Predicting the Winners of Hockey Games

Sam Buttrey*

Abstract
This article describes a method for predicting the outcome of National Hockey League (NHL) games. We combine a model for goal scoring and yielding, and one for penalty commission, in a Markov-type computation and a simulation model that produce predicted probabilities of victory for each team. Where these differ substantially from the market probabilities, we make "bets" according to a simple strategy. Our return on investment is both positive and statistically significant.

1. Introduction

1.1 Background

The National Hockey League (NHL) consists of 30 teams, each of which plays 82 regular-season games per year. In this article we focus on four years' worth of game data, from the 2010-2011 through 2014-2015 seasons (omitting the 2012-13 season which was cut short by a lockout). These data are available from the NHL's website, `www.nhl.com`. Our goal is to use the first few years' data to predict each team's probability of winning each game in the final year's data. Where our predicted probabilities differ from the ones produced by the market, as measured by publicly available money lines, we place "bets" according to a simple strategy.

Our interest in this work grew out of the long-standing problem of "pulling the goalie." This tactic involves replacing the goalie with a sixth skater, leaving the net unprotected. It is therefore a high-risk maneuver, generally attempted in the last minute or so of a game already being lost. The literature on this question goes back to Morrison (1976) and remains active (see, for example, Beaudoin and Swartz (2010)).

Of course to determine the right moment at which to pull the goalie, it is necessary to know the rate at which goals are scored into an unguarded net, and at which goals can be scored in the 6-skaters-versus-5 scenario, and naturally these rates can be expected to vary by team. MacDonald (2012) presents evidence that statistics beyond goals, like shots attempted, make for better predictions of future goals than just goals, although his emphasis is on rating individual players. Indeed a lot of statistical work in hockey is aimed at evaluating players. For example, the well-known "plus-minus" rating is computed by comparing the number of goals scored by each team during a particular player's tenure on the ice; Corsi and Fenwick measures (NHL, 2015) are computed in a similar way from shots taken. Here we are interested only in predicting the outcome of a particular game.

1.2 Our Approach

We model the goal-production process by a pair of independent Poisson processes, producing goals at a rate that is assumed constant for each manpower situation over the course of a game. The independent Poisson model was examined by Maher (1982) for soccer, and found to be "reasonably accurate," although some improvement was seen in adding correlation. Goals in hockey can be seen as "rare events," though not quite so rare as in soccer. Thomas (2007) and others have suggested more sophisticated models that include, for example, the current score in determining the rate, but he treats the two teams' processes as

*Department of Operations Research, Naval Postgraduate School, Monterey, CA 93950

independent. We keep the assumption of independent Poisson processes as a reasonable starting point. However the goal-scoring process is modeled, an accurate estimate of each team's goal-scoring (and goal-yielding) must aid in estimating each team's probability of a win in any particular matchup, and modeling the rate at which goals are scored and yielded has been the focus of a number of studies. Beyond goal-scoring rates, a second very important factor in determining game winners is penalties. At full strength each team has six players – a goalie and five "skaters" – on the ice. (Manpower situations are named with respect to skaters, so full strength is called "five-on-five.") When a minor penalty is assessed against a team, that team plays with one fewer skater for two minutes, or until the other team scores. It is also possible for another penalty to be called during the time the first is active, in which case the manpower situation would be 5-on-3 or 4-on-4. Other situations are possible as well, of course. Teams never play with fewer than three skaters; a penalty called on a team with three skaters is delayed until one of the current penalties has expired.

Teams playing with the 5-on-4 advantage (the so-called "power play") scored at a rate of 6.0 goals per sixty minutes in 2010-11, a little more than double the 5-on-5 rate of about 2.7 goals per sixty minutes. The team playing with 4-on-5 may also score, but naturally these "short-handed" goals are rare. To predict the winner of a game, then, it would help to know both the rate at which teams score (including during penalty situations) and also how often they can be expected to commit penalties. The recent paper of Abrevaya and McCulloch (2014) demonstrates that penalties are not called uniformly at a constant rate; rather, it shows evidence of "make-up calls" and other effects. Nonetheless, the penalty commission rates here will be taken to be constant.

In this paper we describe our efforts to predict the outcomes of NHL games. This effort has three parts: first, we model goal scoring as a pair of Poisson processes with rates determined by the two teams' goal-scoring rates in the 5-on-5 and 5-on-4 situations, and their goal-yielding rates, together with a parameter to account for home-ice advantage. Second, we model the penalty-commission rate for each team as a function of its recent penalty history. Next we pool together the training and validation data to estimate "adjusters" for scoring rates in situations other than 5-on-5 and 5-on-4. Finally, we use those rates in a simple Markov-type model, as well as in a more complex Monte Carlo simulation, to predict probabilities of victory. We create a betting strategy that compares our predictions to those of the market, as indicated by money lines reported by a betting house.

In section 3 we lay out the models for game predicting and the details of the simulation; section 4 describes how we determined the market prices and betting strategy. Finally, section 5 gives results, conclusions, and directions for further inquiry.

2. Data and Models for Goals and Penalties

This paper considers four years' worth of data. The first two, from the 2010-11 and 2011-12 seasons, are the training data, used to determine certain parameters of the models decribed below. The 2012-2013 season was shortened because of a lockout and is ignored here. The third season, 2013-14, is used with the parameters created in the training set in order to produce the betting strategy (although we also pooled that season with the first two for computing scoring rates at different manpower situations). We will refer to this season's data as the "validation set." Finally, data from the 2014-15 season were not used for anything other than evaluating the betting strategy. These become our test set.

2.1 Goal Data

Data for this effort were extracted from the NHL website in the form of "Game Summary" pages for each of the 1230 games played every year. For the 2010-2011 season, one game summary was empty and that game was omitted. These summary pages list all of the goals scored during a game with the manpower situation in force at the time (but see "complications" below), and all the penalties assessed in the game. The game summaries also give the amount of time spent by each team in each manpower situation. An example of a game summary, regarding the game between the Columbus Blue Jackets (CBJ) and Anaheim Ducks (ANA) played at Anaheim on November 19, 2010, can be found at `http://www.nhl.com/scores/htmlreports/20102011/GS020282.HTM` (NHL, 2010). This game was won by the Blue Jackets in regulation, 4-3. We will refer to this example game, using the NHL's convention, as "game 282."

2.2 Complications

A number of complications present themselves. First, regular-season games that are tied at the end proceed to overtime. If a game is tied after 60 minutes, a five-minute overtime is played. If neither team scores in the overtime, the tie is broken with a shootout procedure. The team that wins the shootout wins the game by a final score that is reported as if the shootout had produced exactly one goal. Overtime goals go into our goal-production models, but the "goals" awarded in the shootout are not of interest to us and we ignore them.

A second complication involves empty-net goals. As we noted above, it is common for a team that is losing late in the game to replace its goalie with a sixth skater. This increases a team's chances of a quick goal, but, by leaving its own goal unprotected, greatly increases the chance of its being scored on. In game 282 the Ducks played with an empty net for 1:03. The league reports goals scored into an empty net (there were 228 in 2010-2011, representing about 3.4% of goals). It also reports the total amount of time the net was empty. The league does not directly identify goals scored in the 6-on-5 situation, though this information is accessible in an indirect way. (It also does not record the period in which the net was empty; occasionally a goalie will be replaced during a delayed penalty before the third period.) For our purposes we ignore empty-net goals entirely, and treat the rare 6-on-5 scores as if they were scored in the usual 5-on-5 manpower situation.

Finally, penalty shots are awarded for infractions like tripping a skater who is "on a breakaway." In the 2010-11 season, 27 goals arose from the 77 penalty shots imposed. We believe that a penalty shot does not inform us about scoring rates, so we ignore them for the purposes of modeling goals.

2.3 Model for Scoring and Yielding Goals

We can now describe our model for the scoring and yielding of goals. This modifies a related model presented in Buttrey, Washburn, and Price (2011). Each game is broken down into intervals of constant manpower. Our model produces individual estimates for each team for the common 5-on-5 and 5-on-4 manpower situations, and global adjusters for the other ones. Each interval appears twice in the data set, one with each team as "scorer" and the other as "yielder." For example, table 1 shows the intervals from game 282. In that game, the two teams played at 5-on-5 for 48 minutes, 42 seconds, during which time Anaheim scored one goal (top row) and Columbus three (second row). Anaheim played 5-on-4 for 5:36 during which no goals were scored (third and fourth rows), and 5-on-3 for 0:35, during which it scored one goal to Columbus's none (seventh and eighth rows).

Columbus played 5-on-4 for 5:07 and scored one goal during that time (sixth row), but Anaheim scored a "short-handed" goal as well during that interval (fifth row). So this game contributes six observations to the data set; the seventh and eighth rows are not put into the model because they involve one of the rare manpower situations. Finally, where Anaheim is the Scorer we have set Home = True.

Table 1: Intervals for Game 282

Duration	Scorer	Yielder	Goals	Home	Used
48:42	ANA.5v5	CBJ	1	True	True
48:42	CBJ.5v5	ANA	3	False	True
5:36	ANA.5v4	CBJ	0	True	True
5:36	CBJ.4v5	ANA	0	False	True
5:07	ANA.4v5	CBJ	1	True	True
5:07	CBJ.5v4	ANA	1	False	True
0:35	ANA.5v3	CBJ	1	True	False
0:35	CBJ.3v5	ANA	0	False	False

Each manpower situation contributes two segments per game, one in which the home team is the scorer (so that the "Goals" column gives the number of goals scored during that manpower situation by the home team), and another in which it is the yielder (where "Goals" refers to scores by the visitors). We built a table like table 1 using all the data from the 2010-11 and 2011-2012 seasons (but to predict any individual game we use only data recorded before that game was played; see "weights for goal data" below). On average, about 96% of minutes in hockey games are spent in the 5-on-5 or 5-on-4/4-on-5 situations. The small number of minutes for other situations makes scoring rate estimates for these situations highly variable. (For example, only one 3-on-3 goal was scored in the 2010-11 and 2012 seasons combined.) So rather than try to estimate scoring rates for each of the nine possible manpower situations for each team, we focus only on the 5-on-5 and 5-on-4 scoring rates. We also estimate a single yielding rate for each team. The "Used" column in table 1 shows the set of intervals that serve as input to our model. Then we estimate "adjusters" for the other manpower situations by pooling scoring rates in those situations across all teams and all games in 2010-2011, 2011-2012, and 2012-2013 seasons.

In detail, our model says that the number of goals scored in unit time is a Poisson random variable whose rate (expressed in goals per 60 minutes) depends on the scoring and yielding rates of the two teams involved plus the home-ice effect. In game 282, for example, the model says that the number of goals Columbus should score in sixty minutes of 5-on-5 play has the Poisson ($\lambda_{CBJ.5v5,ANA,Home}$) distribution, and more generally, the number of goals Columbus should score in t minutes is Poisson ($(t/60)\lambda_{CBJ.5v5,ANA,Home}$), where

$$\log(\lambda_{CBJ.5v5,ANA}) = \mu + \alpha_{CBJ.5v5} + \beta_{ANA}. \tag{1}$$

Here μ denotes the intercept, $\alpha_{CBJ.5v5}$ is the parameter that describes Columbus's 5-on-5 scoring rate, on the log scale, and β_{ANA} describes Anaheim's goal-yielding rate. During the same t minutes, the model says that number of goals scored by Anaheim is Poisson ($[t/60]\lambda_{ANA.5v5,CBJ,Home}$), where

$$\log(\lambda_{ANA.5v5,CBJ,Home}) = \mu + \alpha_{ANA.5v5} + \beta_{CBJ} + \delta, \tag{2}$$

and the final parameter δ measures the home-ice advantage. The parameters in the model are estimated using R (R Core Team, 2013) and the add-on `glm2()` package (Marschner, 2014).

This model, then, is characterized by three parameters for each team (a 5-on-5, a 5-on-4, and a Yielder), plus an overall intercept and a single home-ice parameter that is constant for all the games on a particular date. (The home-ice parameter changes from day to day, since we re-estimate all the parameters in the model for each new day's games.) Of course this model could be extended to specify a particular "Yielder" parameter for each team's performance in, for example, the 4-on-5 situation, but we have decided to use a single parameter to keep the number of parameters to a manageable size. (Furthermore there is an identifiability issue, since all 5-on-4 intervals are played against 4-on-5).

For manpower situations other than 5-on-5 and 5-on-4, we use as "adjusters" estimates of the manpower effects computed by pooling together the games in the training and validation sets. The yielding estimate is kept constant across all situations. As an example, the scoring rate in the 4-on-5 situation in the training data is about 34% of the 5-on-5 situation. So when Columbus is in the 5-on-4, its scoring rate is estimated directly as a model parameter, while Anaheim's scoring rate during that time is taken to be 34% of its 5-on-5 rate. If Columbus were in the 4-on-3 situation against Anaheim, we would predict its scoring rate to be 5.2 times its 5-on-5 rate, while Anaheim's scoring rate during that time would be estimated at 0.092 times its 5-on-5 rate, since overall in the training data, goals were scored in the 4-on-3 at 5.2 times the 5-on-5 rate, and in the 3-on-4 at 0.092 times the 5-on-5 rate. Since 3-on-3 minutes are limited, we combine the goals and minutes for 3-on-3 and 4-on-4 to compute one adjuster for both situations. We elected to adjust 5-on-5 rates, rather than 5-on-4, since the former are the most accurately estimated. The set of manpower adjusters is given below in table 2. (As we will see below, the Markov model considers only four situations.) We note that the 3-on-5 scoring rate, while only 11% of the 5-on-5 rate, is estimated to be slightly larger than the rate for 3-on-4 in these data.

Table 2: Manpower Adjusters

Parameter	Markov	Simulation
Adjuster: 4v5	0.34	0.34
Adjuster: 4v4	1.16	1.16
Adjuster: 5v3		9.07
Adjuster: 4v3		5.19
Adjuster: 3v3		1.16
Adjuster: 3v4		0.09
Adjuster: 3v5		0.11

2.4 Weights for Goal Data

We can fit this model for a whole season's worth of data, but in practice when we make a "live" prediction about a particular game we would be restricted to using data gathered before the game in question. For example, to predict game 282, we should use only data up through November 18^{th}, 2010. Naturally we would only use data from the 2010-11 season, since teams can change substantially from year to year. Furthermore, it is plausible to suggest that data farther into the past ought to carry less weight than more recent data. So we give games played n days ago a weight w_n that decreases exponentially with distance into the past, so that $w_n = g^n$. We describe how we pick this parameter below.

2.5 Penalty Data and Model

A second model produces the estimates of the rates at which teams incur penalties. More precisely, we are interested in the rate at which penalties give one team a manpower advantage. There are a number of complications here. First, a large number of penalties appear in pairs, one for each team. These are so-called "coincident penalties." Coincident major penalties (usually for fighting) result in two players being penalized, but in no change in the on-ice manpower situation. So we discard pairs of penalties that are assessed at the same time and for the same duration but on opposite teams. We also discard coincident minor penalties, even though they really do lead to a simultaneous reduction in on-ice manpower. We also discard ten-minute penalties that, again, are routinely coincident and generally do not change the on-ice manpower situation.

Another complication ensues when a player is called for a double-minor, which is two minor penalties called against one player at one time. They are served consecutively, the second starting when the first ends (whether by goal or expiration). We treat these as if they were two unrelated penalties.

When a player commits an infraction that leads to a penalty shot, on-ice manpower is not reduced, whether the penalty shot succeeds or not. There were 77 penalty shots awarded in the 2010-2011 season, and this number is small enough that we have ignored these penalties as well. We are left with a data set of unpaired minor penalties by date and team. In 2010-2011, there were, by our count, $8,474$ such penalties, compared to a total of $6,573$ goals that season.

Our model for penalties once again considers these events to be "rare" and depends on the Poisson (as used, for example, in Beaudoin and Swartz (2010)). As with the goals model, we establish an exponential decay parameter d such that weight associated with data from n days ago is $v_n = d^n$. Then our estimate of the rate at which Columbus, say, will commit penalties on November 19^{th} is $\sum v_n p_n / \sum v_n$, where p_n is the number of penalties committed by Columbus n days before.

3. Models for Game Winners

Given a set of goal-scoring and penalty parameters, we are ready to estimate the probability of each team winning. We describe two models; first, the simpler analytic approach, and then the simulation that incorporates some features that are intractable analytically and lead to better performance. (A logistic regression model, of intermediate complexity, did not perform as well as the simulation and is not further discussed here.)

3.1 Analytic Markov Model

For the first model we are grateful to Alan Washburn, who formulated it and programmed one version of it. Each team is represented by a binary Markov process where it is either penalized or not at every time. Let λ and μ give the rates at which penalties are called and expire, respectively. Let $P(t)$ be the probability of being penalized at time t, with $P(0) = 0$. $P(t)$ must satisfy the forward differential equation $P'(t) = \lambda - (\lambda + \mu)P(t)$, the solution of which is

$$P(t) = \frac{\lambda - \lambda \exp(-(\lambda + \mu)t)}{\lambda + \mu}.$$

Now we add home and visitor team subscripts H and V to λ, μ, and P, and assume that the two penalty processes are independent. There are four possible configurations at each time: 5-on-5, 4-on-5, 5-on-4, and 4-on-4. The probability of a 4-on-5 situation at time t, for example, is $P_H(t)(1 - P_V(t))$. From these expressions the average amount of time spent in each of the four configurations can be computed by integration. Let $f(x) \equiv (1-\exp(x))/x$. Then, for example, the mean amount of time spent in the 4-on-5 configuration over the interval [0,T] is

$$\begin{aligned} \tau_{4v5} &= \frac{\lambda_H \mu_V \{1 + f([\lambda_H + \mu_H] T)\}}{(\lambda_H + \mu_H)(\lambda_V + \mu_V)} \\ &+ \frac{\lambda_H \lambda_V \{f([\lambda_V + \mu_V] T) + f([\lambda_H + \mu_H + \lambda_V + \mu_V] T)\}}{(\lambda_H + \mu_H)(\lambda_V + \mu_V)} \end{aligned}$$

Naturally, $\tau_{5v5} + \tau_{4v5} + \tau_{5v4} + \tau_{4v4} = T$. If these times are multiplied by the scoring rates estimated above and summed, the result is the average number of goals scored in regulation by each team. If we now assume that the actual number of goals scored is a Poisson random variable, then computing the winner is a straightforward matter of comparing two such random variables to find the probability that one is larger over sixty minutes, with a provision for breaking ties.

3.2 Monte Carlo Simulation

Our Monte Carlo simulation is more accurate, but less tractable, than the Markov model decribed above. With a set of parameters for a particular game already estimated, we first simulate the penalties for each team from the Poisson distribution with the relevant rate. Those penalties are (assumed to be) distributed uniformly across the sixty minutes of the game. Suppose for a moment that all penalties are called against Columbus and that no two penalties overlap.Then for each penalty we simulate a time-until-goal from the exponential distribution (call it E_i for penalty i). Penalties for which $E_i \geq 2$ do not lead to goals; they represent a period of two minutes of 5-on-4 during which no power-play goal is scored. Penalties for which $E_i < 2$ result in power-play goals. Each of those penalties represents E_i minutes of 5-on-4 time, plus $2 - E_i$ of full-strength play after the goal. The total 5-on-4 time is $T = \sum_{Ei \geq 2} 2 + \sum_{Ei<2} E_i$. So, having recorded any power-play goals, we generate regular-strength goals over the 5-on-5 time, which has duration $60 - T$, for each team, and then we generate short-handed goals for Columbus at their 4-on-5 rate over a period of duration T. The distribution of goals seems intractable in this case. When penalties overlap, the computation becomes even more tedious.

Our code "disentangles" overlapping penalties by separating the duration of the penalties into separate pieces of constant manpower, with penalties expiring with no score at the proper rate, and with power-play goals causing other manpower situations' durations to be increased in the proper way. Penalties that do not overlap with any others can be handled comparatively quickly, as above. Then the sums of regular-strength, power-play, and short-handed goals produce the total score. We discard penalties scheduled to be assessed against a team with three skaters. (In the NHL, such a penalty is delayed, and only imposed when an existing penalty expires.) When a game ends in a tie, we resolve a tie by playing five minutes of 4-on-4, as in an NHL overtime, and, if necessary, simulate the shoot-out with a coin flip. However, our overtime dismisses any penalties that might have been called within the last two minutes of regulation time, and we do not assess penalties in overtime.

The Monte Carlo simulation provides a framework in which we might incorporate even more realism. For example, we know that scoring rates are not exactly constant across all minutes, and that they decrease late in tied games. We might include a "make-up call" rate in penalties in accordance with Abrevaya and McCulloch (2014), and so on, and it is straightforward to include variations like these in the simulation. The simulation might also be useful for estimating attributes of events other than wins, like the probability of going to overtime or the distribution of the total number of goals.

The simulation is implemented in R. Starting with November 1 each year, we simulate each game $10\,000$ times. (This start time of about a month after opening day is to collect enough data to make parameter estimates meaningful.) Our estimate of a team's winning probability is simply the fraction of those replications in which that team won. We will denote these estimates for the home and visiting team as s_h, s_v, the "s" evoking "simulation." Since every game has a winner, $s_h + s_v = 1$. If the simulation's "true" probability of predicting an Anaheim win is s, then, since the trials are independent, the standard error associated with our estimate of s is the usual $\sqrt{s(1-s)/n}$, which takes its maximum at $s = 0.5$. With $n = 10\,000$, the maximum value of the standard error is 0.005. It takes about 40 seconds to run $10\,000$ replications for a single game, on a powerful desktop computer, so we can run a season of about $1\,000$ game predictions overnight.

Figure 1 shows the stability of the simulation over one set of 10,000 replications of game 282 (using parameters selected by the method of the next section). The dots show the proportion of Columbus wins in each set of 500 replications. Over the entire set, the proportion of Columbus wins was about 60.1%; the dots exhibit a standard deviation around that number of about $\sqrt{.601(.398)/500)} = 0.022$. The solid line shows the cumulative proportion settling down to 60% as the number of replications gets large.

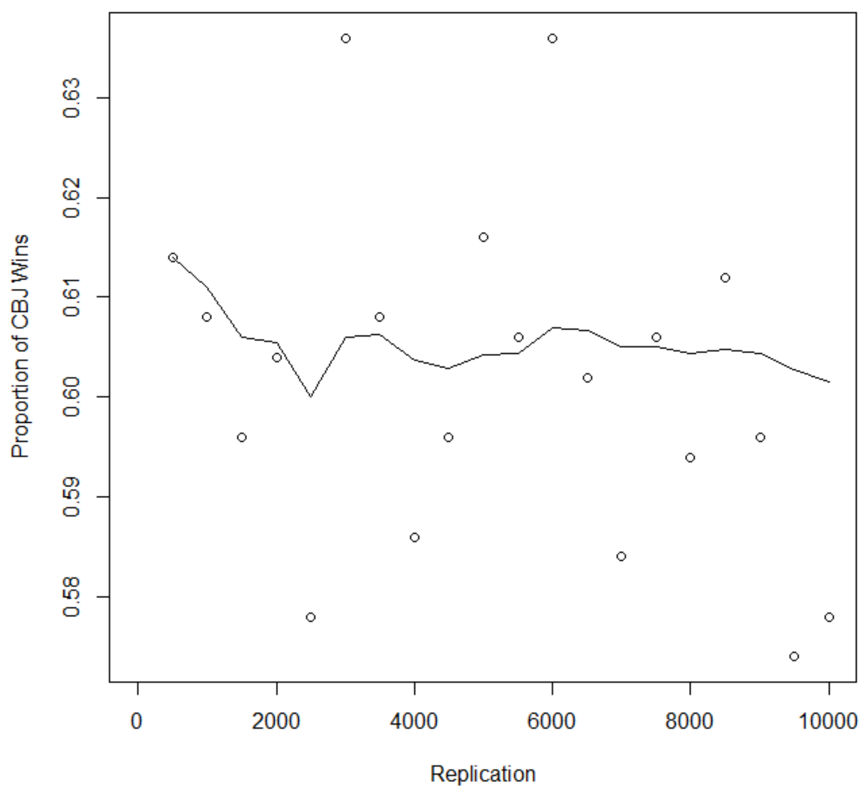

Figure 1: Stability of simulation estimator. Dots show proportion of Columbus wins in 10,000 simulations of game 282 for each set of 500 replications; the solid line shows the cumulative proportion over the entire set.

3.3 Selecting the Decay Parameters

We used the 2010-11 and 2011-12 seasons as training data to select the decay parameters g and d. For each of the 121 pairs of values of $g \in .90, .91, \ldots, 1.00$ and $d \in .90, .91, \ldots, 1.00$, we ran our two models for all games starting on November 1 of each year. For each combination of g and d, we computed the rate at which model predictions of game outcomes (selecting the team whose predicted probability is > 0.5) agreed with the game's actual winner. This provides a basic measure of accuracy of prediction. We also computed the area under the predictions' receiver operating characteristic curve (the "AUC"), using the pROC package in R (Robin, Turck, Hainard, Tiberti, Lisacek, Sanchez, and Mller, 2011). This helps us judge whether the numeric probabilities of a home win produced by the model serve to discriminate between actual home wins and home losses. We ranked each pair of values on both of these criteria, from best to worst, and chose the settings that minimized the sum of the ranks of the two measures. This led to slightly different choices of g and d for the different models. For the Markov model, the best choices were $g = 0.98$ and $d = 0.90$, whereas for the simulation model the best combination was $g = 0.99$ and $d = 0.91$.

To recap, then, here is the approach to predict the games played on date $k + 1$. We select all of the 5-on-5 and 5-on4/4-on-5 rows of the dataset like the one in table 1, for all the games played in a particular season up through date k. To all the entries for games played on day $k - n$ we assign weight g^n. Then we fit our Poisson model and extract the relevant scoring and yielding parameters. For example, Game 282 was played on November 19, 2010, so in order to predict the outcome of that game we accumulate all the 5-on-5 and 5-on-4/4-on5 data up to November 18, apply the weighting, and estimate the parameters for each team. We use $\alpha_{ANA.5v4}$ and β_{ANA} as the baselines for scoring and yielding, respectively, just because "Anaheim" is the first place name in alphabetical order.

In this case, and using the value $g = 0.99$ selected for the simulation model, the parameter estimates (on the log scale, per 60 minutes) are $\mu = 2.08$, $\alpha_{CBJ.5v5} = -0.99$, $\alpha_{CBJ.5v4} = -.91$, $\alpha_{ANA.5v5} = -1.33$, $\alpha_{ANA.5v4} = 0$, $\beta_{ANA} = 0$, $\beta_{CBJ} = -0.10$, and $\delta = 0.069$. That is, Columbus is predicted to score faster in the 5-on-5 ($-.99$ to -1.33), but slower in the 5-on-4 ($-.91$ to 0), than Anaheim, and exhibits lower yielding (that is, stronger defense). The home-ice coefficient is positive, as expected, although it is not contrained to be and in fact is sometimes negative. (The coefficients to be used for the games of November 19, 2010 under the Markov model are, as mentioned, slightly different because the "best" value of g is slightly different for that model.)

Similarly, we compute estimates of the penalty rates by applying weight d^n to games played on day $k - n$. In this example, using the value $v = 0.91$ selected for the simulation model, the Anaheim rate is estimated to be 3.94 penalties per 60 minutes, and the Columbus rate to be 3.51. Table 3 shows the set of decay parameters used in the models.

Table 3: Decay Parameters

Parameter	Markov	Simulation
Decay Rate: Goals	0.98	0.99
Decay Rate: Penalties	0.90	0.91

3.4 Evaluating the Models

Each model produces a predicted probability of winning for each game. As we did when selecting the decay parameters, we evaluate the models based on the proportion of correct predictions and on the area under the ROC. We expect better performance on the training set than on the validation set consisting of 2013-2014 data, but in fact the Markov model produces a slightly higher AUC on the validation set. In every case the simulation performs better than the Markov model.

Table 4: Model Performance

Data and Criterion	Markov	Simulation
Training set Correct rate	54.5%	57.2%
Training set AUC	55.4%	57.6%
Validation set Correct rate	55.1%	55.7%
Validation set AUC	55.1%	56.6%

4. Data and Strategy for Betting

4.1 Earlier Work on Hockey Betting

Although gambling in other venues, like horse-racing tracks, has been studied for some time, the first look at betting on NHL games appears to have been the paper of Woodland and Woodland (2001), which considered data from the 1990-91 season through 1995-96. They describe the market at that time as being "in a state of a flux, with only minimal participation." Their primary finding was that, in contrast to horse-racing, but in line with betting on other sports, the hockey market tended to overbet favorites and underbet underdogs. They find that betting on high-odds underdogs might produce, on average, positive returns. This result was seconded by Gandar, Zuber, and Johnson (2004), who while challenging some of the details in the earlier work nonetheless found the same inefficiency. This runs counter not only to intuition but to the work of (Sauer, 1998), who found that "standard definitions of [sports betting] market efficiency are satisfied. . . [although] there are empirical regularities that are inconsistent with generic notions of efficiency." Of course we might also expect that as hockey betting becomes more popular, and more bettors and more sportsbooks enter the market, opportunities to exploit market efficiencies might shrink. The recent work of Paul and Weinbach (2012) suggests that the market may be closer to perfect efficiency. They do note the persistence of bettor biases, like, for example, betting more on visiting favorites than would be expected based on the odds. They also observe that sportsbooks do not set prices to "balance the book," that is, to ensure that betting volume is equal on both sides of any bet. Instead the prices seem to arise as a forecast of game outcomes, which, they assert, leads to profit for the sportsbook even in the face of bettor biases. Of course our intent, as bettors, is to identify the games where the sportsbook's predicted probabilities deviate from our own estimates, and to bet on those.

4.2 Money Lines

To establish a betting strategy we need to be able to compare our models' predictions with the values specified by the market. We were able to obtain these from the useful website `www.covers.com`, to which we are grateful. As is conventional in North American markets, odds are reported in the "money line" style. In this style, betting favorites' odds are represented as a negative number giving the size of the bet required to win 100 units, while underdogs' odds are represented as a positive number giving the payoff for a 100-unit bet. For example, in our game 282, the November 19, 2010 game between the Ducks and Blue Jackets, the favored Ducks were listed at -127, indicating that a bettor would need to wager 127 units to win 100, or equivalently, win 78.7 units on a bet of 100. Meanwhile the underdog Blue Jackets were listed at 117, so a bettor would win 117 units on a bet of 100. If the two teams are offered at the same value, both teams will usually be listed at -105; in that case the game will have two favorites. Conventionally the money lines end in 0 or 5; we suspect the ones reported here to be averages from a number of providers.

In the example of game 282, the market is assigning a "probability" of 127/227 = 55.9% to the Ducks, and of 100/217 = 46.1% to the Blue Jackets. (We put "probability" in quotation marks because these two percentages sum to a number greater than 100%, which is where the bookmakers make their profit.) As a second example, if two teams are judged by the market to be equally likely to win, the money lines will typically be -105 for each team, corresponding to "probabilities" of 51.2%. The home and visitor "probabilities" from the market are denoted k_h^*, k_v^*, where the k evokes "market" and the asterisks indicate that these are not probabilities. (As a reminder, we always have $k_h^* + k_v^* > 1$, but $s_h + s_v = 1$.) The range of the k's is a bit smaller than of our s values; the most extreme money lines in the two-year training set were for the game of March 31, 2012 between the home Blues (money line -340, $k_h^* = 77.3\%$) and the visiting Blue Jackets (money line 305, $k_v^* = 24.7\%$), a game which, as it happens, the Blue Jackets won. In contrast we had twenty teams over the two years whose predicted win probabilities under the simulation were $\geq 80\%$. Our prediction for the March 31 game, incidentally, was $s_h = 80.3\%$ in favor of the Blues.

There will normally be some change between the initial, "opening" line and a line offered closer to the time at which the game is played. Our intent is to compete, as it were, with the opening line. Changes in the line may reflect external information like late injuries, trades, goaltender changes, or other game-specific information encapsulated in the "smart money." Unfortunately, we suspect that our betting line data reflect these later lines. It is possible to purchase historical data on opening lines, but we have not found any that are free and in the public domain. We might expect that our results would be slightly improved if we were comparing our strategy to the opening line.

4.3 Betting Strategy

Now our betting strategy can be described. We constructed the strategy using the probabilities predicted on the 2013-2014 validation data. These probabilities were constructed using the decay parameters selected with reference to the training set. For each game we have the predicted probability from the model, and the market "probabilities." A natural starting point is to place a bet whenever the former exceeds the latter – in our notation, when $s_h - k_h^* > 0$ or $s_v - k_h^* > 0$. More generally, to ensure that we are not misled by variability in the estimates s, we consider a betting strategy with one additional parameter ϵ; we then place a bet when $s_h - k_h^* > \epsilon$ or $s_v - k_h^* > \epsilon$.

Following Gandar et al. (2004), we always bet 100 units. On an underdog the net profit for a winning bet is the money line; on a favorite, the net profit is $-100 \times 100/(\text{money line})$. For example, in game 282, our estimates of the teams' win probabilities were $s_h = 40.0\%$ on the home Ducks and $s_v = 60.0\%$ on the visiting Blue Jackets. Since the market "probabilities" were $k_h^* = 55.9\%$ for the Ducks and $k_v^* = 46.1\%$ for the Blue Jackets, we place a bet of 100 units on the underdog Blue Jackets (unless ϵ is large). The Blue Jackets won that game and our wager would have returned 217 units, 117 profit. In the case of the March 31, 2012 with the large lines, our estimated win probability of 80.3% was greater than the market value of $k_h^* = 77.3\%$ for the Blues. We would have bet 100 units on the Blues, hoping to make a profit of $-100 \times 100/-340 = 29.4$ units – but in this case the bet would have lost.

We select ϵ using the 2013-2014 data, by finding the value associated with the largest return on investment (ROI, measured as net return divided by cost). For each value of $\epsilon = 0, 0.01, 0.02, \ldots, 0.20$ we compute the ROI associated with applying that value of ϵ tp the 2013-2014 data. For the simulation model, the best choice turns out to be $\epsilon = 0.02$, for which value the ROI is 4.07%. The Markov model's ROI tends to increase with increasing ϵ, presumably because accuracy can be higher when the number of bets is smaller. We select the first local maximum, which is where $\epsilon = 0$. Finally, we apply those values of ϵ to the 2014-2015 data.

5. Results and Conclusions

Table 5 shows the results (rounded to the nearest dollar) when our betting strategy, with $\epsilon = 0$ for the Markov model (top panel) and $\epsilon = 0.02$ for the simulation (bottom panel), is applied to the test set consisting of the 2014-2015 data. These data were used for no other purpose during the model-building and parameter estimation stages. Each row is labeled by a category of games ("dogs" and "faves" are short for "underdogs" and "favorites"). Each row shows, first, the number of teams in each category (Opp, giving the number of opportunities to bet) and the number of bets made (Bets). Recall that in some games both teams are "favorites" for our purposes. Subsequent columns give the total cost of those bets (Cost), the number of bets on which we were correct (Correct), the winnings we would have received (Winnings), and the net gain (Net). Note that our "winnings" include the stake; so if we were to bet 100 units and win 135, for example, our winnings would be 235.

We compute the ROI by comparing the total net return (final column) to the total in the Cost column.Thus the overall ROI for the Markov model in this test set is $-1,851/99,000$, a loss of -1.87%. For the simulation model, we have ROI = $1,921/74,200 = 2.59\%$. This is certainly a positive result, particularly because we developed all model parameters without reference to this test set. (Of course, we do re-fit the goal and penalty models before every day's games, but the decay parameters and manpower adjusters are kept fixed at the values in table 3). However, we have to note that this betting scheme would have lost money in 2010-11 and 2011-12. Despite our efforts, there may be some over-fitting aspect of our technique for which we have not accounted.

Table 5: Results on 2014-2015 Test Set

Markov Model	Opp	Bets	Cost	Correct	Winnings	Net
Home Dogs	305	245	24,500	106	24,930	430
Vis Dogs	721	685	68,500	272	67,079	−1,421
Home Faves	774	19	1,900	9	1,738	−162
Vis Faves	358	41	4,100	18	3,402	−698
Simulation	Opp	Bets	Cost	Correct	Winnings	Net
Home Dogs	305	85	8,500	39	8,646	146
Vis Dogs	721	297	29,700	128	31,541	1,841
Home Faves	774	217	21,700	134	21,575	−125
Vis Faves	358	143	14,300	84	14,359	59

To establish the extent to which these results are statistically significant, we can compare them to the results we might have seen if we had selected teams at random. Since the simulation model bet on 742 games, we selected 742 games from the 2014-15 season uniformly at random, then selected a team from that game uniformly at random on which to bet. (We do this in two steps to ensure we never bet on both teams in the same game.) We follow the scheme of betting 100 units on underdogs and the negative of the money line on the favorite. In 1,000 replications of this random-betting scheme, we averaged about a -13% ROI, with a maximum value of -5.9%. So both models produce ROIs better than could have been expected at random. It is also interesting to notice that the models emphasize certain bets and avoid others. Table 6 shows the odds as presented in the $1,079$ games under consideration in the 2014-15 season, in approximately equal bands. For each band we show the number of betting opportunites, the number of bets made, and the corresponding rate. The overall betting rate for the simulation model was $742/(2 \times 1079) = 34.4\%$, but the individual rates in the table are more spread out than they would have been if bets had been selected at random; in only one of our $1,000$ simulations did we observe a bigger SD across the rates than we see in table 6. It appears that, in general, visiting teams provide better betting opportunities than home ones.

Table 6: Betting Opportunities and Rates

	Opportunities		Bets		Rate	
Odds	Home	Visitor	Home	Visitor	Home	Visitor
-530+	156	27	48	14	30.8	51.9
-200+	146	55	47	26	32.2	47.3
-165+	136	43	33	17	24.3	39.5
-145+	118	50	37	19	31.4	38.0
-130+	118	83	27	28	22.9	33.7
-115+	100	100	25	39	25.0	39.0
100+	82	90	24	41	29.3	45.6
110+	50	81	18	34	36.0	42.0
120+	71	174	23	77	32.4	44.3
140+	63	175	17	74	27.0	42.3
170+	39	201	3	71	7.7	35.3
Total	1079	1079	302	440	28.0	40.8

5.1 Conclusions and Future Work

We have established a method for predicting the outcomes of NHL games that seems to be able to do better than the market. On any day, we model goal-scoring by a Poisson model that uses data up through the preceding day to estimate rates for 5-on-5 and 5-on-4 scoring, goal yielding, and home-ice advantage. Data farther in the past is downweighted by an exponential decay parameter whose value is selected using the training set. Goal-scoring rates for other manpower situations would be too variable in this model, and are instead computed from league-wide data as adjusters to the full-strength rates. A second effort models penalty commission rates (after excluding paired penalties) using past penalty data, again with a parameter that reduces the effect of observations farther into the past. This model presumes that unpaired minor penalties are generated by independent Poisson processes.

These scoring, yielding and penalty rates are then used in models to predict game outcomes. A simple Markov-based model, which considers only four manpower situations, is not quite as successful as a simulation, which we also expect canb be extended to capture more detail. The models produce estimates of win probabilities that, while not particularly accurate individually, seem to be able to be used in a profitable betting strategy. Our betting strategy applied to the 2014-15 season's data produced a return on investment of 2.6% – although, disappointingly, that same strategy would have lost money when applied to the training set.

Much work remains to be done on our approach. We presume that the models for goals and penalties can be improved, and that the simulation can be improved as well by incorporating more detail on the way games are actually played.

References

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

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

Buttrey, S. E., A. R. Washburn, and W. L. Price (2011): "Estimating NHL scoring rates," *J. Quant. Anal. Sports*, 7.

Gandar, J. M., R. A. Zuber, and R. S. Johnson (2004): "A reexamination of the efficiency of the betting market on National Hockey League games," *J. Sports Economics*, 5, 152–168.

MacDonald, B. (2012): "An expected goals model for evaluating NHL teams and players," in *Proceedings of the MIT Sloan Sports Analytic Conference 2012*.

Maher, M. J. (1982): "Modeling association football scores," *Statistica Neerlandica*, 36, 109–118.

Marschner, I. (2014): *glm2: Fitting Generalized Linear Models*, URL `http://CRAN.R-project.org/package=glm2`, r package version 1.1.2.

Morrison, D. (1976): "On the optimal time to pull the goalie: a Poisson model applied to a common strategy in ice hockey," *TIMS Studies in Management Science*, 4, 137–144.

NHL (2010): "Game Summary," URL `http://www.nhl.com/scores/htmlreports/20102011/GS020282.HTM`.

NHL (2015): "Columbus Blue Jackets – Hockey Analytics 101," URL `http://bluejackets.nhl.com/club/page.htm?id=99406`.

Paul, R. J. and A. P. Weinbach (2012): "Sportsbook pricing and the behavioral biases of bettors in the NHL," *J. Econ. Finan.*, 36, 123–135.

R Core Team (2013): *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, URL `http://www.R-project.org/`.

Robin, X., N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez, and M. Mller (2011): "pROC: an open-source package for R and S+ to analyze and compare ROC curves," *BMC Bioinformatics*, 12, 77.

Sauer, R. D. (1998): "The economics of wagering markets," *J. Econ. Literature*, 36, 2021–2064.

Thomas, A. (2007): "Inter-arrival times of goals in ice hockey," *J. Quantitative Analysis in Sports*, 3.

Woodland, L. M. and B. M. Woodland (2001): "Market efficiency and profitable wagering in the National Hockey League: Can bettors score on longshots?" *Southern Economic J.*, 67, 983–995.