1 Performance Comparison

In this section, we compare the performance of our proposed informative missing method with two alternative analysis methods that both ignore information in the missing data. The first is a logistic

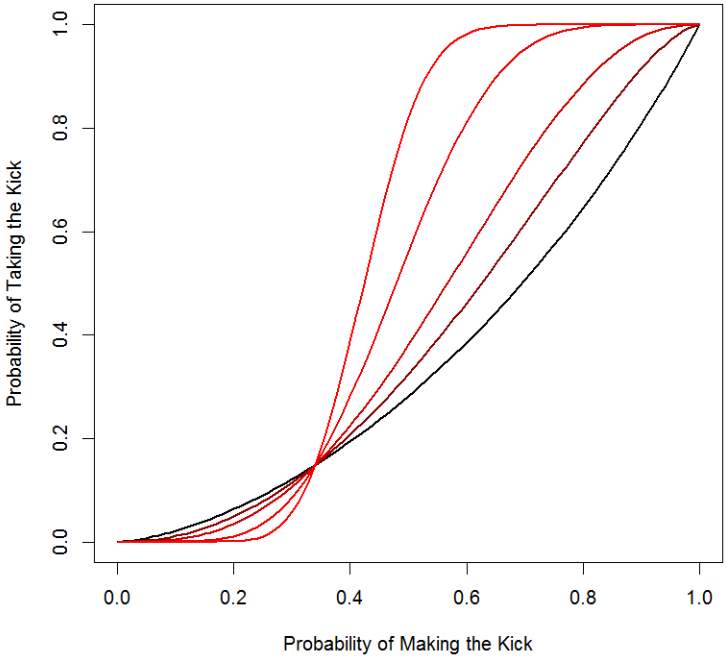


Figure 4.2 Estimates of the probability of taking the kick as a function of the probability of making the kick for kicks of varying importance, ranging from not very important (change in win probability $= 0.01$, Black) to very important (change in win probability $= 0.50$, Red).

regression approach where we assume

$$y_{ik} \overset{ind}{\sim} \text{Bernoulli}(\text{logit}(x_{ik}'\beta_k)),$$

with x_{ik} as defined in Section 3 and β_k treated as a kicker-specific vector of fixed effects. This method, which we refer to as the "separate logistic regressions" (SLR) method, is designed to mimic previously established methods to evaluate field goal kickers that ignore missing values and estimate logistic regression parameters separately for each kicker. To make inference results comparable with our method, we use a Bayesian approach with independent diffuse normal priors for each of the $6 \times K$ parameters in $\{\beta_k : k = 1, \ldots, K\}$ to fit the SLR model.

The second alternative method is a hierarchical probit regression (HPR) approach where we model the observed y_{ik} values exactly as proposed in Sections 2 and 3, except that we ignore the parts of the proposed model that involve missing data. In other words, we assume (4.6) but ignore the missing indicators $\{m_{ik} : i = 1, \ldots, n_k, k = 1, \ldots, K\}$ and (4.7). We again use a Bayesian approach model fitting and inference, with priors as described for the IM approach in the Appendix.

For each of the IM, HPR, and SLR methods, the standard deviations of posterior draws for the intercept and distance coefficient for each kicker were computed and averaged over kickers. These averages, presented in Table 4.2, show greatest precision for the IM approach and least precision for the SLR approach. This finding is consistent with the idea that borrowing information across kickers in the HPR approach can reduce uncertainty in estimates relative to SLR, and that extracting additional information from the missing data indicators in the IM approach can further reduce uncertainty compared to the HPR method. However, this assumes that the posterior standard errors are appropriate measures of uncertainty for all three methods.

To provide an alternative evaluation of the three methods, we use a cross-validation strategy where we apply each method to a training dataset to obtain parameter estimates, and then use the fitted models to predict outcomes of kicks in a test set. We set aside kick opportunities from a random sample of 150 games as a test set and used kick opportunities from the remaining 644 games as training data. The test set error from each of our three modeling strategies is presented in Table 4.3. Our proposed IM approach was more accurate than both the HPR and SLR methods with respect to mean absolute error (MAE) and mean squared error (MSE).

Table 4.2 Posterior standard deviations for intercept and distance coefficients (averaged over kickers) by method.

Parameter	IM	HPR	SLR
Intercept	0.223	0.324	0.590
Distance Coefficient	0.008	0.015	0.024

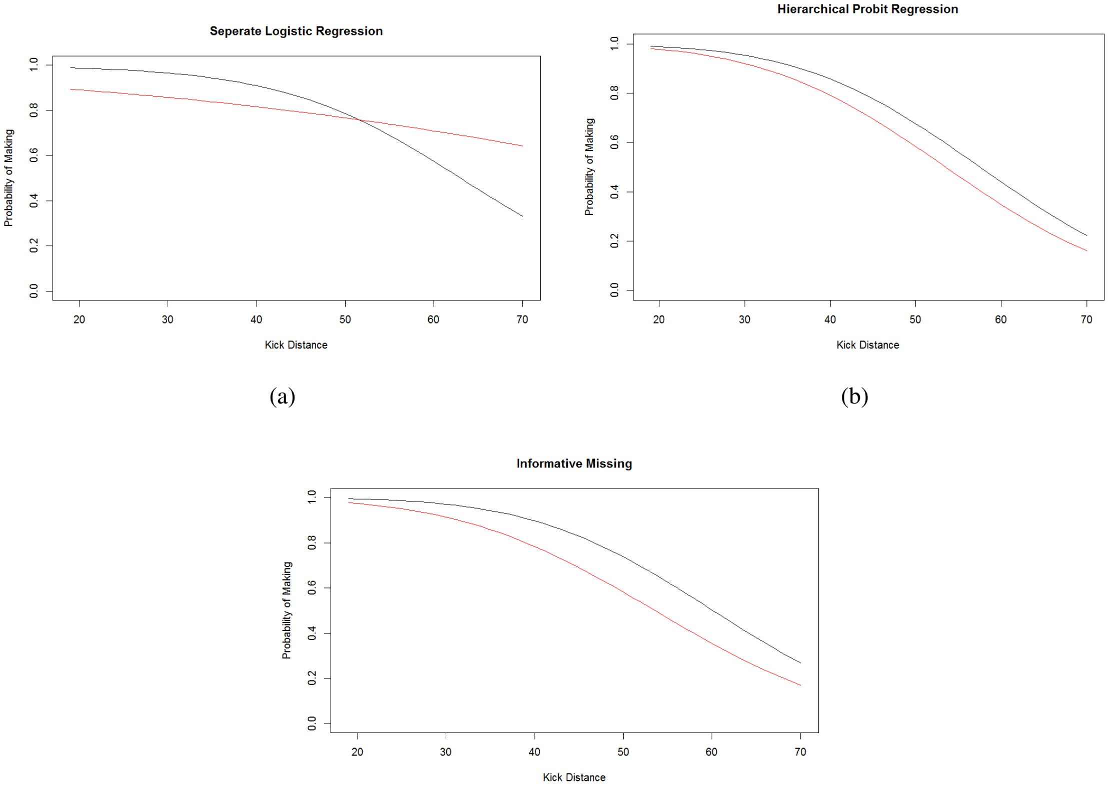
Table 4.3 Test set mean absolute error (MAE) and mean squared error (MSE) for each method.

Measure	IM	HPR	SLR
MAE	0.256	0.261	0.261
MSE	0.136	0.139	0.142

One advantage of our IM approach is that it can provide improved comparisons of kickers whose attempted kicks were very different. This advantage is likely not captured when comparing performance using test set error on observed kicks where extensive extrapolation is often not required. For example, a kicker who attempts primarily short kicks in the training data will typically attempt primarily short kicks in the test set. Thus, methods that do not extrapolate well are not necessarily penalized in evaluations based on cross-validation. However, extrapolation is required to compare a kicker who takes primarily short kicks with a kicker who is frequently called upon to try long kicks.

We demonstrate a comparison of kickers with highly different sets of attempts for each modeling strategy by observing estimated probabilities for two NFL kickers, Garret Hartley and Sebastian Janikowski. Over the three seasons included in our dataset, these two kickers performed similarly in terms of raw field goal percentage ($FG\%_{Hartley} = 0.84, FG\%_{Janikowski} = 0.86$), but Hartley attempted only 2 kicks at a distance greater than 50 yards, while Janikowski attempted over 20 kicks greater than 50 yards (Figure 4.1).

The predicted probabilities of making a field goal by distance in the most common conditions (no precipitation and grass surface) for Hartley (red) and Janikowski (black) from each of our three analysis methods are presented in Figure 4.3. Because Hartley has only 2 attempts greater than 50 yards, estimating success probability in this range with a separate logistic regression (Figure 4.3(a)) requires a dangerous amount of extrapolation and allows for the possibility of inaccurate predictions. Likewise, because Hartley had fewer kicks on a narrow range of distances, the separate logistic regression model also fits poorly at short distances, estimating that he would make 87% of kicks from the extra point distance when he actually succeeded on 50 of 51 (98%) of extra point kicks (which were not included in our field goal dataset). The hierarchical probit regression allows for sharing of information among kickers, which benefits prediction primarily for kickers with a small number of kicks on a narrow band of distances. This improves Hartley's estimated accuracy at shorter distances and shifts his estimated probabilities closer to a typical field goal kicker (Figure 4.3(b)), but still likely overestimates his ability at long distances. After accounting for the information in coaching decisions with our IM approach, we predict that Janikowski dominates Hartley at all distances, especially on longer kicks (Figure 4.3(c)). This not only makes sense based on coaching decisions and performance at similar distances, but matches conventional thinking about these two kickers. Prediction for Janikowski, where we have a large sample size across a wide band of distances, is fairly consistent across each model.



(a)

(b)

(c)

Figure 4.3 Comparing the estimated probability of making an attempt as a function of kick distance with Garret Harlety (red) and Sebastian Janikowski (black) for the (a) SLR, (b) HPR, and (c) IM analyses.

4.4.2 Kicker Evaluation

To arrive at a fair value to evaluate and compare all NFL kickers, we carry out the following steps for each kicker $k = 1, \ldots, K$ to obtain a figure of merit we refer to as adjusted FG%.

(a) Estimate the probability of success for every kick in 2011 as if they were taken by kicker k, and

(b) average these probabilities to generate one number that represents the estimated proportion of kicks made if kicker k was asked to take all the field goal attempts in the 2011 season (adjusted FG%).

By using adjusted FG%, all players are compared on the same set of kicks, and this set should accurately represent the typical distribution of kicks attempted in an NFL season. The top 5 kickers in the NFL

Table 4.4 Top 5 NFL kickers by adjusted FG%, with their raw FG% included.

Kicker	Raw FG%	Adjusted FG%
S. Janikowski	0.857	0.888
N. Novack	0.794	0.887
A. Vinatieri	0.881	0.867
N. Rackers	0.859	0.865
R. Bironas	0.889	0.862

Table 4.5 Comparing the success probability for S. Janikowski and G. Hartley at a distance of 25 yards and 50 yards on a grass surface with no precipitation, and comparing the two on overall adjusted FG% (posterior means with 95% credible intervals in parentheses).

Kicker	$P(y=1\|d=25)$	$P(y=1\|d=50)$	Adjusted FG%
S. Janikowski	0.99 (0.96,1.00)	0.73 (0.67,0.79)	0.89 (0.85,0.92)
G. Hartley	0.95 (0.89,0.99)	0.58 (0.44,0.71)	0.80 (0.71,0.87)
Difference	0.04 (-0.02,0.10)	0.15 (0.02,0.30)	0.09 (0.02,0.18)

during this time frame by adjusted FG% are presented in Table 4.4. Table 4.5 finalizes the comparison of Hartley and Janikowski by comparing the probability of making a short kick (25 yards), probability of making a long kick (50 yards), and overall adjusted FG%.

The previously established raw FG% is compared to the new Adjusted FG% for all 41 qualifying kickers in Figure 4.4. Overall the two measurements have a moderate positive relationship ($r = 0.508$), while some kickers have an adjusted FG% well above or below their raw FG%. Two kickers that stand out are Nick Novack and John Carney, who have similar raw FG% values ($FG\%_{Novack} = 0.794, FG\%_{Carney} = 0.783$) but very different adjusted FG% values ($aFG\%_{Novack} = 0.887, aFG\%_{Carney} = 0.691$). Carney, as a much older kicker in the twilight of his career, had the lowest percentage of attempts over 40 yards (8%) and a low percentage of kick opportunities attempted (47%), while Novak had a high percentage of kicks over 40 yards (53%) and the highest percentage of kick opportunities attempted (68%).

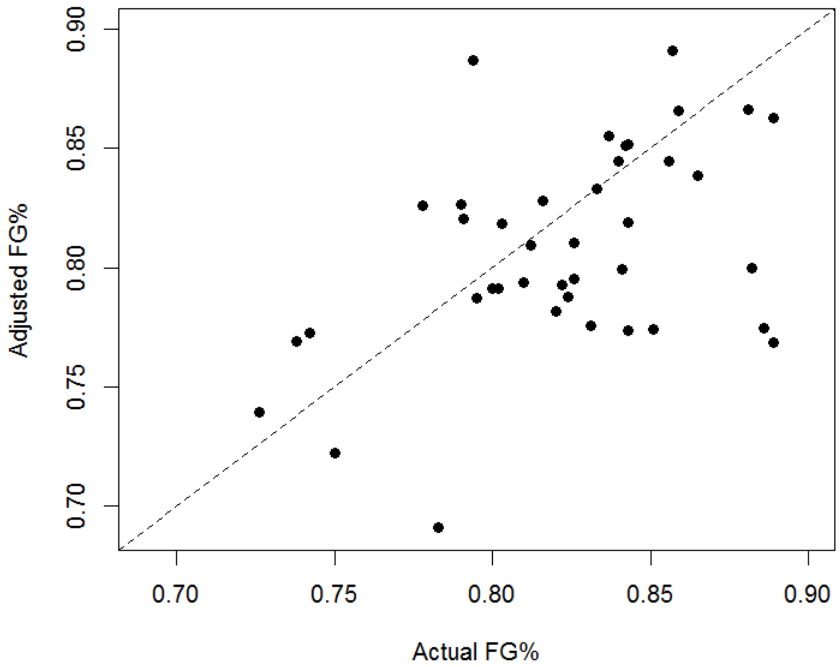


Figure 4.4 Comparing raw FG% to adjusted FG% for all 41 qualifying kickers (dotted line indicates equality between raw and adjusted FG%).

4.5 Discussion

The IM method presented here offers improvements over existing methods for evaluating NFL field goal kickers, especially for assessing accuracy at longer distances where the majority of missing values occur. For some kickers, probabilities estimated without accounting for missing values would greatly overestimate kicking accuracy at longer distances. In the terminology of Little and Rubin (2002), our application provides a clear example where data are "not missing at random" (NMAR) because, as we have demonstrated in Section 4, the probability of attempting a field goal depends on the probability of making a field goal. In this case, proceeding with analyses that assume data are missing at random (MAR) could lead to incorrect inferences.

In addition to the covariates listed in Table 4.1, we considered many other variables to account for the effect of conditions on each kick opportunity, including indoor or outdoor, home or away, elevation, wind speed (without direction, which is missing in our dataset), and temperature. We also tried a separate approach that replaced the variables elevation, indoor/outdoor, home/away, and surface with a unique indicator for each individual stadium. We arrived at the selected subset of variables in Table 4.1 by (a) using AIC to find the best fit probit model to estimate the probability of making the field goal on the observed kicks in the training data, and (b) using AIC to find the best fit probit model to estimate the probability of attempting the field goal on the kick opportunities in the training data, while conditioning on the probability of making the field goal as estimated from (a). Coefficients to account for each coach were also considered when estimating whether or not the kick was attempted, because some coaches are notably more or less aggressive than others. We did not pursue this idea thoroughly due to confounding that arises from kicker and coach combinations that matched all three seasons in our dataset. One season of information was excluded from one team that utilized two kickers (primarily, but not always, using one for shorter kicks and the other for longer kicks). Because this scenario is rare in the NFL, we opted not to extend our method to address data from teams with multiple kickers.

Our approach allows for flexibility in how the response variable is modeled without losing closed-form conditional distributions useful for computational estimation of posterior distributions. In the Appendix, we have detailed the fitting of our IM model specifically for a Bernoulli response and probit link functions because these choices were the most appropriate for our application. However, it is straightforward to alter our approach to cover other exponential dispersion distributions for the response variable and other choices for the link function.

We have demonstrated several advantages of our proposed modeling approach over current techniques for evaluating NFL field goal kickers. Due to its efficiency and flexibility, we believe that our proposed framework for incorporating information underlying missing values could be useful in a wide array of other applications.

A. Fitting the IM Model

In this section, we describe our approach for fitting the model defined in Section 4.2 to the NFL kicker data, using the variables described in Section 4.3 and terms as defined in Section 4.2. We begin by introducing, for each $i = 1, \dots, n_k$ and $k = 1, \dots, K$, a latent random variable

$$\upsilon_{ik} = x_{ik}'\beta + z_{ik}'v_k + \varepsilon_{ik},$$

which is the linear predictor in (4.2) plus $\varepsilon_{ik} \overset{ind}{\sim} N(0,1)$. We connect υ_{ik} to y_{ik} by the relationships

$$y_{ik} = 0 \iff \upsilon_{ik} \leq 0 \quad \text{and} \quad y_{ik} = 1 \iff \upsilon_{ik} > 0,$$

which leads to an equivalence with (4.6). Likewise, for each $i = 1, \dots, n_k$ and $k = 1, \dots, K$, we define a latent random variable

$$\psi_{ik} = \omega_{ik}'\alpha + e_{ik} \tag{4.9}$$

which is the linear predictor in (4.5) plus $e_{ik} \overset{ind}{\sim} N(0,1)$. We connect ψ_{ik} to m_{ik} by the relationships

$$m_{ik} = 0 \iff \psi_{ik} \leq 0 \quad \text{and} \quad m_{ik} = 1 \iff \psi_{ik} > 0,$$

which leads to an equivalence with (4.7). Recall that ω_{ik} in (4.9) includes the linear predictor $x_{ik}'\beta + z_{ik}'v_k$ as a component, so both β and v_k are associated with both υ_{ik} and ψ_{ik} (and thus with both y_{ik} and m_{ik}).

We specify the following diffuse prior distributions for the parameters our IM model. Each element of β and each element of α is assigned a $N(0,10^2)$ prior with mutual independence across all elements. We define the individual elements Σ_v, the variance of each random effect vector v_k by

$$\Sigma_v = \begin{bmatrix} \sigma_{v1}^2 & \sigma_{v12} \\ \sigma_{v12} & \sigma_{v2}^2 \end{bmatrix},$$

and fit diffuse uniform priors on σ_{v1}^2 and σ_{v2}^2, as well as a uniform(-1,1) prior on the correlation term $r_{v12} = \frac{\sigma_{v12}}{\sigma_{v1}\sigma_{v2}}$.

Our model and choice of priors is designed to facilitate efficient and accurate Gibbs sampling from the full conditional distributions for each of our parameters. Let y be the collection of all y_{ik} values, and define m, v, υ, and ψ analogously. Using $\phi(\cdot;\mu,\Sigma)$ to represent the pdf of the normal distribution with mean μ and variance Σ and using $p(\cdot)$ and $p(\cdot|\cdot)$ to denote a prior density and conditional posterior density, respectively, with arguments to such functions implying the random variables involved, we provide the full-conditional densities for α, β and v_k, for all k, below.

The full conditional distributions for α, β, and v_k for any k are

$$p(\alpha|\beta, v, \psi) \propto p(\alpha) \prod_{k=1}^{K} \prod_{i=1}^{n_k} \phi\left(\psi_{ik}; \omega_{ik}\alpha, 1\right)$$

$$\begin{aligned} p(\beta|v, \alpha, \Sigma_v, m, \upsilon, \psi) &\propto p(\beta) \\ &\times \prod_{k=1}^{K} \prod_{i=1}^{n_k} \left[\phi(\upsilon_{ik}; x'_{ik}\beta + z'_{ik}v_k, 1) \times (1(m_{ik}=0) + 1(m_{ik}=1))\right] \\ &\times \prod_{k=1}^{K} \prod_{i=1}^{n_k} \phi\left(\psi_{ik}; \omega'_{ik}\alpha, 1\right) \end{aligned} \tag{4.10}$$

$$\begin{aligned} p(v_k|\alpha, \Sigma_v, m, \upsilon, \psi) &\propto \phi(0, \Sigma_k) \\ &\times \prod_{i=1}^{n_k} \left[\phi(\upsilon_{ik}; x'_{ik}\beta + z'_{ik}v_k, 1) \times (1(m_{ik}=0) + 1(m_{ik}=1))\right] \\ &\times \prod_{i=1}^{n_k} \phi\left(\psi_{ik}; \omega'_{ik}\alpha, 1\right) \end{aligned} \tag{4.11}$$

The indicator functions in (4.10) and (4.11) are included because υ_{ik} holds no information regarding β or v_k if $m_{ik} = 1$. The full conditional distribution for each υ_{ik} is given by

$$\begin{aligned} \upsilon_{ik}|y_{ik} = 1, m_{ik} = 0, \beta, v_k &\sim N(x'_{ik}\beta + z'_{ik}v_k, 1), \text{ truncated at the left by } 0, \\ \upsilon_{ik}|y_{ik} = 0, m_{ik} = 0, \beta, v_k &\sim N(x'_{ik}\beta + z'_{ik}v_k, 1), \text{ truncated at the right by } 0, \text{ and} \\ \upsilon_{ik}|m_{ik} = 1, \beta, v_k &\sim N(x'_{ik}\beta + z'_{ik}v_k, 1). \end{aligned}$$

Similarly, the full conditional distribution for each ψ_{ik} is given by

$$\begin{aligned} \psi_{ik}|m_{ik} = 1, \beta, v_k, \alpha &\sim N\left(\omega'_{ik}\alpha, 1\right), \text{truncated at the left by 0, and} \\ \psi_{ik}|m_{ik} = 0, \beta, v_k, \alpha &\sim N\left(\omega'_{ik}\alpha, 1\right), \text{truncated at the right by 0.} \end{aligned}$$

The full conditional distribution for each element of Σ_v is

$$\begin{aligned} p(\sigma^2_{v1}|v, \sigma^2_{v2}, \sigma_{v12}) &\propto p(\sigma^2_{v1}) \prod_{k=1}^{K} \phi(v_k; 0, \Sigma_k) \\ p(\sigma^2_{v2}|v, \sigma^2_{v1}, \sigma_{v12}) &\propto p(\sigma^2_{v2}) \prod_{k=1}^{K} \phi(v_k; 0, \Sigma_k) \\ p(r_{v12}|v, \sigma^2_{v1}, \sigma^2_{v2}) &\propto p(r_{v12}) \prod_{k=1}^{K} \phi(v_k; 0, \Sigma_k). \end{aligned}$$

Sampling from the full conditional distributions for α is straightforward through conjugacy. The full conditional posterior distribution for β and v_k is a system of independent normal random variables. With independent normal priors for the elements of β and normally distributed v_k, the system results in closed form normally distributed full conditionals for β and v_k through conjugacy. We chose to define independent priors for each term of Σ_v separately, as opposed to a conjugate inverse-Wishart prior on the entire Σ_v covariance matrix because any attempt at a non-informative inverse-Wishart prior had a strong influence on the posterior correlation term. Full conditional draws for σ_{v1}, σ_{v2} and r_{v12} were independently obtained using an accept/reject algorithm.

With an ordinal response the latent variable υ can be adjusted by adding additional cut-points (Albert and Chib (1993)). With a normally distributed response, latent variable υ is replaced with response variable y in each of the conditionals. In both these instances, closed-form full conditional distributions for β and v_k parameters are maintained using the latent variable model on the missing values. The missing value indicator or binary response can also be adjusted to a logit rather than probit link by utilizing a scale mixture of normals (Holmes and Held (2006)).

Bibliography

Albert, J. H. and Chib, S. (1993). Bayesian Analysis of binary and polychotomous response data.

Berry, D. A. and Berry, T. D. (1985). The probability of a field goal: rating kickers. *The American Statistician*, 39.2:152–155.

Clark, T. K., Johnson, A. W., and Stimpson, A. J. (2013). Going for three: Predicting the likelihood of field goal success with logistic regression. *Sloan Sports Analytics Conference*.

Copas, A. J. and Farewell, V. T. (1998). Dealing with non-ignorable non-response by using an 'enthusiasm to respond' variable. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 161:3:385–396.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (methodological)*, pages 1–38.

Holmes, C. C. and Held, L. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis*, 1.1:145–168.

Ibrahim, J. G. (1990). Incomplete data in generalized linear models. *Journal of the American Statistical Association*, 85.411:765–769.

Ibrahim, J. G., Chen, M., Lipsitz, S., and Herring, A. (2005). Missing-data methods for generalized linear models: A compartive review. *Journal of the American Statistical Association*, 100:469:332–346.

Jørgenson, B. (1987). Exponential Dispersion Models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 49:2:127.

Lipsitz, S. R. and Ibrahim, J. G. (1996). A conditional model for incomplete covariates in parametric regression models. *Biometrika*, 83.4:916–922.

Little, R. J. (1982). Models for nonresponse in sample surveys. *Journal of the American Statistical Association*, 77.378:237–250.

Little, R. J. and Rubin, D. B. (2002). *Statistical Analysis with missing data*. John Wiley & Sons.

Lock, D. and Nettleton, D. (2014). Using random forests to estimate win probability before each play of an nfl game. *Journal of Quantitative Analysis in Sports*, 10:2:197–205.

Mason, A., Richardson, S., and Best, N. (2012). Two-pronged strategy for using DIC to compare selection models with non-ignorable missing responses. *Bayesian Analysis*, 7:1:109–146.

Pasteur, R. D. and Cunningham-Rhoads, K. (2014). An expectation-based metric for NFL field goal kickers. *Journal of Quantitative Analysis in Sports*, 10.1:49–66.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensatory score in observational studies for causal effects. *Biometrika*, 70.1:41–55.

CHAPTER 5. SUMMARY AND DISCUSSION

Chapters 2 and 3 demonstrated the use of the random forest to accurately estimate win probability values in head-to-head based competitions. Chapter 2 introduced the approach specifically on the NFL, and Chapter 3 expanded the approach to apply to other sports using the NBA and NHL as examples. Due the lack of observed in-game win probability values measuring performance using a traditional loss function on a test data set was not possible. Within Chapter 2 we introduced a new way to evaluate the performance of any win probability estimator, and within Chapter 3 we used the methodology to validate the performance of win probability estimates using the random forest on each of our 3 sports.

Although we achieved good performance using this random forest approach in all 3 of the sports examined there are many opportunities for additional research and expansion of the method. For example, our method is designed exclusively for head-to-head based competition, but with adjustment the random forest could be utilized to estimate win probabilities in competitions with multiple competitors such as golf or swimming. Another issue, which we examined but did not account for in the method presented here, is the lack of independence among the response variable in training data from sequential observations within the same game. We attempted to account for this issue in our methods but they did not improve performance; further research could identify a methodology that does.

Within Chapters 2 and 3 we also demonstrated several uses of estimated win probability values. These ranged from plots of win probability throughout a game to the evaluation of coaching decisions and player performances. Most notably, we used win probability to evaluate overall NBA player performance and to measure the importance of a made field goal in any 4th down opportunity to support our method in Chapter 4. Further research could be applied to expand the use of estimated win probability values across these 3 sports and all other sports.

Within Chapter 4 we introduced a new method to account for missing response values that are not missing at random in general linear mixed effect models. We applied the method to evaluate NFL field goal kickers while including information on when a field goal was not attempted despite reasonable cause for an attempt. By accounting for the missing values our method improved estimation of field goal kicker accuracy over a previously established logistic regression model and a comparable hierarchical probit model. Due to its flexibility, future research can explore applications of this method in other general linear mixed effect models where missing values of a response variable are related to the linear predictor. Possible examples include, evaluating NHL shooters in the shootout where the coach decides on his 3 best shooters from 20 available players or more generally, any survey where non-response bias is an issue.

We believe the work presented here is innovative and applicable research that furthers the field of statistics, and in particular statistics in sport.