# Markov Counting Process Description of a Sports League Applied to the National Hockey League

# Abstract

This note develops a Markov Counting Processes (MCP) formalism for describing the evolution of scoring totals and scoring rates of sports teams playing in leagues with discrete-scoring rules, then applies the Poisson distribution to the formalism to derive a formula for the probability a National Hockey League (NHL) team transitions to goals-for-per-game (GFg) at games from GFg at games. Using the last 10 NHL seasons of scoring outcomes taken from Hockeyreference.com, transition probability predictions were calculated and compared to the derived MCP transition probability predictions with a linear regression. The resulting fit parameters with 95% confidence interval were [slope, intercept] = [0.9940.071, 0.0320.034], the r-squared value was .887, and the residuals were approximately normally distributed with a mean of approximately zero. The fit parameters, degree of correlation, and residual characteristics are interpreted as implying that that the derived MCP transition probabilities are equal to the NHL data transition probabilities plus noise resulting from limited data and estimations made in the model, suggesting that modeling the NHL as a set of Poisson Markov Counting Processes is justified.

1. Introduction

Applications of statistics to sports leagues are often efforts to inform predictions of outcomes of future league matchups. These efforts typically fit model parameters to past league matchup outcomes, finding strength estimates for each team that, under the chosen model’s assumptions, maximize the likelihood of having observed those past league matchup outcomes. Once the most probable team strengths are determined, the probability one team beats another can be estimated using a function of the two team’s strengths, and the estimating party can use the estimated probability to inform a prediction.

These team strengths are, at a basic level, reflections of team scoring rates – the more scores a team earns per game and the less scores a team allows per game, the stronger the team. Despite that fact, there seems to have been little theoretical work done to describe the evolution of team scoring rates over the course of a season. The question representing the knowledge gap this paper aims fill is therefore, “If a team has scoring rate at games into its season, what is the probability it will have scoring rate at games into its season?”

Section two begins to answer the representative question by introducing the MCP representation of a sports team, defining formalism for representing the probability of observing every possible season scoring total outcome for a team, and presenting actual NHL season scoring total outcomes as example MCP realizations. Section three extends the formalism to define a Markov chain (MC) representation of team season scoring rate outcomes, then presents actual NHL season scoring rate outcomes as example MC realizations. Section four specifies Poisson Counting Processes (PCPs) as approximate representations of NHL teams, then presents the Poisson version of the developed formalism. Section five describes the method used to compare MC scoring rate predictions and NHL data predictions for various parameter sets , then discusses the results. At the end, I provide derivations for starred equations.

# Season Scoring as a Markov Counting Process

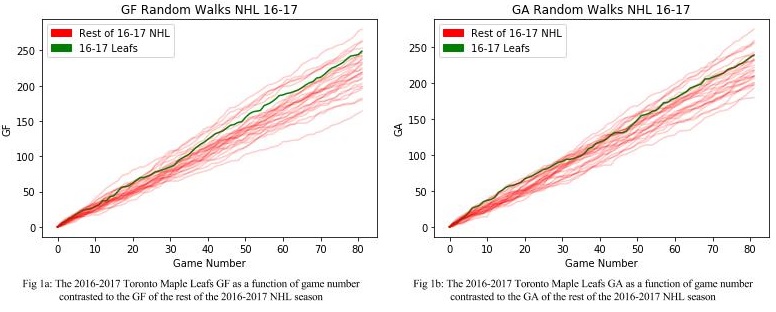
A time-domain Counting Process (CP) counts the number of events that occur during a period of time. The count, defined as the total number of events that have occurred up until time , is non-decreasing and takes only integer values. The state of the process at time is and the state space the process exists in is C(t)’s range. Consider, as an example, repeatedly rolling a regular die. The sum of the values of the roll outcomes can be considered a discrete CP: C(1) – the count after the first roll – would be a random draw from Uniform(1,6), the Uniform distribution that can take values between one and six; C(2) would be the sum of two random draws from Uniform(1,6); and C(t) would be the sum of random draws from Uniform(1,6). The state space the CP exists in would be the set of integers between 1\*t and 6\*t, because those are the smallest and the largest possible outcomes for a single roll, respectively. Because each roll of the die is an independent event, the probability C(t+1) takes some value depends only on C(t), and the die-rolling CP can be considered Markovian – a Markov Counting Process.

In many popular sports such as Football, Soccer, Basketball, Baseball, and Hockey, only a positive integer number of scores can be earned in a game. Additionally, the scores earned by the same team in different games are assumed to be the outcomes of independent events. As a result, the sum of the scores a team has earned in a season can be represented by a MCP. Similarly, the sum of the scores a team allows in a season can be represented by its own MCP. Each game of a season can be compared to rolling two die; one of the results is added to the home team’s scores-for (SF) total and the away team’s scores-against (SA) total, and the other is added to the home team’s SA total and the away team’s SF total. A team’s season scoring total outcome can, therefore, be represented by a point moving through the two-dimensional space of all possible season SF totals and SA totals.

Team 𝑖, for example, can be said to be in state at 𝑛 games into its season, where is the point with components equal to team 𝑖’s SF and SA totals 𝑛 games into its season. In a league with 𝐴 teams, for all teams 𝑖 = *1*. . . 𝐴 the initial state is because all teams start a season with zero SF and SA. Playing the 𝑛th game in its season transitions team 𝑖 from state to state . For an 𝑁 game season, each team 𝑖 ends its season in state , and the components of that state are that team’s end of season SF and SA totals.

Consecutive state components are related by and state components separated by games are related by , where is a vector of random variables representing team i’s 𝑛th pair of MCP increments - the amount of scores it earns and allows in the 𝑛th game of its season.

The set of states [ ∀ 𝑛 = 0 … 𝑁] is a realization of a MCP process and defines team 𝑖’s scoring total “path”. As an example of what realized MCP paths look like, figure 1 plots the 2016-2017 Toronto Maple Leafs SF and SA with respect to season game number along with the rest of the teams in the 2016-2017 NHL season.

****

The probability that component 𝑝 of team 𝑖’s state vector transitions from = 𝑢𝑝 to is

where is the probability that the random variable is equal to 𝑣𝑝 − 𝑢𝑝. The probability that component 𝑝 of team 𝑖’s state vector transitions from = 𝑢𝑝 to is similarly

can be interpreted as the entry in the row and column of a transition probability matrix. Over a season the total number of scores earned and allowed by a team can only increase, and hypothetically any positive integer number of scores can be earned, so the matrix with entries is an infinite dimensional upper triangular matrix with each entry representing a point in .

The amount of goals a team scores and allows in a game are assumed to be independent random variables, as has been shown to be reasonable [1][2]. The transition probability matrix representing the probability team 𝑖 moves from state = 𝑢 =[𝑢0, 𝑢1] to state = 𝑣 = [𝑣0, 𝑣1] is then the outer product of the component transition probability matrices

Each entry of can also be interpreted as the sum of all likelihood functions that transition team from state to state in steps.

# Team Strength and Strength Transitions

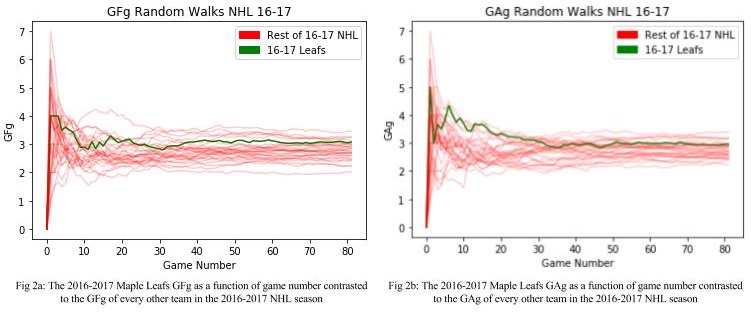
The rates-of-change of a team’s scores-for total and of its scores-against total are that team’s scores-for-per-game (SFg) and scores-against-per-game (SAg). The difference between a team’s SFg and SAg can be considered its “strength”, in the sense that the more scores a team earns per-game and the fewer scores a team allows per-game the stronger the team. For the NHL, the amount of points a team earns in a season – which determines the rank of a team compared to the rest of its league - has been shown to strongly correlate with the difference in that team’s SFg and SAg [3]. If an game NHL season is simulated, as tends to infinity the correlation tends to one. Something something different point-earning scheme as different sorting mechanisms blah blah. Something something find the point-earning scheme that most reliably sorts by the end of a finite length season blah blah.

A team’s scoring rate, being a ratio of a positive-integer-number of scores and a positive integer number of games, has a domain of. If team 𝑖 is in total scoring state , then that team’s scoring rate state is

Consecutive scoring rate state components are related by the same random variable as consecutive total scoring state components

And, similarly, scoring rate state components separated by games are related by

Where the star in the equation number implies a derivation is provided in the appendix. Since each depends explicitly on , the set of values [ ∀ 𝑛 = 0 … 𝑁] is a time-inhomogeneous Markov chain. As an example of what scoring rate Markov chains look like, figure 2 displays the 2016-2017 Toronto Maple Leafs SFg and SAg (denoted as GFg and GAg because scores are called “goals” in the NHL) with respect to season game number along with the rest of the teams in the 2016-2017 NHL season



The probability that component 𝑝 of team ’s scoring rate state transitions from to is . Solving for the sum of random variables in the definition of lets it be stated that

Where are the constraints resulting from requiring the number of goals to be a positive integer. Similarly, the probability that component of team 𝑖’s scoring rate transitions from to is

.

can be interpreted as the entry in the row and column of a transition probability matrix. The transition probability matrix for team 𝑖 to move from strength = 𝑢 =[𝑢0, 𝑢1] to strength = 𝑣 = [𝑣0, 𝑣1] is then the outer product of the component strength transition probability matrices

The non-zero entries of the matrices [ define the complete set of allowable scoring rate outcomes for a season. A non-allowable strength result, for example, would be a team having , because that would imply the team had 25.5 scores.

1. Application of the Markov Chain Model to the NHL

The formalism developed in sections two and three is not sport specific – the transition probabilities can be applied to any sport in which the number of scores earned in a game can be approximated as a draw from some positive-integer-domain probability distribution. To apply the formalism to a specific sport, the probability distribution representing the likelihood that a certain amount of scores are earned in a game for that sport must be inserted into the transition probabilities.

Scoring in hockey has been shown to be approximately Poisson [1][2][4]. That is, the number of goals a team scores and allows in a game can be approximated as independent random draws from Poisson distributions.

Where and represent the number of goals team is expected to score and allow, respectively, in game . Hockey is not a purely Poisson process, though, as expected scoring is not uniform over every slice of time in a game. Goals are scored at a disproportionately high rate at the ends of periods - especially at the end of the third period – and during odd man situations [1][2][5].

Ignoring the NHL’s deviation from a purely Poisson process and dropping the component notation for clarity, the probability team ’s total scoring state transitions from state to state in game can be written as

The sum of Poisson variables is a Poisson variable with a mean of the sum of the constituent variable means. Consequently, the probability team ’s total scoring state transitions from state in game to state in game can be written as

And the scoring transition probability matrix can be written as

Scoring rate transition probabilities depend on the same Poisson random variable as total scoring transition probabilities. As such, the probability team ’s scoring rate transitions from state in game to state in game is

And the strength transition probability matrix entries can be written as

1. Comparing Predicted Transition Probabilities to NHL Data

In order to determine if the transition probability actually represents the probability an NHL team goes from scoring rate at game to scoring rate at game , transition probabilities estimated using equation 15 must be compared to values calculated using actual NHL data. Ideally, the Markov chain probabilities calculated using equation 15 could be directly contrasted to Monte-Carlo probabilities calculated using an NHL data sample population. For example, the set of A NHL teams that reached state and the subset of B teams that subsequently reached state would be counted. The probability an NHL team reaches state given it was in state could then be estimated by and compared to .

Unfortunately, for most transitions, there haven’t been enough teamsthat reached the exact same final state from the exact same initial state in order to directly make meaningful estimates of transition probabilities. Instead, in order to make transition probability estimations with NHL data, a range of potential and values were tolerated.

Allowing for a tolerance in and results in the following method for estimating transition probabilities with NHL data: the set of A NHL teams with in the range and the subset of B teams with in the range are counted. The probability a team reaches a state in the range given it had a state in the range is then estimated by

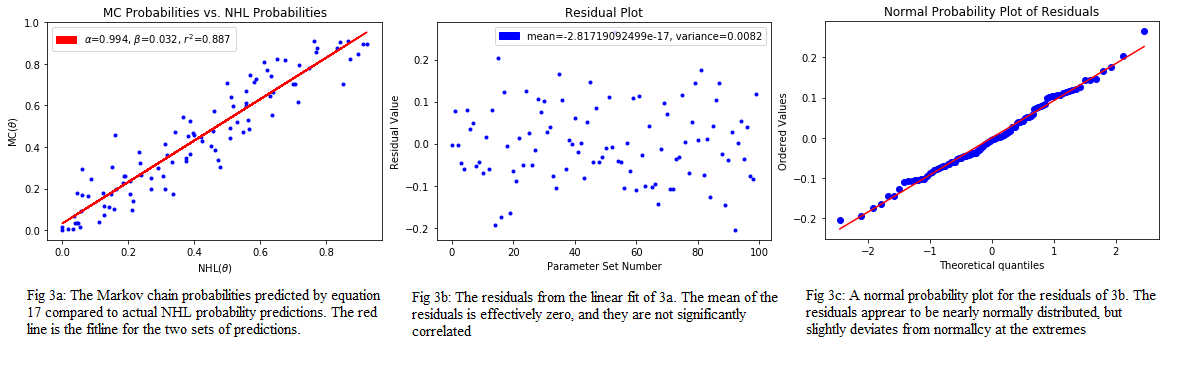
For example, in the last 10 NHL seasons there have been 52 teams that had between 2.9 and 3.1 GFg 41 games into their season. Of those teams, 19 had between 2.9 and 3.1 GFg 82 games into their season. NHL data therefore estimates that the probability a team reaches a state in the range given it had state in the range to be Probabilities produced with this method can be contrasted to a modified version of (15):

Where the domain of is all allowable values of in the range ; is an approximation of the expected amount of goals in a game for the average NHL team, taken to equal the last 10 NHL seasons’ average GFg; and MC stands for Markov Chain. Applying (17) to the preceding sample parameter set results in

100 parameter sets were pseudo-randomly generated, and the values of and were calculated for each. A linear regression was then run between the two sets of resulting probabilities. The fit takes the form

where and are constants and is a Normally distributed error term. If were equal to for every parameter set, then the fit parameters would be , the r-squared value would be and each error term would be zero. Only a finite number of NHL games have been played, is only an approximation, and hockey is not a purely Poisson process, though, so both and are expected to only be approximations of the true probability corresponding to the parameter set As a result, some deviation of the data from the ideal fit parameters, an r-squared value less than 1, and non-zero error variance should be expected.

Average values of the fit parameters were found to be , the average r-squared value was found to be , and the errors were approximately Normally distributed with a mean of zero and a variance of .0082. The parameter results are only point estimates, but the true parameters are in the ranges and with 95% confidence. These results suggest that is equal to plus noise resulting from the imperfect approximations made in calculating the two predictions:



Residual analysis of the linear fit suggests that, while the relation between and is not perfectly linear, the linear fit is a reasonable approximation. The success of the Markov chain equations in modeling NHL data suggests that modeling the NHL as a set of Markov chains with Poisson transition probabilities is justified.

Derivations of Starred Equations

(6):

(8):

The sum of goals must be a positive integer, therefore allowable values of and for given are

(15):

(17): By the definition of conditional probability

This derivation finds expressions for both the denominator and the numerator of the above quotient individually, then combines them to form an expression for .

Denominator: The probability a team will have is the probability that the sum of the goals scored in the first n games equals :

For some expected amount of goals in a game . The probability a team will have is then the sum of over all values of such that

Numerator: For a team to go from to , there are a minimum amount of goals and a maximum amount of goals that can be scored. The minimum amount of goals would take the team to , and the maximum amount of goals would take them to . The probability a team has given it had is then

Since goals in different games, and therefore different clusters of games, are assumed to be independent variables, the probability a team will go from to and from to is. The probability a team will have and is the sum of over all values of such that . Then,

Then, by combining the derived numerator and denominator with the definition of conditional probability,

Citations

[1] Dayaratna, K. and Miller, S. (2012). *Pythagorean Win-Loss formula and Hockey*

[2] Ryder, A. (2004). *Poisson Toolbox: a review of the application of the Poisson Probability Distribution in hockey*

# [3] Found, R. (2016). *Goal-based Metrics Better Than Shot-based Metrics at Predicting Hockey Success*

[4] Morrison, D.G. (1976). *On the optimal time to pull the goalie: A Poisson model applied to a common strategy used in ice hockey.* TIMS Studies in Management Science, Vol. 4.

[5] Thomas, A. (2007). *Inter-Arrival Times Of Goals In Hockey*