Biased Penalty Calls in the National Hockey League

David Beaudoin, Oliver Schulte and Tim B. Swartz *

Abstract

This paper investigates penalty calls in the National Hockey League (NHL). Our study shows that there are various situational effects that are associated with the next penalty call. These situational effects are related to the accumulated penalty calls, the goal differential, the stage of the match and the relative strengths of the two teams. We also investigate individual referee effects across the NHL.

Keywords : Logistic regression, National Hockey League, Referee evaluation.

*David Beaudoin is Associate Professor, Département Opérations et Systèmes de Décision, Faculté des Sciences de l'Administration, Pavillon Palasis-Prince, Bureau 2636, Université Laval, Québec (Québec), Canada G1V0A6. Oliver Schulte and Tim Swartz are Professors, School of Computing Science and Department of Statistics and Actuarial Science, respectively, Simon Fraser University, 8888 University Drive, Burnaby BC, Canada V5A1S6. Each author has been supported by the Natural Sciences and Engineering Research Council of Canada. The authors thank both Editors who provided constructive comments on an earlier version of the paper. The authors also thank Kurt Routley for his work developing the web crawler and Yifan (Lucas) Wu for programming assistance.

1 INTRODUCTION

The National Hockey League (NHL) is the premier hockey league in the world. Accordingly, officiating in the NHL is taken very seriously. There is a lengthy process involved in the training of NHL officials for which there are fewer than 80 positions (referees and linesmen). Apart from the general comportment and decision making aspects of officiating, an NHL official requires an athleticism that exceeds what is needed in most other sports. Officials require power and endurance to skate at high speeds in order to keep up to the pace of play. Furthermore, the NHL employs a Director of Officiating (Stephen Walkom appointed in 2013) who administers officiating performances. Every NHL match is monitored for questionable calls and NHL teams are able to submit complaints regarding officiating decisions (www.nhlofficials.com).

However, despite the rigorous standards and the monitoring of officials, it appears that there are ingrained practices of NHL officiating that defy fair play. Our view concerning fair play suggests that only on-ice infractions are relevant in the assignment of penalties. Extraneous factors such as the so called "home team advantage" should not be relevant in the determination of penalties.

To our knowledge, there have been very few quantitative investigations of penalty calls in the NHL. From an economics perspective, Allen (2002) investigated the behavioral aspects of "crime" by studying aggregate penalty calls in the 1998-99 NHL season where an additional on-ice referee was deployed. The study was more concerned with theories of criminology than the game itself. In contrast, Beaudoin and Swartz (2010) developed a simulator for NHL matches where their main focus was the timing involved in "pulling the goaltender". In a side comment (Remark #2), they noted that road teams are called for more penalties than home teams in a 11:10 ratio. This observation coincides with the perception that referees are influenced by the home crowd (chapters 10 and 11 of

Moskowitz and Wertheim 2011). In contrast to our interest in biases (i.e. to which team the next penalty is assigned), Schuckers and Brozowski (2012) were primarily concerned with penalty rates in the NHL. For example, they found that penalties dramatically decrease as the game progresses, a "swallowing of the whistle" effect. In addition, they found that fewer penalties are called in tight games (i.e. as the score differential between the two teams decreases). More recently, Abrevaya and McCulloch (2014) considered a large dataset consisting of all penalties in the NHL that occurred from the 1995-1996 regular season through the 2001-2002 regular season. Using advanced statistical techniques from the machine learning literature, they observed that the next penalty call depends on a number of factors including the team that was assigned the last penalty call, the current time of the match, the time since the last penalty and the home team.

This paper builds on the work of Abrevaya and McCulloch (2014) in several ways. We consider more recent seasons in our data analysis, 2009/2010 through 2013/2014. This is important because a lockout occurred during the 2004/2005 season, and from that time forward, important rule changes in the NHL were implemented. Many of these rule changes were concerned with an attempt to emphasize the skill and speed of the game. And with harsher standards for penalties, the determination of penalties changed dramatically in the season following the lockout (Vesper 2007). Whereas Abrevaya and McCulloch (2014) observed that the next penalty is more likely to be called against the team that was not assigned the most recent penalty, we present a nuanced view of this finding. Specifically, it is the totality of penalties that have been called on each team that plays a role in the determination of the next penalty. The team which has fewer penalties in aggregate is more likely to receive the next penalty. Also, our paper investigates the tendencies of individual referees with respect to their decision making. The proposed methodology has the potential to assist in the performance evaluation of referees.

There is a secondary area of literature that is related to our work. It concerns refereeing bias in soccer. There are a number of papers that suggest that refereeing bias exists where the dependent variable is either cards assigned (yellow and red) or the number of minutes of extra time added to a game. Some of the interesting findings include an officiating bias in favour of the home team (Buraimo, Forrest and Simmons 2010 and Rocha et al. 2015) and that the composition of the crowd affects favoritism (Dohmen 2008 and Garicano, Palacios-Huerta and Prendergast 2005).

In Section 2, we begin by describing the dataset. The data was taken from the 2009/2010 through 2013/2014 regular seasons of the NHL. Hence, the results are only directly applicable to the NHL. We then carry out logistic regression analyses which investigate the effect of various covariates on the next penalty call. The results are then compared to the results obtained through a boosting algorithm and are found to be comparable. Section 3 provides a discussion of possible causes of the inferences obtained in Section 2. We distinguish between biases that are due to refereeing decisions and those that are due to situational changes in the way the game is played. Then in Section 4, we expand the investigation to individual referees by introducing a performance metric for referees. The metric introduces no additional parameters into the model. We make the distinction between referees (who call penalties) and linesmen (who call only restricted types of penalties such as "too many men on the ice" and who may sometimes *report* infractions to referees). We conclude with a short discussion and concluding remarks in Section 5.

2 ANALYSES USING LOGISTIC REGRESSION

Our penalty data were obtained from the website www.nhl.com and involve match information from the 2009/2010 through 2013/2014 NHL regular seasons. A web crawler was used to scrape the data. Considerable effort was required to convert the data to a useable format and to check for errors. We have omitted 10-minute penalties as they do not confer a manpower advantage. Although summary statistics on penalties are available, our data is more comprehensive as we obtained covariates of interest at the times that the penalties occurred. With four full regular seasons of NHL data (82 games per team) and the 2012/2013 lockout season (48 games per team), we collected penalty data for 5640 matches. Our resultant penalty data file consists of 42424 rows, where each row corresponds to a penalty call. Therefore, on average, there were $42424/5640 = 7.5$ penalties called per match.

With the above data, we fit various logistic regression models. We do not consider every penalty that occurred; only those "penalty occasions" that provided a manpower advantage. For example, consider a situation where both Team A and Team B received offsetting 5-minute penalties. Since neither team was awarded a powerplay advantage, we do not consider this as a penalty occasion. As another example, consider a situation where Team A received a 5-minute penalty and Team B received a 5-minute penalty and a 2-minute penalty. We count this as a single penalty occasion where Team A was awarded a powerplay. The resulting dataset consists of $n = 38084$ penalty occasions.

We fit logistic regression models where $y_i = 1(0)$ according to whether the ith penalty was called against the home (road) team, $i = 1, \ldots, n$. The dependent variable y_i is distributed according to $y_i \sim \text{Bernoulli}(p_i)$ and we considered the following covariates:

$x_{1i} \equiv$ total road penalties minus total home penalties in the particular match when the ith penalty was called

$x_{2i} \equiv$ total road goals minus total home goals in the particular match when the ith penalty was called

$x_{3i} \equiv$ the time in the match when the ith penalty was called; x_{3i} ranges from the 0th minute to the 65th minute which is the end of overtime

$x_{4i} \equiv$ team strength parameter where values 1/0/-1 correspond to a stronger home team, evenly matched teams and a stronger road team based on regular season points

In Table 1, we present the results from fitting various logistic regression models where each model includes an intercept term. The Akaike Information Criterion (AIC) incorporates a penalty term so that the complexity of models are taken into account. We observe that the best fitting model is the one which consists of all four covariates x_1, x_2, x_3 and x_4. We also note that most of the variability is explained by the covariates x_1 and x_2.

Model	AIC	Model	AIC
x_1, x_2, x_3, x_4	51622	x_2, x_4	52598
x_1, x_2, x_3	51628	x_2, x_3, x_4	52598
x_1, x_2, x_4	51628	x_2	52601
x_1, x_2	51634	x_2, x_3	52602
x_1, x_3, x_4	51689	x_3	52739
x_1, x_3	51690	x_4	52739
x_1, x_4	51692	x_3, x_4	52741
x_1	51693		

Table 1: Logistic regression models initially considered and the resulting AIC diagnostic where each model includes an intercept term. The models are listed in increasing order of AIC such that the best fitting models appear at the top.

Using the full model based on all four covariates, we then considered the contribution of interaction terms to see if model improvements could be obtained. The six possible interaction terms were x_1x_2, x_1x_3, x_1x_4, x_2x_3, x_2x_4 and x_3x_4. We began adding the interaction terms to the model in an attempt to improve AIC. We also dropped terms from the model that were not statistically significant. Through this iteration procedure, the "best" model is given by

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = -0.1237 + 0.4014x_1 - 0.0520x_2 - 0.0299x_4 - 0.0046x_1x_3 \tag{1}$$

which has a greatly improved AIC = 51548. The five estimates in model (1) have associated p-values $2.0 * 10^{-16}$, $2.0 * 10^{-16}$, $1.4 * 10^{-15}$, 0.0048 and $2.0 * 10^{-16}$, respectively. In terms of the implications for hockey, the partial effects are summarized as follows:

- the next penalty is more likely to be called on the road team
- the next penalty is more likely to be called on the team that has accumulated fewer penalties in the match
- the next penalty is more likely to be called on the team that is leading in the match
- the next penalty is more likely to be called on the weaker team
- as the match progresses, the effect due to penalty differential (i.e. the second bullet point) is slightly reduced

When considering the fitted logistic regression model (1), it is instructive to look at both the size of the estimated parameters $\hat{\beta}_i$ and the potential values of the associated covariates. When doing so, it is apparent that penalty differential x_1 has the greatest effect on the outcome of the next penalty call. For example, consider the case where the road team has accumulated $x_1 = 3$ more penalties than the home team, and default values are set for the remaining covariates (i.e. $x_2 = 0$, $x_3 = 30$ and $x_4 = 0$). Then the probability that the home team receives the next penalty is a stunning $\hat{p} = 0.66$. The intercept term is also of particular interest as it conveys a type of home team advantage; i.e. the rate which penalties are assigned to the home team. With all things being equal (i.e. $x_1 = 0$, $x_2 = 0$, $x_3 = 30$ and $x_4 = 0$), the probability that the next penalty is called on the home team is $\hat{p} = 0.47$. This estimate agrees with Remark #2 from Beaudoin and Swartz (2010) which notes that penalties are called on the road team in a 11:10 ratio (i.e. with probability 0.476).

2.1 A Machine Learning Approach

Whereas logistic regression has been the long-standing approach for analyzing Bernoulli data, there has been increasing interest in analyses based on machine learning algorithms in the context of large datasets. Machine learning analyses often suffer from a lack of direct interpretability. On the other hand, these modern approaches tend to provide superior predictive capability and do not require the specification of a parametric relationship involving the covariates (Hastie, Tibshirani and Friedman 2009).

The particular machine algorithm which we considered is gradient boosting which is based on an ensemble of decision trees. In our implementation, we used the *gbm* function from the `gbm` package in R (Ridgeway et al. 2015). The computations were based on the same response variable and covariates used in the analyses based on logistic regression. Without going into the details regarding our choices, the *gbm* tuning parameters were set according to shrinkage = 0.005, n.trees = 1350, interaction.depth = 2 and learning rate = 0.005.

In Table 2, we rank the importance of the four covariates using the *relative influence* variable provided by *gbm*. Relative influence is scaled so that its sums to 100 over all covariates. We observe a similar pattern as obtained in logistic regression. Namely x_1 (penalty differential) is by far the most important predictor and x_2 (goal differential) is the second most important covariate.

Covariate	Relative Influence
x_1	79.18
x_2	11.11
x_3	8.93
x_4	0.78

Table 2: The relative influence of the four covariates ranked from most important to least important based on the gradient boosting algorithm.

We now provide a comparison between the predictions obtained via logistic regression and gradient boosting. For ease of notation, the fitted logistic regression model (1) can be expressed as $\text{logit}(\hat{p}) = x'\hat{\beta}$. The expression can then be inverted to solve for the predictive value $\hat{p} = \exp\{x'\hat{\beta}\}/(1 + \exp\{x'\hat{\beta}\})$. In Table 3, we provide the predictions for various covariate settings. When focusing on a particular covariate, say x_1, we set the remaining covariates at their default values. The default values (i.e. standard values) for the four covariates are $x_1 = 0$, $x_2 = 0$, $x_3 = 30$ and $x_4 = 0$. We observe that the predictions obtained via logistic regression and gradient boosting are comparable.

3 INTERPRETATION OF RESULTS

We have observed that the team which is assessed the next penalty depends on which team is the home team, the accumulated penalty differential, the goal differential, the time of the match and the relative strengths of the two teams. There are various explanations for the effects, and we must be careful in assigning causal relationships.

The tempting conclusion is that the observations are a result of officiating biases such

Covariate	GB	LR	GB	LR	GB	LR	GB	LR
x_1	$x_1 = 4$		$x_1 = 2$		$x_1 = -2$		$x_1 = -4$	
	0.70	0.72	0.59	0.60	0.37	0.34	0.29	0.24
x_2	$x_2 = 2$		$x_2 = 1$		$x_2 = -1$		$x_2 = -2$	
	0.44	0.44	0.45	0.46	0.50	0.48	0.50	0.50
x_3	$x_3 = 10$		$x_3 = 20$		$x_3 = 40$		$x_3 = 50$	
	0.47	0.47	0.48	0.47	0.47	0.47	0.47	0.47
x_4	$x_4 = 1$		$x_4 = -1$					
	0.47	0.46	0.48	0.48				

Table 3: The predicted probabilities of Prob($y = 1$) for gradient boosting (GB) and logistic regression (LR) for specified covariate values. When focusing on a particular covariate (left column), default values are assigned to the remaining covariates.

as those that have been established in soccer. On the other hand, it is possible that teams modify their playing behavior according to the match situation.

For example, it is possible that when a team is ahead in a match, they decide to play rougher and incur more penalties. It is also possible that when a team has more penalties in a match, they decide to be more careful and are less likely to have the next penalty called against them. Personally, we find these two explanations unlikely. However, a more tenable explanation is that teams that are winning play more conservatively and spend more time in their defensive zone and with less possession. This may contribute to more penalties such as holding and hooking. To investigate this conjecture, Figure 1 provides a graph of the proportion of hooking and holding penalties plotted against the goal differential attained by the penalized team. We observe that as the goal differential of the penalized team increases, more of their penalties tend to be of the holding/hooking variety. This is indicative that leading teams begin to play cautiously, trying to preserve their lead and are perhaps less willing to venture into the offensive zone. It is suggestive that playing style may change as a team builds a lead.

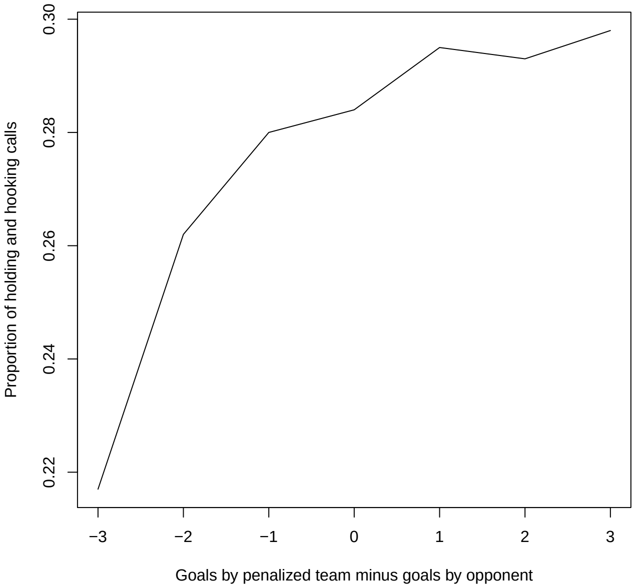

Figure 1: Proportion of holding and hooking calls plotted against the goal differential attained by the penalized team.

As a further study of how goal differential x_2 may affect playing style, Figure 2 provides a graph of the percentage of shots on goal by the home team plotted against x_2 in even manpower situations. There is a clear increasing pattern which demonstrates that as the road team's goal differential increases, the home team takes a greater percentage of shots. The implication is that teams play more cautiously and play more in their own end as their lead increases.

Analogous to Figure 2, Figure 3 investigates how penalty differential affects playing style. Figure 3 provides a graph of the percentage of shots on goal by the home team plotted against x_1 in even manpower situations. Unlike Figure 2, there is no clear pattern in Figure 3 and this suggest that the effect on penalty calls due to penalty differential can be mainly attributed to officiating.

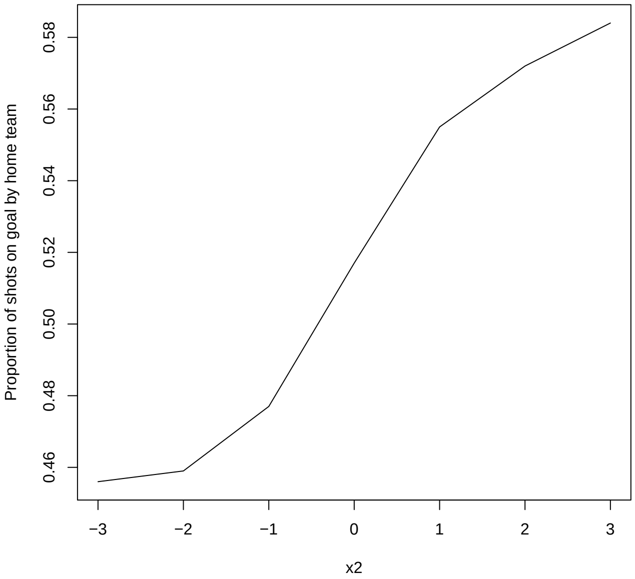

Figure 2: Proportion of shots on goal by the home team versus goal differential x_2.

The logistic model (1) also suggests that penalties are more likely to be called on the road team (intercept term) and on the weaker team (x_4). It is plausible that these observations may also have an explanation involving playing style, and are not exclusively attributed to refereeing bias. For example, both road teams and weaker teams may feel that their opportunity to win is diminished, and that playing a conservative style and preventing goals is in their best interests. As mentioned above, a conservative playing style where one has less possession, may lead to increased penalties.

To investigate the situational changes simultaneously, we fit the same logistic model (1) to the data arising from the first half of matches (i.e. the first 30 minutes). We suspect that teams are less likely to modify their playing behavior and strategies early in the match due to situational circumstances. The first half of a game is too early for

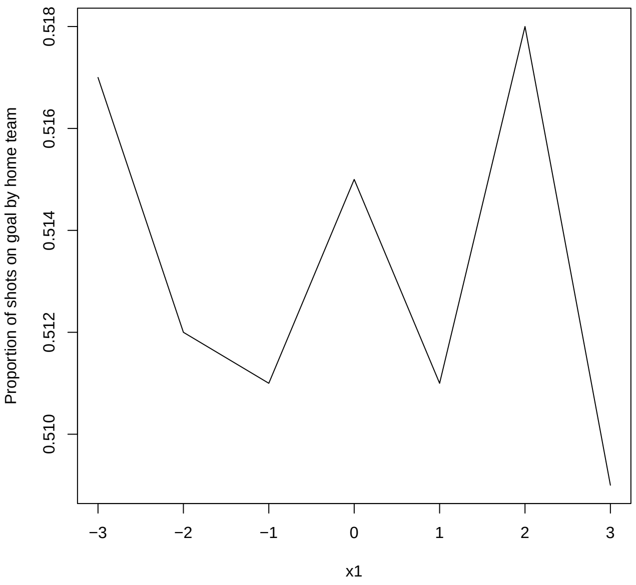

Figure 3: Proportion of shots on goal by the home team versus penalty differential x_1.

teams to think about preserving leads and playing more cautiously. The first half fitted logistic regression model is given by

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = -0.1132 + 0.4886x_1 - 0.0471x_2 - 0.0426x_4 - 0.0089x_1x_3 \qquad (2)$$

The estimates in (2) are all statistically significant. Moreover, we observe only small differences between the estimates in (2) and those given in (1). In particular, the estimates in (2) are all within two standard deviations of the estimates in (1). This suggests that the situational effects on penalty calls are primarily due to officiating biases.

In summary, there are arguments that can be put forward that suggest that situational effects may influence playing style and hence the probability of receiving the next penalty. On the other hand, the similarity between the estimates in (1) and (2) suggest that refereeing bias is the major explanation for the effect of x_1, x_2, x_3 and x_4 on Prob$(y = 1)$.

4 INVESTIGATION OF INDIVIDUAL REFEREES

In this section, we use the same dataset but restrict our analysis to the 26 referees who officiated at least 300 games. Referees work together in pairs where the partner assignments are rotated throughout the season.

Referring back to the covariates considered in logistic regression, and following the discussion in Section 3, we believe that the intercept term and the x_1 term are the terms most strongly associated with officiating bias. Recall that the intercept term describes the home team advantage whereas x_1 is the accumulated penalty differential in favor of the road team. Therefore, for the evaluation of referees, we consider only these two terms, and the associated logistic regression model

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = -0.1154 + 0.2302x_1 \ . \tag{3}$$

Model (3) is simpler than the previously considered model (1) but we note from Table 2 that x_1 is the covariate which explains most of the variation in $\text{Prob}(y = 1)$. In our evaluation of referees, we state two assumptions which we posit are indicative of good officiating:

1. the home team and road team should be penalized at equal rates

2. the penalty differential should not be predictive of the next penalty

Under these two assumptions, the logistic regression model (3) reduces to

$$\log\left(\frac{p}{1-p}\right) = 0 \quad \longleftrightarrow \quad p = 0.5 \tag{4}$$

Now let $\hat{\beta}_0 = -0.1154$ and $\hat{\beta}_1 = 0.2302$ be the estimated parameters in (3). Then

$$\hat{p}(x_1) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_1)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_1)} \tag{5}$$

is the estimated probability that the home team is assessed the next penalty when the total road penalties minus home penalties is x_1.

Our proposed measure of refereeing performance has a cross-validation aspect to it where for a given referee j, we fit the same logistic regression model (3) except that we exclude all observations involving referee j. Accordingly, let $\hat{\beta}_0^{(j)}$ and $\hat{\beta}_1^{(j)}$ be the estimated parameters omitting the data corresponding to referee j. Then

$$\hat{p}^{(j)}(x_1) = \frac{\exp(\hat{\beta}_0^{(j)} + \hat{\beta}_1^{(j)} x_1)}{1 + \exp(\hat{\beta}_0^{(j)} + \hat{\beta}_1^{(j)} x_1)} \tag{6}$$

is the estimated probability (omitting referee j) that the home team is assessed the next penalty when the total road penalties minus home penalties is x_1.

Referring to (4), (5) and (6), the idea underlying our proposed performance metric is that referee j is making better than average refereeing decisions under situation x_1 if

$$|\hat{p}^{(j)}(x_1) - 0.5| > |\hat{p}(x_1) - 0.5|$$

and is making worse than average refereeing decisions under situation x_1 if

$$|\hat{p}^{(j)}(x_1) - 0.5| < |\hat{p}(x_1) - 0.5| \ .$$

The rationale is that if referee j's exclusion pulls the logistic model further away from the idealized surface, then referee j is making good decisions.

For the jth referee, we therefore propose the following statistic as a refereeing performance metric:

$$Q_j = \sum_{x_1} w(x_1) \Big(\ |\hat{p}^{(j)}(x_1) - 0.5| - |\hat{p}(x_1) - 0.5| \ \Big) \tag{7}$$

where the weight $w(x_1)$ is the proportion of penalties in the entire dataset corresponding to x_1. For example, consider the $n = 38084$ penalty occasions that are recorded in the dataset. Let $n(2) = 3883$ be the number of those occasions when the road team had $x_1 = 2$ more total penalties than the home team. In this case, $w(2) = n(2)/n = 0.102$. To illustrate formula (5), we note that $\hat{p}(2) = 0.585$. This means that when $x_1 = 2$, 58.5% of the time (which is remarkably high), the next penalty is called on the home team. Suppose that the jth referee has $\hat{p}^{(j)}(2) = 0.600$, i.e. this is the corresponding probability when penalties called by the jth referee have been deleted from the dataset. Then, the contribution to the metric Q_j due to the $x_1 = 2$ term is $0.102(|0.600-0.5|-|0.585-0.5|) = 0.00153$. The contribution is positive meaning that referee j is making better than average decisions under $x_1 = 2$. Referees for whom Q_j is positive have an above average officiating metric. Conversely, referees for whom Q_j is negative have a below average officiating metric. In (7), more weight is given to the more common situations that occur in hockey games.

Now a potential difficulty with (7) is that we do not want a performance metric that depends greatly on the number of matches officiated. And we note that (7) is based on estimates which may be sensitive to sample size. Since we have limited our investigation to only those referees who have officiated at least 300 matches, this should not be a great problem. Also, we note that the maximum number of matches officiated by any referee in our dataset is 345. However, to provide a less sensitive measure (in case we wish to consider referees with fewer matches), we suggest a variation to Q_j based on a bootstrapping-type procedure (Efron and Tibshirani 1993). For the jth referee, we resample $m = 200$ of his matches without replacement. These m matches are then excluded in the estimation of $\hat{p}^{(j)}(x_1)$ using logistic regression. For each resample of size m, Q_j is calculated, and this procedure is repeated 5000 times. The average value of Q_j is then used as the performance measure. The rationale is that every referee's performance measure is based on the same number m of excluded matches.

An alternative approach for assessing officiating performance may involve introducing an effect for each referee. However, this increases the parametrization of the model significantly.

Although the measure (7) does not have a straightforward interpretation, it does permit the ranking of referees. In Table 4, we rank the 26 NHL referees from the dataset based on the bootstrapped performance measure Q_j in (7). Tim Peel is the top referee, and at the bottom of the list lies Kevin Pollock. We also observe that Tim Peel and Ian Walsh are more extreme in the positive sense than Kevin Pollock is in the negative sense. We found that the results were not greatly sensitive to modest departures from the choice of $m = 200$.

We are able to provide some context to the performance measure (7). Consider two hypothetical referees. We define Referee A as the perfect referee and Referee B as the completely biased referee. Recall that $m = 200$ is the number of games that are excluded for referee j when calculating $\hat{p}^{(j)}$. Since there are on average 6.753 penalty occasions per match, we set the number of penalty calls in situation x_1 by Referees A and B equal to $6.753(200)w(x_1)$ rounded to the nearest even integer. For the perfect Referee A, he calls half of his penalties in situation x_1 on the home team and half of his penalties on the road team. For the completely biased Referee B, he calls penalties in the same manner as Referee A, except

- if $x_1 > 0$, he calls all the penalties on the home team
- if $x_1 < 0$, he calls all the penalties on the road team (8)

To elaborate on the decision making of Referee B, consider the first bullet point in (8). At the time of the next penalty, the road team has more penalties than the home team. To “even things up”, Referee B calls the penalty on the home team.

Augmenting our dataset with the calls made by the two hypothetical referees, we obtain $1000Q_j = 2.56$ for Referee A and $1000Q_j = -6.67$ for Referee B. These values provide some context for the results presented in Table 4. For example, it appears that some of the top referees (e.g. Peel and Walsh) are very good. That is, they are unbiased in terms of penalty differentials. On the other hand, the referees at the bottom of Table 4 are nowhere close to being as bad as they could be. This is somewhat expected and also reassuring.

To further aid in the interpretation of the Q_j values reported in Table 4, the bootstrapping approach also allows us to get a sense of the associated variability. The standard deviations from bootstrapping are given in Table 4. We see that the standard deviations are of similar magnitude. We suggest that differences of less than one standard deviations should not be regarded as highly meaningful. Therefore, perhaps Peel and Walsh are in a class of excellence by themselves. In addition, since $Q_j = 0$ corresponds to an average referee, perhaps Pollock, McCauley, Martell, Van Massenhoven and Kowal can be viewed as slightly below average. Other clear divisions are less obvious.

Furthermore, it would seem good practice for the NHL to assign their best referees to important matches. We recorded the refereeing assignments for the 18 games that consisted of semifinals and finals matches during the 2013/2014 NHL playoffs. In the NHL, there are two referees who officiate each game. Therefore there were 36 refereeing assignments for the 2013/2014 playoff games under study. Of the referees recorded in Table 4, Kozari, O'Rourke, Watson, Joannette, O'Halloran, McCauley and Pollock received match assignments. According to Q_j in Table 4, only Kozari and O'Rourke were amongst the top 10 referees. Obviously, this suggests that the NHL are using other criteria (Fraser 2013) for determining refereeing assignments in the playoffs.

Finally, let's return to the statistic (7). The statistic was proposed because we believe that the impact of the covariate x_1 on penalty calls is mainly due to refereeing bias.

However, it is plausible that a small portion of the effect is due to a change in playing style. We quantify this by saying that refereeing bias ought to be instead measured by a comparison of $|\hat{p}^{(j)}(x_1) - 0.5 + \Delta|$ with $|\hat{p}(x_1) - 0.5 + \Delta|$ where $|\Delta|$ is small relative to both $|\hat{p}^{(j)}(x_1) - 0.5|$ and $|\hat{p}(x_1) - 0.5|$. In this case, it is easily seen that

$$Q_j = \sum_{x_1} w(x_1) \left(\, |\hat{p}^{(j)}(x_1) - 0.5 + \Delta| - |\hat{p}(x_1) - 0.5 + \Delta| \, \right)$$

and (7) are equivalent if $\text{sign}(\hat{p}^{(j)}(x_1) - 0.5) = \text{sign}(\hat{p}(x_1) - 0.5)$ This provides added support for the use of (7) as an appropriate measure of refereeing bias.

5 CONCLUDING REMARKS

Two of the primary observations in this paper are that (1) teams that have taken more penalties in a match are less likely to have the next penalty called against them and (2) teams that are leading in a match are more likely to have the next penalty called against them (also noted by Abrevaya and McCulloch (2014)). Both of these observations may be suggestive of poor officiating. However, as discussed in Section 3, we need to be careful about our conclusions. It may be possible that NHL teams behave differently under different situations leading to biased penalty calls. It would be good to explore the causal relationships in more detail and attempt to disentangle the reasons behind biased penalty calls. For example, one could investigate European professional hockey leagues where the referees are European. It may be safe to assume that European teams play hockey in the same manner as NHL teams, and therefore differences in penalty calls may be strictly attributed to officiating decisions.

It may also be possible that there exist other sources of bias. For example, one could imagine refereeing bias due to teams, players, coaches, the fan-base, etc. Although we have not tested for any of these, such biases (if they exist) are more objectionable from the point of view of fair play. Hence, we suspect they are less likely to be present, at least in a way that greatly disrupts the integrity of the game. Extending our model to investigate different sources of refereeing bias is a valuable direction for future work.

The evaluation of the performance of NHL referees is a serious and detailed exercise that involves the scrutiny of film, and where the evaluation is carried out by top-level staff (Fraser 2013). Whereas there is no obvious substitution for the careful and time-consuming review process, we believe that the methodology presented in Section 4 provides a simple way to evaluate aspects of refereeing that is objective and can be replicated by anyone who possesses basic statistical expertise. Our methodology detects that not all referees are the same.

6 REFERENCES

Abrevaya, J. and McCulloch, R. (2014). "Reversal of fortune: a statistical analysis of penalty calls in the National Hockey League", *Journal of Quantitative Analysis in Sports*, 10, 207-224.

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.

Buraimo, B., Forrest, D. and Simmons, R. (2010). "The 12th man?: refereeing bias in English and German soccer", *Journal of the Royal Statistical Society, Series A*, 173, 431-449.

Dohmen, T.J. (2008). "The influence of social forces: Evidence from the behavior of football referees", *Economic Inquiry*, 46, 411-424.

Efron, B. and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*, Monographs on Statistics and Applied Probability, 57, Chapman and Hall/CRC: New York.

Fraser, K. (2013). Fraser: How officials are graded and chosen for the playoffs. In *TSN BLOGS*, http://www2.tsn.ca/blogs/kerry_fraser/?id=423337.

Garicano, L., Palacios-Huerta, I. and Prendergast, C. (2005). "Favoritism under social pressure", *The Review of Economics and Statistics*, 87, 208-216.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction, Second Edition*, Springer: New York.

Moskowitz, T.J. and Wertheim, L.J. (2011). *Scorecasting: The Hidden Influences Behind how Sports are Played and Games are Won*, Crown Archetype: New York.

Ridgeway, G. with contributions from others (2015). gbm: Generalized Boosted Regression Models. R package version 2.1.1. http://CRAN.R-project.org/package=gbm.

Rocha, B., Sanches, F., Souza, I. and da Silva, J.C.D. (2015). "Does monitoring affect corruption? Career concerns and home bias in football refereeing", *Applied Economic Letters*, 20, 728-731.

Schuckers, M. and Brozowski, L. (2012). "Referee analytics: An analysis of penalty calls by National Hockey League officials", *MIT Sloan Sports Analytics Conference, March 2-3, Boston, MA.*

Vesper, A. (2007). "What's new in the new NHL? A comparison of the effect of team performance on the probability of winning and goal scoring efficiency in the pre- and post-lockout NHL. *MA Thesis, Department of Economics, University of Georgia.*

Referee	Matches	Measure	Standard
		$1000Q_j$	Deviation
01. Peel, Tim	341	1.26	0.32
02. Walsh, Ian	343	0.96	0.30
03. Devorski, Paul	336	0.57	0.26
04. Pochmara, Brian	343	0.51	0.31
05. Dwyer, Gord	329	0.36	0.29
06. Morton, Dean	329	0.36	0.27
07. Kozari, Steve	345	0.34	0.34
08. O'Rourke, Dan	343	0.18	0.30
09. Lee, Chris	343	0.17	0.28
10. St. Pierre, Justin	318	0.10	0.23
11. Watson, Brad	342	0.07	0.32
12. Meier, Brad	343	0.02	0.28
13. Joannette, Marc	342	-0.06	0.24
14. Kimmerly, Greg	338	-0.08	0.27
15. Furlatt, Eric	340	-0.11	0.26
16. LaRue, Dennis	337	-0.12	0.20
17. Sutherland, Kelly	337	-0.13	0.24
18. Leggo, Mike	339	-0.16	0.23
19. O'Halloran, Dan	340	-0.17	0.20
20. Rooney, Chris	340	-0.20	0.26
21. St. Laurent, Francois	341	-0.24	0.27
22. Kowal, Tom	341	-0.49	0.22
23. Van Massenhoven, Don	315	-0.54	0.23
24. Martell, Rob	331	-0.63	0.22
25. McCauley, Wes	344	-0.64	0.20
26. Pollock, Kevin	340	-0.87	0.19

Table 4: Performance measures and standard deviations for referees with at least 300 games officiated during the 2009/2010 through 2013/2014 NHL regular seasons. The referees are listed according to decreasing levels of performance.