Reversal of fortune: a statistical analysis of penalty calls in the National Hockey League

Jason Abrevaya Robert McCulloch *

January 2014

Abstract

This paper analyzes a unique data set consisting of all penalty calls in the National Hockey League between the 1995–6 and 2001–2 seasons. The primary finding is the prevalence of "reverse calls": if previous penalties have been on one team, then the next penalty is more likely to be on the other. This pattern is consistent with a simple behavioral rationale based on the fundamental difficulty of refereeing a National Hockey League game. Statistical modeling reveals that the identity of the next team to be penalized also depends on a variety of other factors, including the score, the time in the game, the time since last penalty, which team is at home, and whether one or two referees are calling the game. There is also evidence of differences among referees in their tendency to reverse calls.

*Jason Abrevaya is Professor of Economics, Department of Economics, The University of Texas at Austin, abrevaya@eco.utexas.edu. Robert E. McCulloch is Professor of Econometrics and Statistics, University of Chicago Booth School of Business, robert.mcculloch@chicagobooth.edu.

1 Introduction

Ice hockey is a unique game. Skating on ice makes the players incredibly fast and maneuverable. The boards, the net, the ice, and the players in their equipment combine with the speed to create dramatic collisions. For better or worse, the National Hockey League (NHL) (the major North American professional league) has a tradition of violence. As legendary owner of the Toronto Maple Leafs Conn Smythe said, "If you can't beat 'em in the alley, you won't beat 'em on the ice."[1] According to author Paul Gallico, hockey is "a fast body-contact game played by men with clubs in their hands and knives laced to their feet."[2]

Due to the speed of the game, the constant body contact, and the possibility of violence flaring up, hockey is extremely difficult to officiate. The *National Hockey League Official Rules* specifies a set of infractions for which players can be penalized. And, when an NHL referee calls a penalty on a player, that player is sent to the penalty box (usually for two minutes but sometimes longer) and his team plays with four skaters and a goalie instead of the usual five. The penalized team is said to "short-handed" and the team with the man advantage is "on the power play." If the team on the power play scores, the penalty ends. A game is sixty minutes long and is played in three twenty minute periods.

For each type of penalty, the NHL rulebook contains a specific description of what constitutes an infraction. As one example, a "holding" penalty (which accounts for 12% of penalty calls) is defined in the rulebook as follows:

> *A minor penalty shall be imposed on a player who holds an opponent by using his hands, arms or legs. (Note: A player is permitted to use his arm in a strength move, by blocking his opponent, provided he has body position and is not using his hands in a holding manner when doing so.)*

As another example, a "hooking" penalty (11% of calls) is described as follows:

> *Hooking is the act of using the stick in a manner that enables a player to restrain an opponent. (Note: When a player is checking another in such a way that there is only stick-to-stick contact, such action is not to be penalized as hooking.)*

[1] Source: `www.legendsofhockey.net`.

[2] Source: `www.hockey.tribute.com/77-famous-ice-hockey-quotes.html`.

Such "rules" are obviously open to interpretation and, therefore, lead to the referee having a great deal of discretion over what should constitute a penalty. Players *expect* to get away with some amount of cheating. The referee is faced with a constant flow of infractions imbedded in fast paced violent action. The referee calls the egregious infractions he sees and a subset of the rest to control the game and avoid blame. According to long-time NHL coach Ken Hitchcock, "there could probably be a penalty call on every NHL shift."[3]

This situation is somewhat analogous to that of a highway patrol officer. Since most people do not obey the posted speed limit, it is impossible for the officer to cite every driver for a violation. Instead, the officer will pull over the most egregious offenders. Even in cases where there are no blatant offenders, however, it makes sense for the officer to occasionally pull over a mild offender in order to have a deterrent effect on other drivers and keep things under control. For the NHL referee, the easy part of the job is to call penalties on clear offences (a fighting penalty); the more difficult task is to pick and choose when to call penalties for minor infractions like hooking, holding, and interference. If the referee makes "too many" calls, he will be criticized for not "letting the players play." If he makes too few, he risks losing control of the game.

Penalty calls can have a large effect on the outcome of the game. The short-handed team is far more likely to allow a goal to its opponent and almost certain not to score. (We are, of course, not the first to document this fact; see, for example, Table 2 of Beaudoin and Swartz (2010).) Thus, in addition to enforcing the rules and keeping the game under control, the referee is also forced to consider the issue of fairness. In particular, a referee does not want to be seen as unduly influencing the outcome of the game. He will be blamed for the loss by the loser.

One approach to appearing to be fair is to avoid repeated penalty calls on the same team, an issue that this paper analyzes in great detail. This method enables the referee to control the game and not be severely criticized. In commenting on a game between the Calgary Flames and the Edmonton Oilers, announcer Glen Healy perfectly captured this idea in saying: "Referees are predictable. The Flames have had three penalties, I guarantee you the Oilers will have three."[4] Put in a negative light, people sometimes refer to this method as "make-up calls," whereby a referee makes a bad call on one team and fixes it by making a subsequent bad call on the other team. Alternatively, it may also be viewed as a reasonable way of

[3]Source: `Toronto Star, February 8, 2004`. A "shift" is a player's turn on the ice and generally lasts less than a minute.

[4]Source: `Hockey Night in Canada broadcast, October 25, 2003`.

handling a very difficult situation. In connection with the fairness issue, referees may be influenced (consciously or subconsciously) by the home crowd. Beaudoin and Swartz (2010) note that road teams are called for more penalties than home teams, and we will investigate in this paper whether the home team's identity additionally affects the likelihood of a make-up call.

While Healy's "guarantee" is a bit strong of course, this paper finds strong evidence for his basic hypothesis. If previous penalties have been on one team, then the next penalty is more likely to be on the other. While this finding will not be surprising to sports fans in general, this study is the first to document and quantify the phenomenon. Statistical modeling also reveals that the identity of the next team to be penalized depends on a variety of other factors, including the score, the time in the game, the time since the last penalty, which team is at home, and whether one or two referees are calling the game.

Allen (2002), Levitt (2002), and Heckelman and Yates (2003) have previously examined penalty calls in the NHL. They used data from the 1998-99 and 1999-2000 seasons when the NHL experimented with the number of referees, using either one or two referees in a game. These papers study the referee effect with the idea that more referees may be able to detect more infractions and their goal is to detect all infractions (like a homicide detective). Using analysis analogous to the economic approach to crime in society (Becker (1968)), these papers discuss how players may adapt their behavior given a change in the probability of detection. These economic studies posit simplistic models in which the probability of an infraction being committed and the probability of detection (conditional on an infraction) are constant parameters. Under the implicit assumption that infractions and detections are i.i.d. events, aggregate data (such as number of penalties called) are used in order to test their theories. Only the change in referee system (one referee to two referees) is presumed to affect the probabilities of infraction and detection. In contrast, this study focuses upon a dynamic statistical analysis and examines how referee behavior is affected by specific game situations. Overall, the practical significance of our results suggest that previous studies have oversimplified matters greatly by failing to account for the complex nature of referee behavior.

The paper is organized as follows. Section 2 describes the unique data set of NHL penalty calls and briefly summarizes the frequency and timing of calls. Section 3 examines the frequencies of the *sequences* of penalty calls. This section documents a dramatic tendency for penalties to "reverse": when one team has had more recent penalty calls on them, it becomes far more likely that the next penalty call will be on the other team. Section 4 provides ex-

tensive statistical modeling of the reverse-call probability using a wide range of conditioning variables. Several different approaches are considered, but due to the many interactive effects that influence penalty calls, we find that the most flexible methods (here, the Bayesian additive regression tree (BART) and boosting) perform the best. The results confirm that various game situations (score, time, time since last penalty, etc) and characteristics (home team, one versus two referees, etc) have extremely significant effects on the reverse-call probability. Section 5 investigates differences in penalty calling amongst individual referees, and Section 6 concludes.

2 The Data

The data for this paper were taken from individual game boxscores in the on-line archive of the *USA Today* (`www.usatoday.com`), specifically for games played from the 1995–96 season through the 2001–02 season. We focused on these years because it covers the time period when the league switched from using one referee to using two referees. As discussed in Section 1, previous studies have examined the effect of switching to two referees. Up until the 1997-98 season, all NHL games were officiated by just one referee. The league then experimented with a two-referee system in 1998-99 (about 20% of games) and 1999-2000 (about 60% of games). Beginning with the 2000-01 season, almost all games have used two referees. Since our basic story focuses on the viewpoint of the referee, we wanted to be able to study the effect of the switch and have some data from games with just one referee.

Since each game boxscore was stored on a separate webpage in the *USA Today* archive, a "web crawler" program was used in order to download each boxscore as an individual `html` file. The `html` files were then converted to text files, each of which was processed by a C++ computer program to retrieve all of the relevant game information. Appendix A contains an example of an original boxscore that was processed by the C++ program. There is a large amount of data contained in each boxscore, including the teams playing, details (including time) on each goal scored and each penalty call, and the identity of the referee(s). The resulting dataset consists of 7,821 regular-season games over seven seasons, which represents 98.8% of the 7,913 regular-season games that were played. A small fraction of the games were not included because the boxscores were not available or could not be processed.

Penalty calls are very common in NHL games. In the 7,821 games in our sample, there were only five games with no penalties called. There were a total of 92,348 penalties called

(an average of 11.8 penalties per game). For the analysis in this paper, only "non-matching penalties" are considered. A "matching penalty" occurs when the referee(s) calls multiple penalties at the same time, with the same number of penalties called on both teams. Although we view matching penalties as another way for referees to control the game and exhibit fairness, only non-matching penalties lead to a change in playing strength for the two teams. For this reason, if multiple penalties are called by the referee(s) at the same time, a "penalty" observation is included in the sample only if a different number of penalties are called on each team.[5] When matching penalties are dropped, a total of 66,030 (non-matching) penalties remain in the sample. For the remainder of the paper, we will use the term "penalty" to denote a non-matching penalty.

To give a sense of when penalties are called during the course of an NHL game, Figure 1 displays a minute-by-minute histogram of penalty calls. A game consists of three 20-minute periods. Overtime play, which occurs at the end of a tied game, is omitted from the figure.

The first and second periods appear quite similar, with the number of calls increasing as the period goes on. The third period is different in two respects: (1) there are far fewer calls than in the first two periods, and (2) the number of calls first increases and then decreases. Both of these facts are consistent with the theory that it's more difficult to make a penalty call when the game is on the line. Apparently, in the third period the referee "puts away the whistle" and "lets the players play." In each of the three periods, there is an increase in calls in the final minute of play (exhibited by the slight spikes at the minutes 20, 40, and 60). The spike at the end of the game can be explained by players who commit infractions once the game's outcome is no longer in question, but the spikes in the first two periods are a bit more puzzling.[6]

3 Penalty-call Sequences

A striking feature of the penalty-call data is the frequency with which calls alternate — that is, if a team is called for a penalty, it is more likely that the opposing team will get called for the following penalty. The term *reverse call* will be used to describe a penalty call on a team

[5]For example, looking at the sample boxscore given in the Appendix A, matching penalties occurred at 17:07 of the first period and 0:58 of overtime. For this game, the four matching penalties would not appear in the analysis sample.

[6]One possible explanation is that penalty calls near the end of a period are less costly since the end of the period temporarily breaks up the power play.

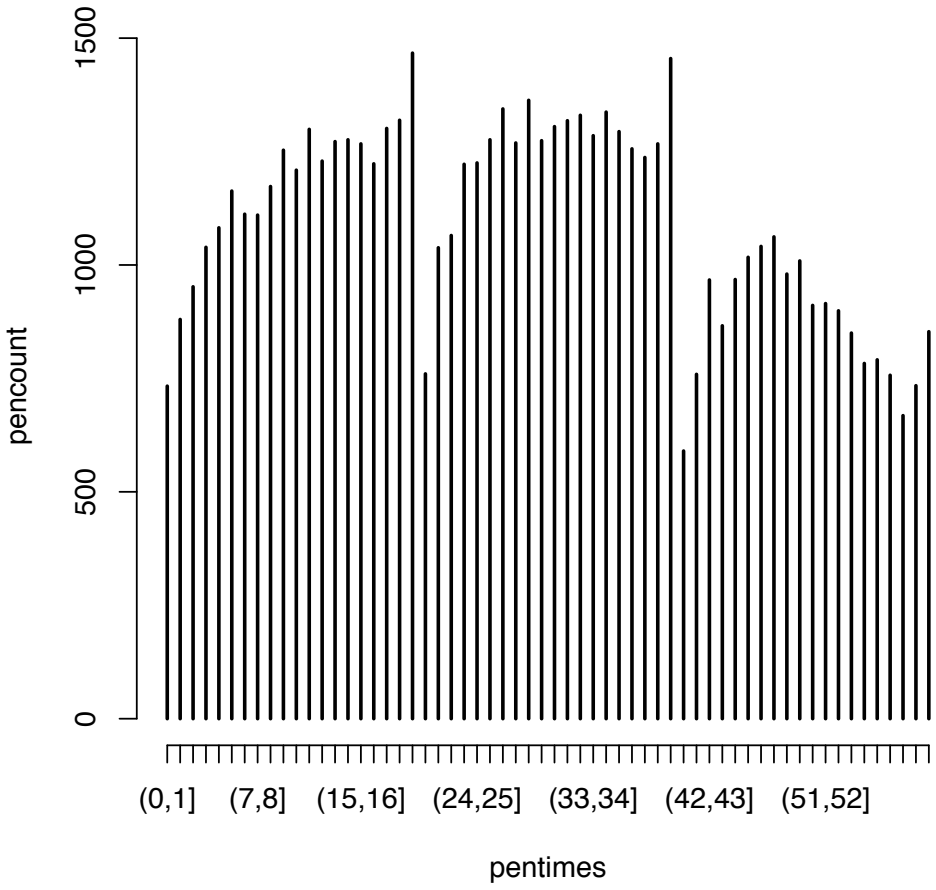

Figure 1: *Frequency of penalty calls, one-minute intervals.*

that was not the last team penalized. The overall fraction of reverse calls in our sample is 58.9%.

To get a more complete picture of reverse calls, we look at the sequence in which penalties are called. Table 1 reports the frequencies of every possible sequence of penalties. These frequencies are reported for the first two penalties, the first three penalties, and so on through the first six penalties of the game. Note that the number of games with at least six penalties (6,608) is still a significant fraction of the full sample of 7,821 games. To simplify matters, we label the team called for the first penalty of the game as Team A. As such, every sequence begins with an "A." For the first two penalties of the game, the sequence "AB" indicates that the second penalty was a reverse call (which occurs 63.4% of the time). There are several interesting features of the sequence frequencies. First, a naïve model of penalty calls as totally random events (i.e., coin flips) is strongly rejected by the frequencies. For instance, if calls were i.i.d. Bernoulli with $p = 0.5$, the frequencies for the "First 3 Penalties" column should all be close to 25%. Second, the modal frequency for each sequence length is the sequence associated with *perfectly alternating calls*. For the 6-penalty sequence, the most likely call sequence is "ABABAB," corresponding to five reverse calls in a row. Third, as would be expected given the prevalence of reverse calls, the least likely sequence (for each sequence length) is the one in which the same team is called for every penalty. Fourth, a general pattern in Table 1 is that the sequences with more alternating calls and less inequity in the total number of calls tend to have higher observed frequencies.

Viewing the frequencies from Table 1 in a slightly different way, Table 2 reports the frequencies of reverse calls conditional on the sequence of the prior penalty calls. As always, a reverse call corresponds to the most recent penalty. For instance, for the third penalty call of the game, the possible sequences of prior calls are "AA" and "AB." From Table 2, the probability of a reverse call after an "AA" sequence (68.6%) is much higher than the probability of a reverse call after an "AB" sequence (52.9%). To make the table easy to read, the sequences have been sorted based on the conditional probabilities in descending order. For the fourth penalty, a reverse call is most likely when a team was called for the first three penalties of the game (a sequence of "AAA"). The conditional probability in this case is 75.9%, quite a bit higher than the probability of a reverse call on the second penalty of the game. In general, the highest conditional probabilities are associated with sequences in which the last two or three calls were made on a single team. For the sixth penalty, the three highest reverse-call probabilities correspond to the sequences "ABBBB" (last four calls on Team B), "ABAAA" (last three calls on Team A), and "AAAAA" (all five calls

Table 1: Penalty Sequence Frequencies

First 2 Penalties ($n = 7,806$)		First 3 Penalties ($n = 7,756$)		First 4 Penalties ($n = 7,602$)		First 5 Penalties ($n = 7,264$)		First 6 Penalties ($n = 6,608$)	
Sequence	Freq.	Sequence	Freq.	Sequence	Freq.	Sequence	Freq.	Sequence	Freq.
AB	63.4%	ABA	33.5%	ABAB	19.5%	ABABA	10.1%	ABABAB	6.1%
AA	36.6%	ABB	29.9%	ABBA	18.3%	ABBAB	9.8%	ABBABA	5.9%
		AAB	25.1%	ABAA	14.1%	ABABB	9.5%	ABABBA	5.7%
		AAA	11.5%	AABB	12.6%	ABAAB	9.2%	ABBAAB	5.3%
				AABA	12.4%	ABBAA	8.6%	ABAABA	4.8%
				ABBB	11.7%	ABBBA	7.7%	ABAABB	4.3%
				AAAB	8.7%	AABAB	7.6%	ABBBAA	4.1%
				AAAA	2.8%	AABBA	6.8%	AABABB	4.0%
						AABBB	5.9%	ABABAA	3.9%
						AAABB	5.0%	ABBABB	3.8%
						ABAAA	4.8%	AABBAB	3.8%
						AABAA	4.8%	ABABBB	3.7%
						ABBBB	3.9%	AABABA	3.6%
						AAABA	3.7%	ABBBAB	3.6%
						AAAAB	2.0%	AABBBA	3.6%
						AAAAA	0.8%	ABBAAA	3.4%
								ABAAAB	3.2%
								AABAAB	3.1%
								AABBAA	3.1%
								ABBBBA	2.9%
								AAABBA	2.6%
								AAABAB	2.4%
								AABBBB	2.4%
								AAABBB	2.3%
								AABAAA	1.6%
								ABAAAA	1.4%
								AAAABB	1.3%
								AAABAA	1.3%
								ABBBBB	1.2%
								AAAABA	0.7%
								AAAAAB	0.5%
								AAAAAA	0.2%

Team A denotes the team called for the first penalty in the game. Team B is the other team.

on Team A). The lowest probability corresponds to the sequence "AAAAB". Even though the last penalty was on team B, the referee remembers that he gave team A a run of four penalties before that and still has to make it up to B.

Table 2: Reverse-Call Frequencies Conditional on the Sequence of Previous Calls

2nd Penalty		3rd Penalty		4th Penalty		5th Penalty		6th Penalty	
Sequence of Prior Calls	Freq. of Reverse Call	Sequence of Prior Calls	Freq. of Reverse Call	Sequence of Prior Calls	Freq. of Reverse Call	Sequence of Prior Calls	Freq. of Reverse Call	Sequence of Prior Calls	Freq. of Reverse Call
A	63.4%	AA	68.6%	AAA	75.9%	AAAA	72.5%	ABBBB	71.4%
		AB	52.9%	ABB	61.1%	ABBB	66.1%	ABAAA	70.4%
				ABA	58.0%	ABAA	65.7%	AAAAA	68.8%
				AAB	49.5%	AABA	61.2%	AAABA	65.7%
						AABB	53.6%	AABAA	65.3%
						ABBA	53.3%	ABABA	61.2%
						ABAB	51.5%	ABBAB	60.7%
						AAAB	42.5%	ABBAA	60.7%
								ABABB	60.4%
								AABBB	60.4%
								AABBA	55.4%
								AAABB	52.9%
								ABAAB	52.8%
								AABAB	47.9%
								ABBBA	47.3%
								AAAAB	36.8%

Team A denotes the team called for the first penalty in the game. Team B is the other team.

To investigate the extent to which penalty calls even out by the end of the game, we computed the absolute value of the difference between the number of penalties called on the two teams for each game. In 20% of the games, the two teams receive the same number of penalties. In 34% of the games, the two teams differ by exactly one penalty. In 25% of the games, the two teams differ by exactly two penalties. With only 21% of games having a penalty difference greater than two, these summary statistics suggest an evening-out phenomenon consistent with the reverse-call prevalence seen in Tables 1 and 2.

4 Modeling Reverse Calls

In Table 2 we see that the frequency of a reverse call depends strongly on the prior penalty calls. In this section we look for other characteristics of the game situation which could affect the identity of the team which gets the next penalty. For example, we shall see that the score is an important factor. It may be hard for a referee to give a team a penalty if that team had the last two penalties. It may be *really hard* to give a team a penalty if they had the last two penalties *and* they are losing the game!

For every penalty in a game after the first one, we let the variable `revcall` be equal to 1 if the current penalty and last penalty are on different teams and 0 otherwise. Our goal is to model reverse calls by estimating p(`revcall`=1 | x) where x denotes variables capturing the game situation at the time the penalty is called. Table 2 does exactly this with x consisting solely of previous values of `revcall`.

Table 3 summarizes the variables used in our analysis. The first variable is our dependent variable `revcall` and the rest of the variables are the covariates making up x. This table separates the covariates into three categories: (1) indicator variables, (2) categorical variables, and (3) numeric ("other") variables. The only categorical variable is `season`, which takes on seven possible values corresponding to the number of seasons in our sample. The last eight variables listed in Table 3 (labeled `gf1` through `pa2`) are controls for the overall performance characteristics of the two teams. The `gf` and `ga` variables measure a team's offensive and defensive ability, respectively, whereas the `pf` and `pa` variables measure a team's propensity to commit and draw penalties, respectively. The analysis is based on a sample of 57,883 penalty calls. The sample size is the original 66,030 observations less the 7,816 first-penalty observations less the 331 overtime penalties.

The variables `inrow2`, `inrow3`, and `inrow4` provide information about which team got the each of the last four penalties. Thus, seeing how p(`revcall` | x) depends on these x covariates looks for the same kind of effects as Table 2. The potential for the other variables to affect the probability of a reverse call is intuitive. For example, in Section 4.4.3 we shall see that when the last two penalties were on the same team (`inrow2`), that team is losing the game (`goaldiff`), it has not been long since the last penalty (`timebetpens`), it is early in the game (`timeingame`), and that team is the home team (`home`), the probability of a reverse call is estimated to be 80%!

Table 3: Variable Descriptions

Variable	Description	Mean	Min	Max
Dependent variable				
`revcall`	1 if current penalty and last penalty are on different teams	0.589	0	1
Indicator-Variable Covariates				
`ppgoal`	1 if last penalty resulted in a power-play goal	0.157	0	1
`home`	1 if last penalty was called on the home team	0.483	0	1
`inrow2`	1 if last two penalties called on the same team	0.354	0	1
`inrow3`	1 if last three penalties called on the same team	0.107	0	1
`inrow4`	1 if last four penalties called on the same team	0.027	0	1
`tworef`	1 if game is officiated by two referees	0.414	0	1
Categorical-variable covariate				
`season`	Season that game is played (e.g., 1995 for 95-6 season)		1995	2001
Other covariates				
`timeingame`	Time in the game (in minutes)	31.44	0.43	59.98
`dayofseason`	Number of days since season began	95.95	1	201
`numpen`	Number of penalties called so far (in the game)	5.76	2	21
`timebetpens`	Time (in minutes) since the last penalty call	5.96	0.02	55.13
`goaldiff`	Goals for last penalized team minus goals for opponent	-0.02	-10	10
`gf1`	Goals/game scored by the last team penalized	2.78	1.84	4.40
`ga1`	Goals/game allowed by the last team penalized	2.75	1.98	4.44
`pf1`	Penalties/game committed by the last team penalized	6.01	4.11	8.37
`pa1`	Penalties/game by opponents of the last team penalized	5.97	4.33	8.25
`gf2`	Goals/game scored by other team (not just penalized)	2.78	1.84	4.40
`ga2`	Goals/game allowed by other team	2.78	1.98	4.44
`pf2`	Penalties/game committed by other team	5.96	4.11	8.37
`pa2`	Penalties/game by opponents of other team	5.98	4.33	8.25

4.1 Model Choice

To simplify the notation a bit, we shall use p(`revcall` | x) to denote p(`revcall`=1 | x), the probability that a reverse call is made given the covariates x. The most common approach to estimating p(`revcall` | x) is linear logistic regression (log odds is a linear function of x). However, given that certain game situations may be particularly stressful for the referee(s), we might expect non-linearities and interactions in p(`revcall` | x). In addition, our data set is "big" enough that we can hope to uncover such relationships.

In order to uncover a potentially complex relationship we followed a common practice in data-mining and tried several predictive modeling approaches. The different approaches are assessed by their out-of-sample predictive performance. In addition to linear logistic regression we tried decision trees, random forests, boosting, and Bayesian Additive Regression Trees (BART). Over the past several years these methods have been shown to be extraordinarily powerful in both the statistical and machine learning literatures. For decision trees, random forests and boosting, the reader is referred to Chapter 8 of James, Witten, Hastie, and Tibshirani (2013); for BART, see Chapter 16 of Murphy (2012).

Details of our assessment of the out-of-sample predictive performance are in Appendix B. The best performing approaches were boosting and BART. We chose to present results based on BART because it conveniently gives measures of uncertainty (posterior intervals) rather than just point estimates and is more automatic in uncovering the level of interaction. In Section 4.4 we shall see that there are interesting interactions.

A basic problem with flexible fitting procedures such as boosting and BART is that the form of the fitted model is necessarily complex and hence difficult to interpret. Consequently, we present the results from a simple decision tree (Section 4.2) and linear logistic regression (Section 4.3) because they are relatively easy to understand and hopefully familiar to many readers even though their fit is inferior to that of BART. See Appendix B for fit comparisons. In Section 4.4 we present the BART results, but our presentation is necessarily more complicated. For a recent interesting discussion of the fit/interpretability trade-off, see Carvalho and Hahn (2014).

Note that we are also using more than one inferential paradigm in that our presentation of the logit results relies on the standard frequentist interpretation while our BART results are Bayesian and report posterior distributions. All BART results use the default prior discussed in detail in Chipman, George, and McCulloch (2010).

4.2 A Decision Tree

We first use a binary decision tree to see how `revcall` is related to the other variables. Figure 2 displays a tree with ten bottom nodes. For each non-bottom node, a decision rule is printed out. The rule specifies the condition for going left down the tree. The set of bottom nodes corresponds to a partition of the observations. For the subset of observations corresponding to a given bottom node, the percentage of observations where `revcall = 0` (the top number) and `revcall = 1` (the bottom number) is printed out.

The tree illustrates the importance of variables like `goaldiff`, `inrow2`, and `timebetpens`, which appear in the decision rules toward the top of the tree. In addition, the tree describes scenarios under which a reverse call is more or less likely. For example, if we go left, right, left, and then left down the tree, we find a scenario (and corresponding subset of the data) where 72% of the calls are reversals. The scenario is described by the path we took: the last penalized team does not have a lead (`goaldiff` < 0.5) *and* that team has had the last two penalties (`inrow2` $= 1$) *and* it has been less than 6.8 minutes since the last call (`timebetpens` < 6.8) *and* there is only one referee (`tworef` $= 0$). A high reverse-call percentage for this scenario is consistent with the theory of referee behavior described in Section 1. Under these circumstances, the referee may find it difficult to give the same team a third penalty.

Alternatively, if we go right (last penalized team is ahead), right (time since the last penalty is more than 3.4 minutes), and left (the last three penalties have not been called on the same team) down the tree, we have a scenario where only 48% of the calls are reversals.

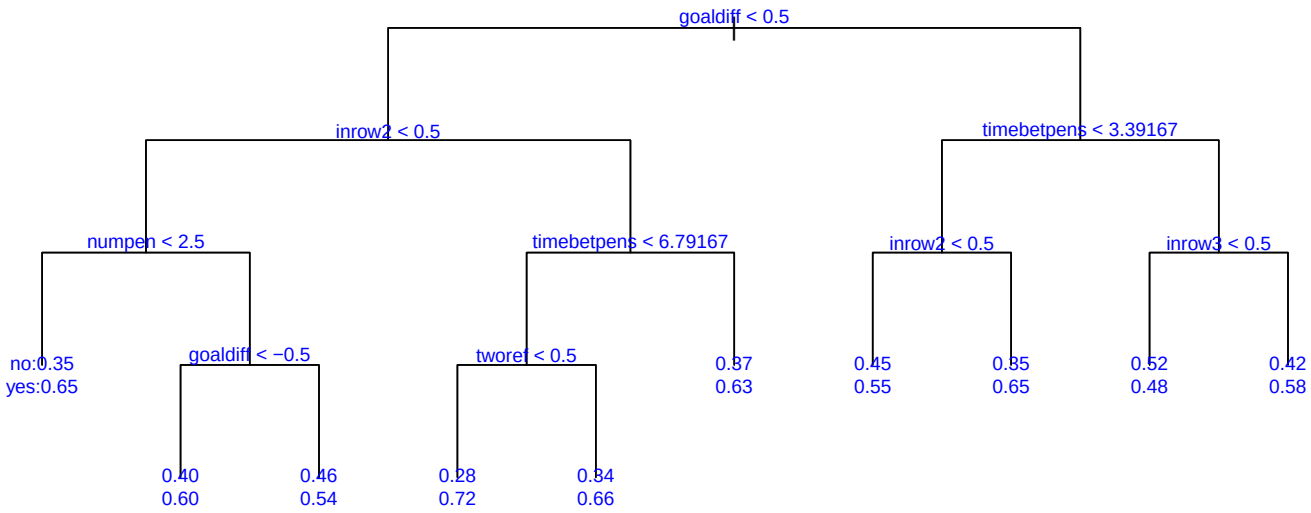

Figure 2: *Tree fit. The bottom nodes correspond to a partition of the data while all the other nodes give the decision rules leading down to the bottom nodes. At each bottom node two numbers are displayed. The top number give the fraction of observations in that bottom node where the penalty was not a reverse call and the bottom number gives the fraction of reverse calls. The wide range in these fractions shows that even though the tree is not our best performer in terms of out-of-sample prediction, it does find substantial fit.*

4.3 Logistic Regression

Table 4 reports the results from a standard logistic regression of `revcall` on the other variables (no interactions). The indicator variables `inrow2`, `inrow3`, `inrow4`, `home`, and `tworef` are all highly significant statistically. The size of the coefficients indicate they are of practical significance as well. Since the probabilities of a reverse call are not extreme, the logistic function is roughly linear so that we can gauge the change in probability as approximately .25 times the change in the log odds (which is linear in the variables). So, for example, the coefficent .2 for `home` indicates an increase of .2 in the log odds when the last penalty was on the home team. This translates into an approximate (and substantial) increase in the probability of a reverse call of .2*.25=.05. Even more substantial are the effects of the `inrow` variables. The coefficient estimate of `inrow2` is .31. The corresponding change in the probability of a reverse call is .31*.25=.08. The effect of three calls in a row is (.31+.28)*.25=.15, and the effect of four calls in a row on the same team is (.31+.28+.17)*.25 = .20. Having two referees rather than one gives a probability change of -.11*.25=-0.0275, which is smaller but still appreciable.

Some of the seasonal dummies are significant. Note, however, that because one referee was used in the early seasons and two in the latter seasons, `tworef` is partially confounded with season. The last season is used as the "base case" in that its dummy was not included in the regression. If we compare a two-referee game in the 2001-2002 season with a one referee game in the 1995-1996 season we have a change in p(`revcall`) of (.14+.108)*.25 = 0.062.

`ppgoal` is barely significant at a 5% level, and the size of the coefficient suggests a small increase in the probabability of a reverse call when the last power-play resulted in a goal. The numeric variables `numpen`, `timebetpens`, and `goaldiff` are highly significant statistically and large in magnitude. An increase in the lead of the last penalized team of four (two goals up versus two goals down) decreases the probability of a reverse call by -.12*4*.25 = -0.12. If it has been 10 minutes since the last call, the reverse call probability goes down by -.023*10*.25 = -0.06. Again, these effects make intuitive sense: if the last penalty was on the winning team or it has been a while since the call was made, it is easier to give them another penalty.

`numpen` is also significant, but care is needed in interpreting the coefficient since it is difficult to envisage a change in `numpen` holding other variables (such as `timeingame`) constant. The `pf` and `pa` variables are also high significant suggesting that these measurements of the two team's propensities to commit offenses serve as useful controls.

Table 4: Logistic Regression Results

Variable	Coeff. Est.	Std. Err.	p-value
(Intercept)	0.8884	0.2139	3.27e-05
inrow2	0.3095	0.0206	$<$ 2e-16
inrow3	0.2768	0.0364	3.01e-14
inrow4	0.1662	0.0639	0.0093
home	0.2055	0.0173	$<$ 2e-16
ppgoal	0.05306	0.02464	0.0312
tworef	-0.1078	0.0370	0.0035
timeingame	-0.001337	0.000981	0.1731
dayofseason	0.0003010	0.0001511	0.0464
numpen	-0.02595	0.00488	1.04e-07
timebetpens	-0.02282	0.00194	$<$ 2e-16
goaldiff	-0.1240	0.0053	$<$ 2e-16
gf1	0.03615	0.0283	0.2019
ga1	-0.05820	0.02487	0.0193
pf1	-0.2985	0.0196	$<$ 2e-16
pa1	0.2612	0.0243	$<$ 2e-16
gf2	-0.06192	0.02829	0.0286
ga2	0.02805	0.02517	0.2652
pf2	0.2879	0.0195	$<$ 2e-16
pa2	-0.3074	0.0242	$<$ 2e-16
S1995	0.1405	0.0680	0.0387
S1996	0.1385	0.0549	0.0116
S1997	0.1091	0.0544	0.0449
S1998	0.0628	0.0443	0.1567
S1999	0.1182	0.0366	0.0013
S2000	0.0137	0.0328	0.6761

4.4 BART Results

As discussed in Section 4.1 (and in more detail in Appendix B), BART gave us the best results in our out-of-sample experiment. There are large "main effects," and the linear logit is adequate for getting a take on those, but BART uncovers more nuanced effects that are missed by the logic model. From BART, we can obtain draws from the posterior distribution of p(`revcall` $|x$) for any x. In this section, we present approaches for obtaining interpretable results from the BART procedure.

First, in Section 4.4.1, we see what x configurations give us high values for p(`revcall` $| x$) and which x configurations give us low values. In the following two subsections, we design various "experiments" in which x is carefully varied and then study the resulting posterior distributions of p(`revcall` $|x$). Results are presented for two different kinds of x experiments. In our first experiment (Section 4.4.2), we vary one component of x (that is, one variable) at a time to examine the "partial effects" of the variables. This is what we did to interpret the logistic results. In our second set of experiments (Section 4.4.3), we pick a subset of important variables and vary them jointly.

4.4.1 Average x for High and Low Probability of a Reverse Call

Given x, we estimate p(`revcall` $| x$) with the posterior mean, which is computed using the average of the BART draws. As a simple way of understanding the BART fit, we took the 1,000 observations with the highest estimates of p(`revcall` $| x$) and averaged the variables using those observations. Then, we did the same for the 1,000 observations with the lowest estimates. This exercise allows us to easily see what kind x values make reverse calls likely and unlikely, respectively.

Table 5 reports the results. The fraction of reverse calls (i.e., the average value of `revcall`) is 0.85 for the 1,000 observations with the highest estimated probability of a reverse call and 0.34 for the 1,000 observations with the lowest. This huge difference reflects the strong fit of BART; when p(`revcall` $| x$) is estimated to be high, 85% of the penalties are reverse calls.

The averages of the explanatory variables tell us what kind of game situation corresponds to a high or low probability of a reverse call. For example, 86% of the high probability observations have `inrow2` $= 1$ (last two calls on the same team) while only 7% of the low probability observations do. Many of the explanatory variable means are quite different. A higher probability of a reverse call is associated with higher averages for all three `inrow`

variables, the last penalty being on the home team, the last penalized team being behind in the game, a short time since the last penalty, having one referee in the game, and a higher chance that a goal was scored on the last powerplay. All of these conditions are consistent with a high stress environment for the referee(s). We omitted the season variables since there are several of them and no obvious pattern in the effects.

Table 5: Mean of explanatory variables for low and high p(`revcall`).

Variable	Low p(`revcall`)	High p(`revcall`)
`revcall`	.34	.85
`inrow2`	.07	.86
`inrow3`	.00	.53
`inrow4`	.00	.18
`home`	.17	.70
`goaldiff`	2.1	-1.1
`timebetpens`	9.7	3.0
`tworef`	.47	.14
`ppgoal`	.045	.318
`numpen`	6.73	5.38
`timeingame`	38.3	26.2
`dayofseason`	91	103

4.4.2 Changing One Variable at a Time

To examine the variables' partial effects, we first set all variables to a base setting and then change one variable at a time. The base setting has the following variable values:

- `gf` = `ga` = 2.8, `pf` = `pa` = 4 (average teams playing)
- `dayofseason` = 100 (middle of the season)
- `timeingame` = 30 (middle of the game)
- `numpen` = 4 (four penalties have been called)
- `timebetpens` = 6 (six minutes since the last penalty)
- `goaldiff` = 0 (tie score)
- `inrow2` = 0 (last two penalties on different teams)
- `tworef` = 0 (one referee)
- `home` = 0 (last penalty not on home team)
- `season` = 1997

At the base setting, the probability of a reverse call is .51, which is lower than the overall rate of .59. Since `inrow2` = 0, `goaldiff` = 0, and it is the middle of the game, we might expect that this base setting represents a "low stress" environment for the referee(s).

Figure 3 displays the effects of the binary variables `inrow(2,3,4)`, `tworef`, `ppgoal`, and `home`. The top panel shows the posterior distributions of p(`revcall` $\mid x$) where x represents various settings for these variables. In this figure, and throughout the paper, each posterior is represented by a black solid dot at the posterior mean and a line covering the 90% posterior interval. The labels along the horizontal axis indicate the x setting. For example, the label "`home=1`" means x is the base setting with `home` set to 1. At the settings "`tworef=0`", and "`tworef=1`", we also change the season to be 1998 since this season had both one-referee games and two-referee games. For "`inrow3=1`", both `inrow2` and `inrow3` are set to 1. For "`inrow4=1`", `inrow2`, `inrow3`, and `inrow4` are set to 1.

The bottom panel shows the posterior distributions of the *difference* in p(`revcall` $\mid x$) due to a change in x. For example, the posterior labelled "`tworef`" depicts the posterior of the

difference between p(`revcall` | x) at x with `tworef=1` and x with `tworef=0`. So, the first posterior in the bottom panel correspond to the posterior of the difference of the quantities represented in first two posteriors of the top panel. The posteriors labelled "`inrow(2,3,4)`" depicts the posterior of the difference between p(`revcall` | x) at x with the corresponding `inrow` variables set to 1 with x such that all `inrow` variables are set to 0. The posterior means depicted in the bottom panel of Figure 3 are, from left to right, -.023, .0067, .06, .09, .2, and .24. Clearly, the `home` effect is large and the `inrow` effects are very large. The BART estimates are in line with those obtained from the logistic regression, but even larger for the `inrow` variables.

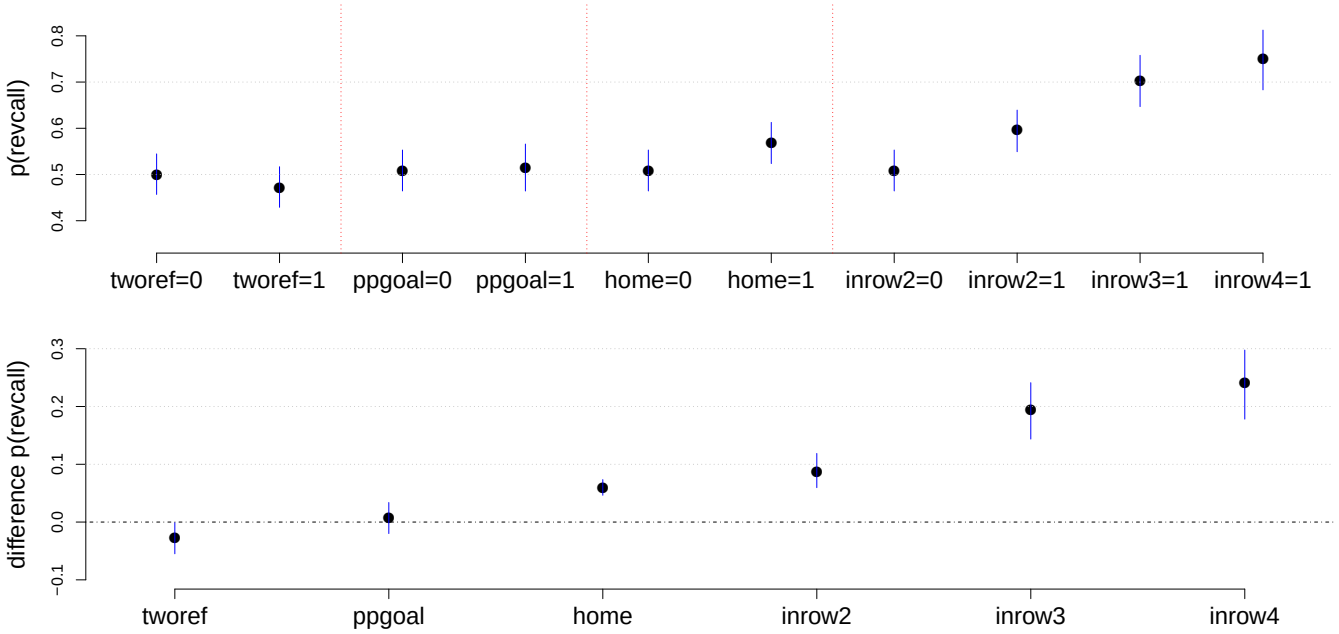

Figure 3: *Effects of binary predictors. The top panel shows the posterior distribution of the probability of a reverse call at various settings of the binary predictors which are indicated by the label on the horizontal axis. The bottom panel shows the posterior distribution of the difference in the probability of a reverse call due to a change in the variable indicated by the label on the horizontal axis. The solid dot is at the posterior mean and the vertical lines indicate 90% posterior intervals. The inrow variables and home are clearly important.*

Figure 4 displays the effects of `goaldiff`, `timebetpens`, `dayofseason`, and `season`. `tworef` is also changed to correspond to the season. `goaldiff` and `timebetpens` are important variables with intuitive effects. If the last penalized team was behind, you don't want to

call them again. The longer it has been since the last penalty, the smaller the tendency to reverse call. `timebetpens` exhibits an interesting non-linearity which may be due to the fact that a penalty lasts two minutes and it takes a bit of time after the penalty is over to "forget". There also appears to be a `season`/`tworef` effect in that the last two seasons exhibit a downward shift.

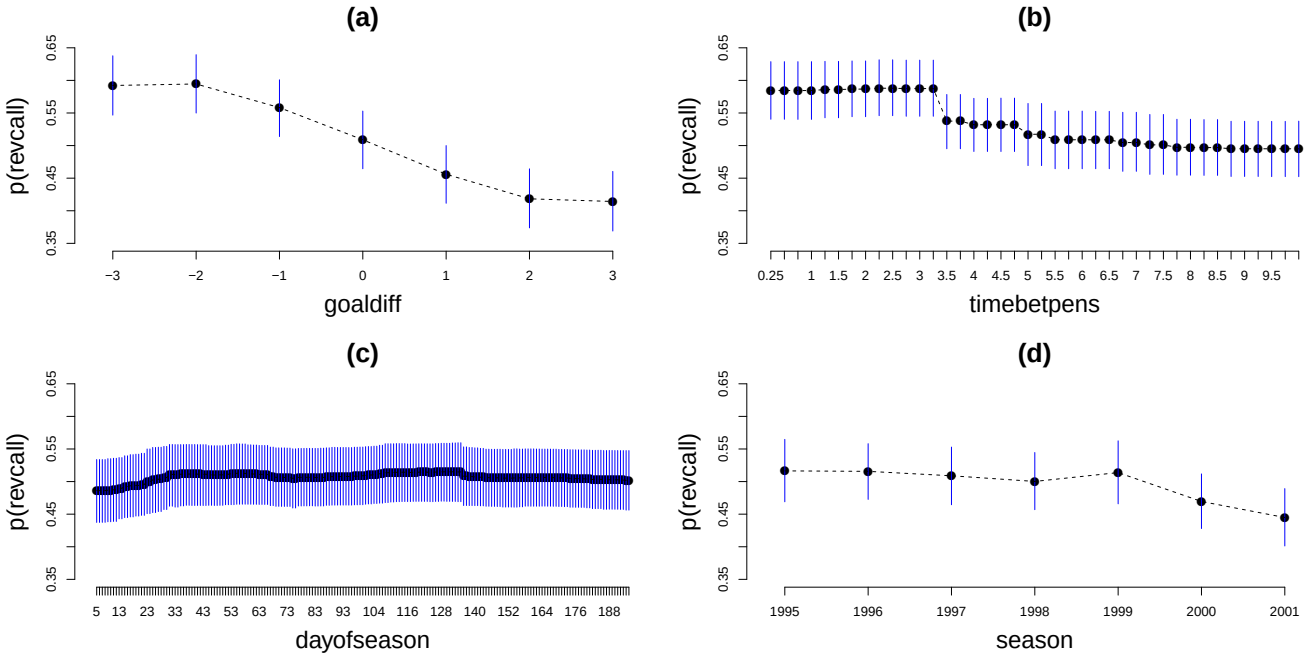

Figure 4: *Effects of non-binary predictors. Each panel shows how the posterior distribution of the probability of reverse call varies as a single x variable changes. Panel (a): goaldiff (goal differential); Panel (b): timebetpens (time between penalties, in minutes); Panel (c): dayofseason (day of the season); Panel (d): season (year). All four panels are on the same vertical scale. The solid dot is at the posterior mean, and the vertical lines indicate 90% posterior intervals. goaldiff and timebetpens have important effects.*

4.4.3 Joint Variation of Variables

In this section, we focus on five factors and consider setting each of them at two possible levels. Table 6 describes the five different factors as well as the two levels considered for each. The first three factors are the variables `goaldiff`, `inrow2`, and `timebetpens`. The fourth factor (associated with the time in the game) varies both `timeingame` and `numpens`

together due to their obvious dependence. The fifth factor is the variable `home`. We use the term "factor" to emphasize that a considered change may involve more than one variable.

Setting all other variables to the base scenario considered in the previous section, the varying of the levels for the five factors in Table 6 gives us 32 possible game scenarios. In order to denote a given scenario, a letter code is used for each factor. The case of the letter code then denotes the level, lower case for a low value and upper case for a high value. Table 6 summarizes the coding scheme. For example, the code `gRtnH` denotes the game scenario where the last penalized team is behind by a goal (`g`), the last two penalties were on the same team (`R`), it has been a short time since the last penalty (`t`), it is early in the game with few penalties called (`n`), and the last penalty was on the home team (`H`).

Table 6: Description of Scenario Design

Quantity	Code Meaning	Code Meaning
goal differential	g `goaldiff` $= -1$ *(last penalized team behind by a goal)*	G `goaldiff` $= 1$ *(last penalized team ahead by a goal)*
consecutive calls	r `inrow2` $= 0$ *(last two calls on different teams)*	R `inrow2` $= 1$ *(last two calls on same team)*
time between penalties	t `timebetpens` $= 2$ *(short time since last penalty)*	T `timebetpens` $= 7$ *(long time since last penalty)*
time in the game	n `timeingame` $= 10$, `numpens` $= 3$ *(early in the game)*	N `timeingame` $= 55$, `numpens` $= 12$ *(late in the game)*
penalty on home team	h `home` $= 0$ *(last penalty not on home team)*	H `home` $= 1$ *(last penalty on home team)*

Figure 5 shows the 32 posterior distributions of p(`revcall` $|\, x$) for the 32 different x obtained by setting the 6 variables (`goaldiff`, `inrow2`, `timebetpens`, `timeingame`, `numpens`, and `home`) to the 32 possible settings described in Table 6 and setting all other variables to the base level used in Section 4.4.2. The variation in the posteriors is remarkable — the posterior means range from a high of .80 to a low of .43. *Yes Mr. Healy, referees are predictable!*

The game scenario which generates the highest probability of a reverse call (.80) is the 10^{th} posterior in the top group, `gRtnH`:

- g: the last penalized team is behind,

- R: they just had two calls in a row against them,
- t: is has not been long since the last penalty,
- n: it is early in the game, and
- H: they are the home team.

The two game scenarios which generates the lowest probabilities of a reverse call (.43 and .45) are the 5^{th} and 7^{th} posteriors in the bottom group, `GrTnh` and `GrTNh`:

- G: the last penalized team is ahead,
- r: the last two calls were on different teams,
- T: is has been a while since the last penalty,
- nN: early or late in the game,
- h: they are not the home team.

Clearly, the high probability scenario describes a situation where it may be stressful for the referee to make another call on the same team. The low probability scenario is almost the reverse situation, except for the time in game.

The four game scenarios with the largest p(`revcall` $\mid x$) are

`gRtnh` `gRtnH` `gRtNH` `gRTnH`

It is clear that `rR` (=`inrow2`) is a key factor, with the score, the time since last penalty, the time in game, and the home team also having important effects.

In Figure 5 the posteriors are ordered so that successive pairs change `h` to `H`. Hence, we can readily see the `hH` effect (last penalty on home team versus not). The size of the effect is about .05 and it does not depend much on the levels of the other factors. The effects of the other factors are not easily seen in Figure 5.

The top panel of Figure 6 shows the posterior distribution of the difference p(`revcall` $\mid x_R$) − p(`revcall` $\mid x_r$) where x_R is a setting of the explanatory variables such that `inrow2`=1 and x_r is a setting of the explanatory variables such that `inrow2`=0. The 16 posteriors depicted

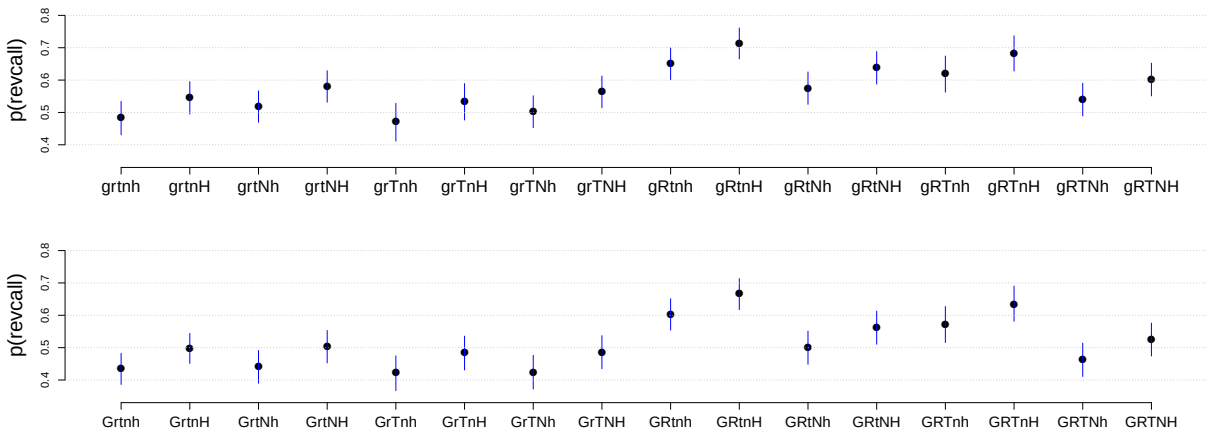

Figure 5: *Posterior distribution of p(*`revcall` $\mid x$*) at each of the 32 game scenarios. The game scenario is indicated by the label on the horizontal axis. The solid dot is at the posterior mean and the vertical lines indicate 90% posterior intervals. The probability of a reverse call changes dramatically with the game situation indicating the strong BART fit.*

correspond to varying the remaining factors at their two levels. Thus, the first posterior in the top panel of Figure 6 is for the `rR` (`inrow2`) effect at the setting `gtnh` for the other factors. The effect of `rR` depends mostly on `nN`. The posterior mean of the effect is about .18 at `n` and .08 at `N`. The `rR` effect is always very large, and it depends strongly on the time in the game. It may be that early in the game the referee is "setting the tone", but late in the game, when the outcome may be on the line, he has to make the "real calls".

The second panel of Figure 6 shows the posterior distribution of the difference p(`revcall` $\mid x_G$) - p(`revcall` $\mid x_g$) where x_G is a setting of the explanatory variables such that `goaldiff`=1 and x_g is the same except that `goaldiff`=0. The posterior mean of the `gG` effect is always big, and depends mostly on `nN` as well. The posterior mean of the effect is about .10 at `n` and .13 at `N`. This suggest that late in the game, the score is a more important factor.

Finally, we introduce `tworef` as a sixth factor. The effect of `tworef` is particularly interesting because such an effect supports our basic story about refereeing. The variable `tworef` seems important in Figure 2, Table 4, and Table 5, but does not exhibit a large effect in Figure 3. Perhaps the size of the effect is very dependent on the game situation.

Figure 7 shows the posterior distributions of the `tworef` effect at each of the 32 possible settings of the five other factors. Each posterior is that of the difference in the probability of

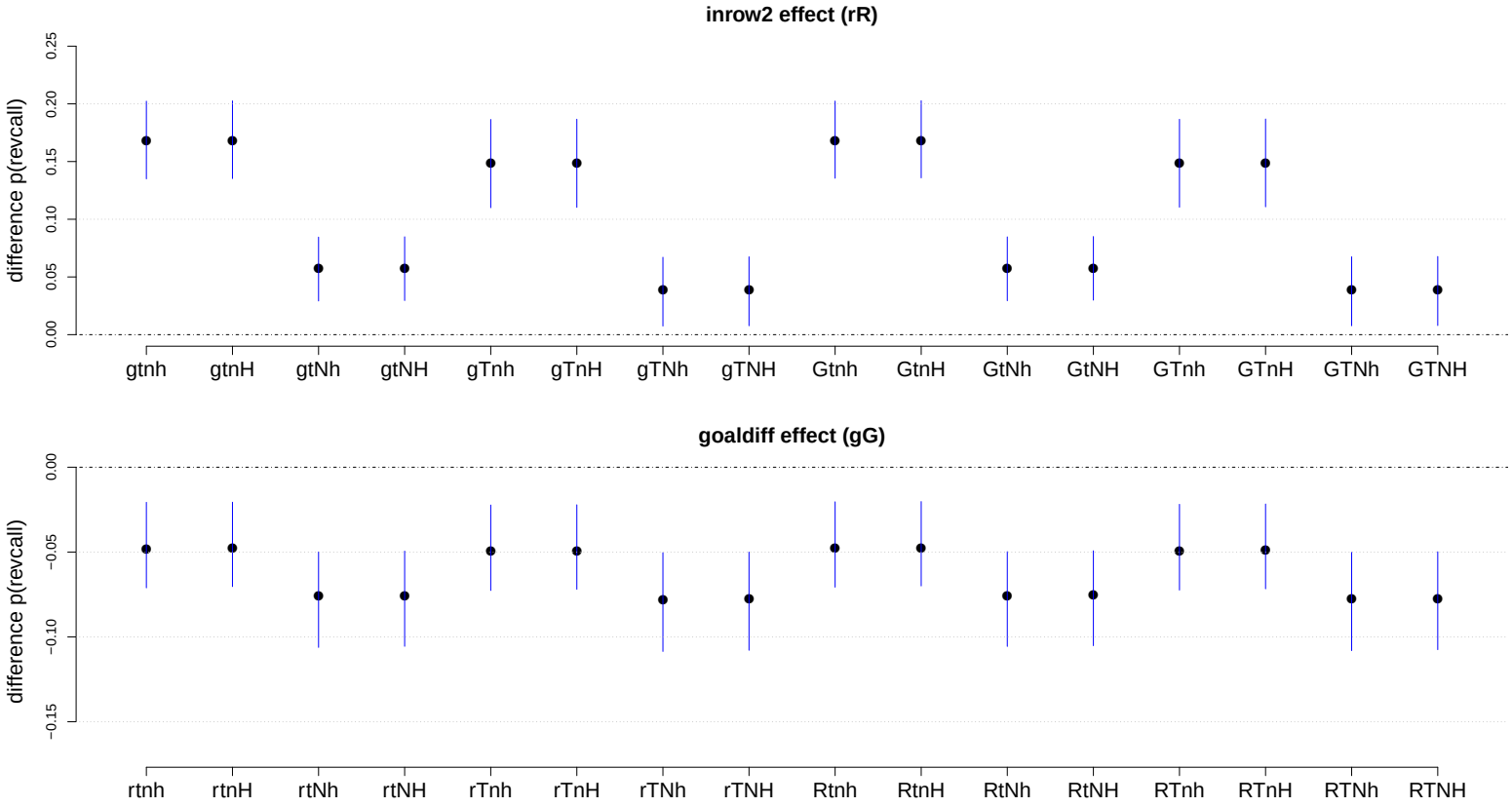

Figure 6: *Posterior distributions of the effects of inrow2 (rR) (top panel) and goaldiff (gG) (bottom panel). There is a larger (rR) effect early in the game and a larger (gG) effect late in the game. The solid dot is at the posterior mean and the vertical lines indicate 90% posterior intervals.*

a reverse call between the base setting with `tworef`=0, `season`=1997, and setting `tworef`=1, `season`=2001. The 32 possible settings of the five variables in Table 6 are indicated by the label on the horizontal axis. We see that `tworef` strongly interacts with `goaldiff` (`gG`), `inrow2` (`rR`), and `timebetpens` (`tT`). For example, the four x with `GrT` exhibit no `tworef` effect, while at the oppposite settings (`gRt`) there is a large effect of about -.1 `gRt` corresponds to the high stress game situation where the last penalized team is behind, the last two calls have been on the same team, and it has not been long since the last call. Thus, it makes sense that the effect of including an additional referee is large at this setting and negligible at the reverse setting.

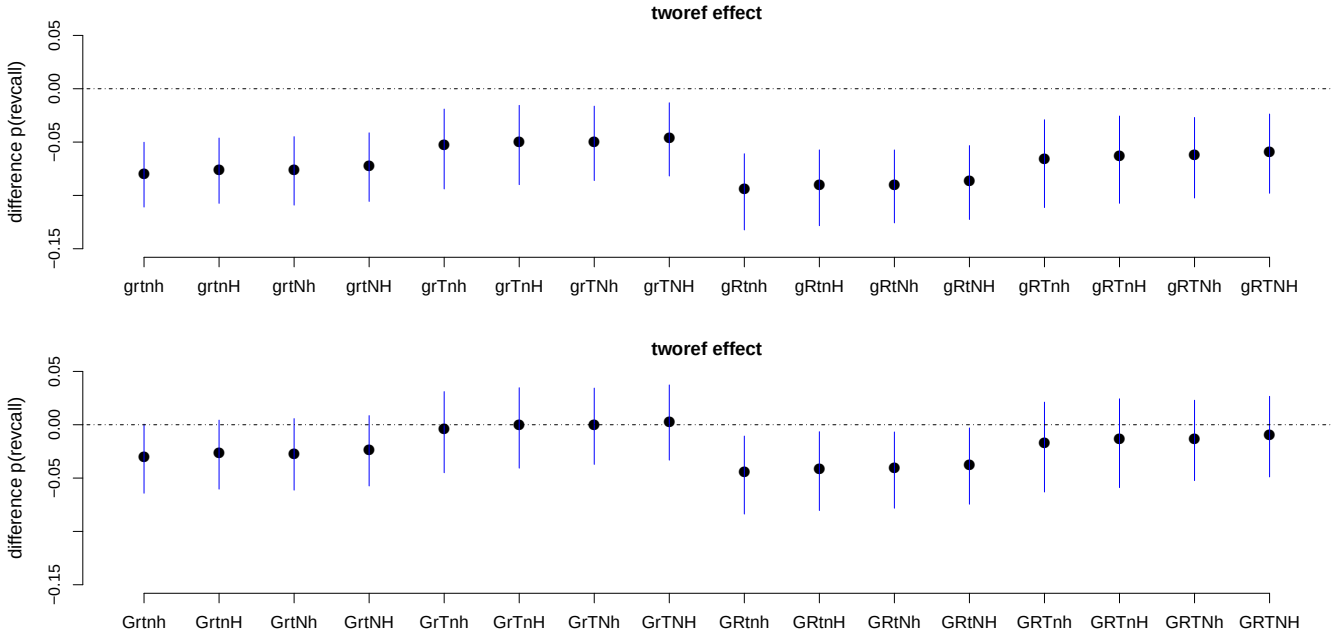

Figure 7: *Posterior distributions of tworef effect at various settings of the other variables. Settings indicated by the label on the horizontal axis. tworef interacts with goaldiff (gG), inrow2 (rR), and timebetpens (tT). The solid dot is at the posterior mean and the vertical lines indicate 90% posterior intervals.*

5 How Different are Referees?

In this section, we consider the behavior of individual referees. In Section 5.1, we use only the data from one-referee games to see if there are "referee effects." That is, do different referees have different propensities to reverse call? In Section 5.2, we look at referees who refereed alone and with a partner, and investigate whether the addition of a partner affected the probability of a reverse call. If the psychological strain is eased with a partner, perhaps the tendency to reverse call is less.

5.1 Comparing Referees, One-Referee Games

In this section, we investigate how individual referees vary in their tendency to reverse call. In order to model the "referee effect" simply, attention is restricted to the 30,918 penalties in the sample for one-referee games between the 1995-96 and 1998-99 seasons. Only the 20 referees with at least 100 penalty calls in the data were considered. All others were lumped into the the referee category "other," giving a total of 21 referee categories. We analyzed the data using the same variables as in Table 3 and Section 4 except that the variable `tworef` is gone, there are only four seasons, and dummy variables are added for the referee identities.

We first ran a logistic regression using "other" as the excluded category. Only two of the referees had significant estimates (at a 5% level). The regression output for those referees is given in Table 7. The next largest coefficient for the referees not shown is -.19 (p-value = .07), so these two referees really stand out.

Table 7: Logistic Regression Results with Referee Identities

Variable	Coeff. Est.	Std. Err.	z-value	p-value
referee 18	0.2546	0.1071	2.378	0.0174
referee 20	-0.5552	0.2090	-2.656	0.0079

We then used BART following the approach in Section 4.4, choosing a base scenario for prediction and then varying only the referee identity. For the base scenario, we chose the `gRtnH` scenario (the "high-stress scenario") described in Section 4.4. (The other interaction scenarios yielded very similar results and are therefore not reported.) Figure 8 shows the posterior distributions of p(`revcall`) for each of the 21 referees. Again, as suggested by the

logit results, only referees 18 and 20 standout as potentially different. Given that there is a "multiple comparisons" issue in that we are looking through the figure for the most different referees, a reasonable summary conclusion is the there is not strong evidence for a referee effect.

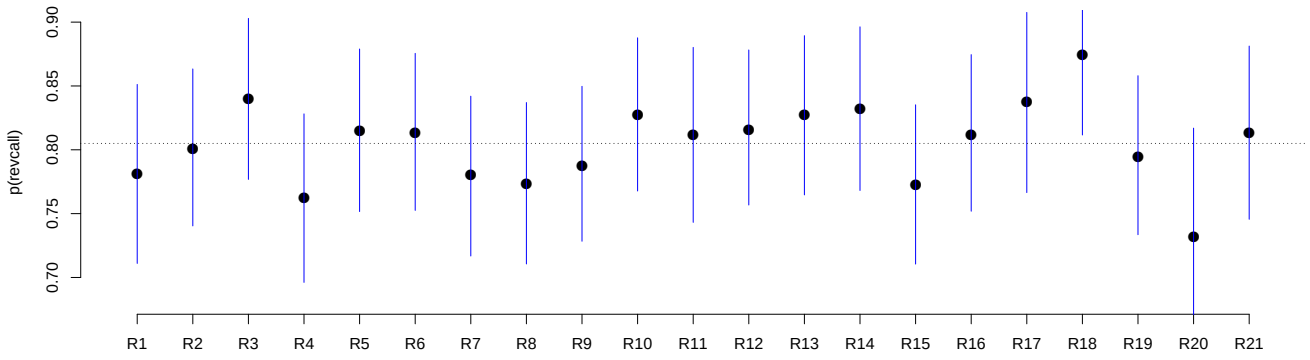

Figure 8: *Posterior distribution of p(*`revcall`*) for each referee. Only referees 18 and 20 seem to be different from the rest. The solid dot is at the posterior mean and the vertical lines indicate 90% posterior intervals.*

5.2 Comparing Referees, Before and After Two Referees

Figure 9 looks at how the reverse-call propensity varies across referees comparing the one-referee system with the two-referee system. For all referees that officiated in at least 140 games under both the one-referee system and the two-referee system, we calculated the percentage of reverse calls in one-referee and two-referee games. Each point in Figure 9 corresponds to a specific referee, with the one-referee reverse call percentage on the horizontal axis and the two-referee percentage on the vertical axis. The line $y = x$ is also drawn through the plot. The symbols plotted are the same identification numbers used in in the previous section, but it is not the same set of referees since not all refereed under both systems. So, for example, referee 3 made reverse calls 60% of the time when by himself and 57.5% of the time when working with a partner.

The reverse call percentages under the one-referee system are higher overall as well as more disperse. All referees except 4 and 19 have lower reverse-call percentages under the two-

referee system. Referees 10, 12, and 18 are about 5% less likely to reverse call with a partner! The amount of dispersion across referees is quite high. Looking at the one-referee numbers, the reverse-call percentages range from 56% to 65%. The positive relationship between the reverse-call percentages across the two systems is also evident.

The results in Figure 9 strongly suggest that referee behavior was changed by the addition of a partner. There are several possible explanations. One possibility is that with a partner, the psychological stress is diminished and so is the tendency to reverse call. Another possibility is that two referees are able to more accurately detect actual offenses and, therefore, fewer reverse calls are needed in the name of fairness.

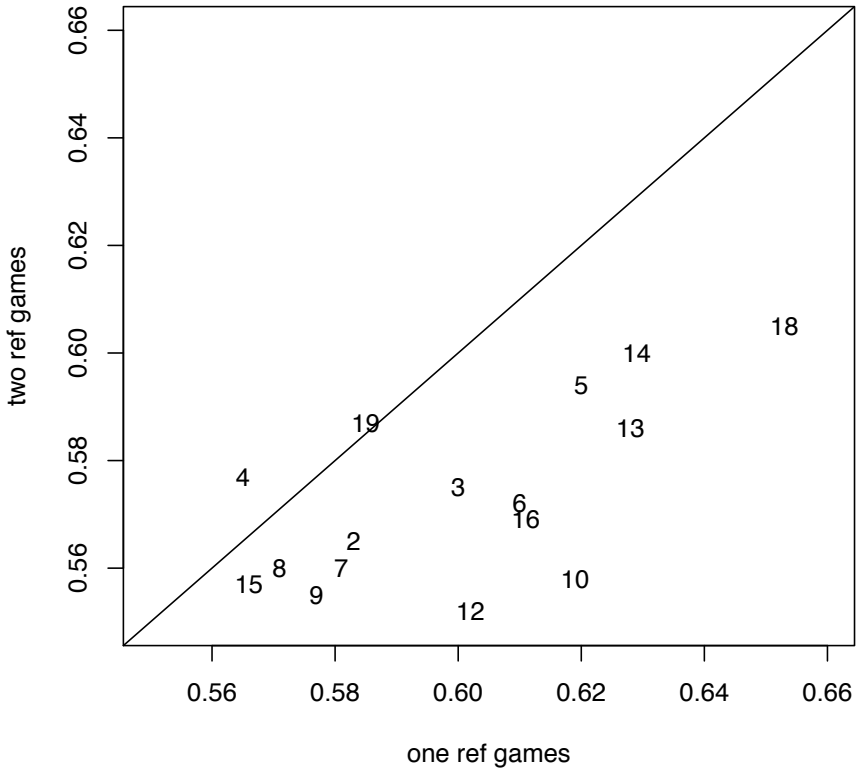

Figure 9: *Percentage of reverse calls, by referee, one-referee games versus two-referee games. Most of the points fall below the line y=x, indicating that referees call a greater percentage of reverse calls in one-referee games.*

6 Discussion

Using a variety of techniques, we have found strong evidence for the existence of patterns in penalty calls. Most striking is the tendency to not call a penalty on the same team "too many" times in a row. We also find evidence that whether a team is the home team, the score, the time since the last penalty, the time in the game, and the number of referees have effects on which team is penalized next. We compared games using one referee with those using two and found that there is less of a tendency to reverse call with two referees.

Our basic hypothesis is that the speed and physical nature of the NHL game, combined with the expectation that many infractions will go uncalled, put the referee in a very difficult situation. Since penalty calling can be largely subjective, it is easy for teams, coaches, and fans to view referees' calls as unfair to their side. In order to keep control of the game and avoid being blamed for the outcome of the game by the loser, it would seem logical that referees would adopt the following strategies:

- make fewer calls later in the third period when the game is on the line
- avoid making repeated calls on the same team
- avoid penalizing the team that is behind
- avoid penalizing the home team
- avoid calls on the same team in quick succession

The empirical findings of this paper are consistent with each of these strategies. In addition, there is evidence that these factors may interact with each other. For instance, the avoidance reverse calls appears to be more acute early in the game and the effect of the score is larger late in the game. We also find that the overall tendency to reverse call is lower in games with two referees.

Of course, this paper represents an "observational study," and there may exist other explanations for the patterns we find. In particular, some of our findings may be due to player behavior rather than referee behavior. Here are some possible player-based explanations of the patterns in penalty calls:

- *Prevalence of reverse calls*: If a penalty call on one team causes the other team to play more aggressively (or seek retribution), a reverse call would be more likely than a

repeat call. This theory seems less convincing than the referee-based theory since the team with the power play has a huge incentive to avoid taking a penalty and losing their man advantage. In contrast, our analysis shows that reverse calls would be most likely during the power play (a low value for `timebetpens`). Also, the reverse-call probability gets larger with additional repeat penalties.

- *Penalty calls influenced by score*: The basic pattern is that penalties are more likely to be called on the team that is ahead and less likely to be called on the team that is behind. One player-based explanation would be that the team with the lead may feel less apprehensive about giving the other team a power play; they may take "lazy penalties" (hook a player rather than skating hard to catch him) or take the opportunity to even a score (not the one on the scoreboard). However, other plausible player-based explanations would be consistent with the opposite pattern of penalty calls (more calls on the team behind). For instance, the team behind may play a more desperate physical style that could lead to penalties. The evidence that the effect of the score on the reverse call is larger when it has been a long time since the last penalty seems more consistent with a referee-based theory than any player-based theory we can think of.

- *Fewer calls on the home team*: The "home-field advantage" is a phenomenon common throughout sports. In hockey, if players play better at home and commit fewer infractions in front of their fans, the finding of fewer penalties on the home team could be explained without a home-biased referee. Alternatively, if the visiting team plays very aggressively in an attempt to get a tie or win on the road, it would be reasonable to expect more calls on the visiting team. Unlike the referee-based theory, however, there are plausible stories regarding player behavior that would be consistent with *more* penalty calls on the home team. For instance, the home team might play more aggressively in an effort to entertain its fans, or the visiting team might play more carefully to avoid penalties and playing shorthanded on the road.

- *Different play during the penalty kill*: It is possible that a team on a power-play plays differently, in a way that encourages penalties. It may also be that the team killing the power-play is more defensive, does not have control of the puck as much and is more likely to be guilty of a subsequent infraction.

- *Fewer calls late in the game*: Both teams may play more carefully when the game is on the line.

Finally, we point out that two of our basic findings have no obvious player-based explanation: (1) the significant difference in reverse-call probabilities between one-referee and two-referee games, and (2) the variation in reverse-call probabilities across referees.

It would be interesting to perform an experiment to explore these issues further. For example, referees could be shown videos of infractions coupled with a description of the game situation. Would referees make the same calls watching a video in safety as they do on the ice?

Appendix A

Sample Game Boxscore in Raw Form

+++ National Hockey League - Lightning vs. Stars - 10/22/1995 0:58AM ET +++

Tampa Bay 1 1 1 0--3

Dallas 2 1 0 0--3

FIRST PERIOD -- Scoring: 1, Dallas, Klatt 2 (Matvichuk), 5:04. 2, Tampa Bay, Gavey 1 (shorthanded) (Tucker, Hamrlik), 12:13. 3, Dallas, Adams 4 (Ledyard), 16:08. Penalties: Matvichuk, Dal (charging), 8:13; Wiemer, T.B. (high sticking), 10:47; Burr, T.B. (charging), 17:07; Zmolek, Dal (high sticking), 17:07; Hamrlik, T.B. (holding), 17:38.

SECOND PERIOD -- Scoring: 4, Dallas, Modano 5 (power play) (Ledyard, D Hatcher), 8:58. 5, Tampa Bay, Gratton 5 (power play) (Klima, Cullen), 17:00. Penalties: Borschevsky, Dal (Obstr hooking), 0:47; Gratton, T.B. (slashing), 1:12; Bradley, T.B. (Obstr hooking), 7:06; Charron, T.B. (high sticking), 8:21; D Hatcher, Dal (closing hand on puck), 15:05.

THIRD PERIOD -- Scoring: 6, Tampa Bay, Houlder 1 (Klima, Ysebaert), 6:39. Penalties: Tampa Bay bench, served by Selivanov (too many men on the ice), 2:43; Zmolek, Dal (Obstr hooking), 10:49; Klima, T.B. (Obstr tripping), 11:45; Ciccone, T.B. (slashing), 13:49.

OVERTIME -- Scoring: None. Penalties: Burr, T.B. (roughing), 0:58; Churla, Dal (roughing), 0:58; Hamrlik, T.B. (hooking), 1:53.

Shots on goal:

Tampa Bay 13 10 13 2--38

Dallas 14 9 5 2--30

Power-play Conversions: Tam - 1 of 4, Dal - 1 of 9. Goalies: Tampa Bay, Puppa (30 shots, 27 saves; record: 2-1-2). Dallas, Wakaluk (38, 35; record: 2-1-1). A:16,789. Referee: Roberts. Linesmen: D Mccourt, Mcelman.

Appendix B

Rather that fitting one type of model to the data, we consider a variety of strategies for learning the relationship between `revcall` and the other variables.

In order to gauge how well a model worked, we did a simple out-of-sample experiment. We randomly selected 11,000 observations to be our out-of-sample "test" data and then used the remaining data (57,883-11,000 = 46,883) observations as "training" data with which to estimate the models.

The "models" we tried were:

- (i) decision trees with various numbers of bottom nodes,
- (ii) random forests with various numbers of trees,
- (iii) linear logistic regression,
- (iv) boosting with various numbers of trees and interaction depths, and
- (v) BART: Bayesian Additive Regression Trees with the default prior.

Model fitting was performed in `R` using the `tree`, `randomForest`, `glm`, `gbm`, and `BayesTree` packages or functions, respectively.

Figure 10 displays the out-of-sample loss for the modeling strategies, where loss is measured by the deviance ($-2 \times$ *log-likelihood value*). Because the deviance has the opposite sign of the likelihood and a bigger likelihood is better, a smaller deviance indicates a better fit. For a textbook discussion of the use of deviance in model selection, see Chapter 6 of James, Witten, Hastie, and Tibshirani (2013). Again, we fit models using the training data and then evaluate the likelihood of the fitted model on the test data. The top panel displays the loss for all models, while the bottom panel only displays the results for logistic regression, boosting, and BART since these models performed the best. The best models in terms of out-of-sample loss are boosting with 250 trees and interaction depth 6 and BART.

While it is not easy to interpret the deviance measure, the results in Figure 10 suggest that logistic regression is "not too bad". Figure 11 plots the fitted p(`revcall`) from the BART fit against those obtained using logistic regression. The BART fit is a posterior mean. The line drawn throught the plot has slope 1 and intercept 0. Broadly, the two models agree

on the probabilities, but there is also substantial discrepancy. For both models, the fitted probabilites range from about .25 to about .90, which suggests a great deal of predictability in penalty calls. In Section 4.4 we see that BART does find interesting interactions which the linear logit clearly could not uncover.

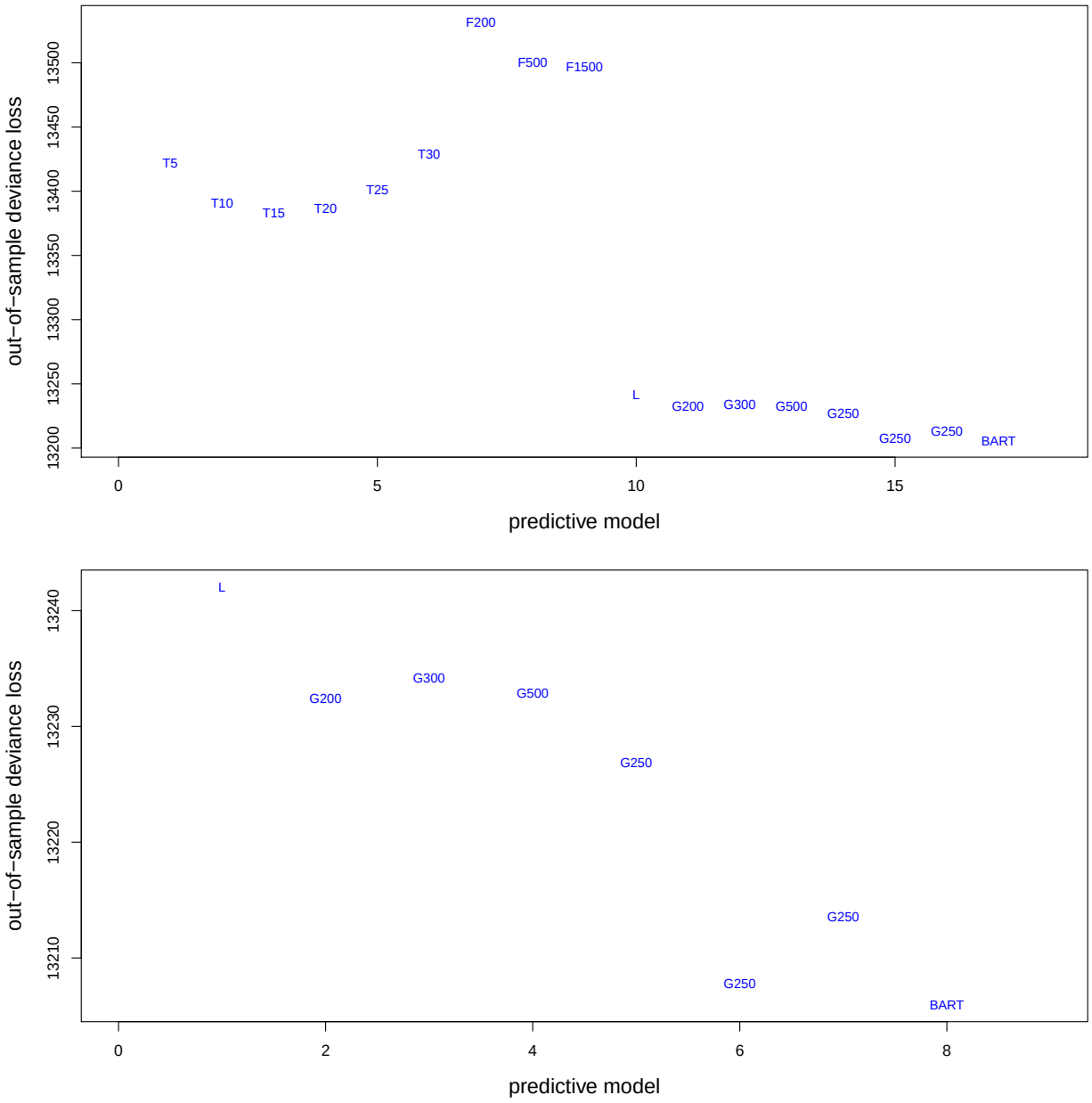

Figure 10: *Out-of-sample deviance loss for various predictive modeling strategies. (i) Ti: tree with i bottom nodes, (ii) Fi: random forest with i trees, (iii) L: linear logistic regression, (iv) Gi: boosting with i trees, for i=200,300,500, trees of depth 3 were used while for i=250 trees of depth 4,6, and 8 were used, (v) BART: Bayesian Additive Regression Trees with default prior. All other parameters set to defaults given in the* `R` *package. In the top panel, all methods are displayed. In the bottom panel we just compare the better ones. The predictive modeling method with the smallest loss is BART, with boosting with 250 trees and tree depth equal to 6 very close.*

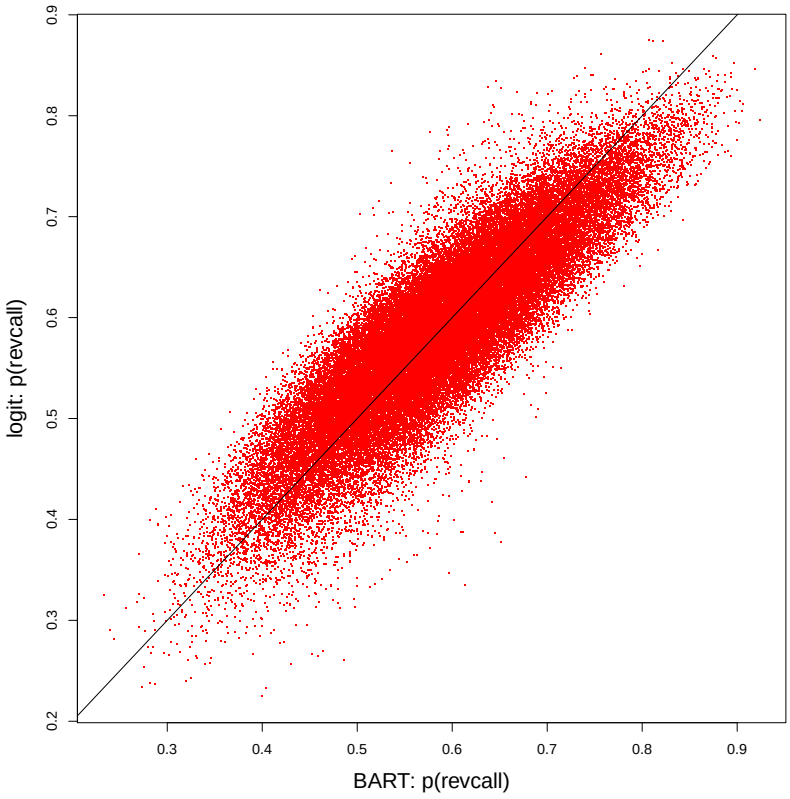

Figure 11: *BART fit (x axis) vs. logit fit (y axis). Both approaches find substantial fit, but there are also some big differences.*

References

Allen, W.D. (2002), "Crime, Punishment, and Recidivism, Lessons from the National Hockey League", *Journal of Sports Economics*, 3, 39-60.

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

Becker, G. (1968), "Crime and punishment: an economic approach", *Journal of Political Economy*, 76, 169-217.

Breiman, L. (2001), "Statistical Modeling: The Two Cultures", *Statistical Science*, 16, 199-231.

Carvalho, C., and Hahn, P.R. (2014), "Decoupling Shrinkage and Selection in Bayesian Linear Models", *working paper*.

Chipman, H., George, E., and McCulloch, R. (2010), "BART: Bayesian Additive Regression Trees", *Annals of Applied Statistics*, 4,1,266-298.

Heckelman, J. and Yates, Y. (2003), "And a Hockey Game Broke Out: Crime and Punishment in the NHL", *Economic Enquiry*, 41, 704-712.

Friedman, J. (2001), "Greedy Function Approximation: A Gradient Boosting Machine", *The Annals of Statistics*, 29,1189-1232.

James, G., Witten, D., Hastie T., and Tibshirani, R. (2013), "An Introduction to Statistical Learning", Springer, New York, New York.

Levitt, S. (2002), "Testing the Economic Model of Crime: The National Hockey League's Two Referee Experiment", *Contributions to Economic Analysis and Policy*, 1.

Murphy, K. (2012), "Machine Learning", The MIT Press, Cambridge, Massachusetts.

R Development Core Team (2003). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3, URL http://www.R-project.org.

R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, http://www.R-project.org/.